Editorial

THIS issue of *Polibits* includes a thematic selection of papers on Information Retrieval and Natural Language Processing. Information Retrieval comprises the technologies that help searching for documents in large text collections or Internet. Natural Language Processing, also known as Computational Linguistics, offers improvements to many applications that involve computers and human language, such as information retrieval, spelling correction, and machine translation, among others.

The **special section on Information Retrieval and Natural Language Processing** includes the first seven papers. First three of them are related with **Information Retrieval** and **Question Answering**. Question answering is a specific kind of information retrieval: instead of searching for whole documents, it goes a step further by providing the user with the answer to a specific question. For example, if the user wants to know *who is the president of Mexico*, an information retrieval application will present them all the documents that mention the words *president of Mexico* and the user themselves can manually look for the name in those documents; a question answering application will do it automatically and return the name: *Felipe Calderon*.

The paper "*TrainQA: a Training Corpus for Corpus-Based Question Answering Systems*" describes the development of lexical resource useful for the question answering task: a corpus of questions and their corresponding answers marked in the text. This corpus can be used for both supervised training of question answering systems and as a benchmark for such system.

The paper "*Cross Language Information Retrieval using Multilingual Ontology as Translation and Query Expansion Base*" evauates an information retrieval system of a special type: cross language information retrieval system, i.e., a system that accepts a search query in one language (English in this case) but searches for relevant documents written in another language (Arabic). The paper demonstrates that using an ontology yields significantly better results than using a simple dictionary for translation.

The paper "*English-to-Japanese Cross-Language Question-Answering System using Weighted Adding with Multiple Answers*" is devoted to a topic that combines the ideas of question answering and cross-language information retrieval: now, not only the question can be formulated in a language different from that of the documents, but the answer can be looked in texts in different languages.

The next two papers are devoted to the internal tasks of natural language processing and information retrieval, namely, to **word sense disambiguation** and **named entity extraction**.

The paper "*Using Sense Clustering for the Disambiguation of Words*" is devoted to disambiguation of homonymous or polysemous words in a specific context. For example, the word *bank* means different things in the contexts *bank account* and *West bank of Jordan*. The paper presents a method for automatically determining the correct sense of the word in context by determining groups of similar senses.

The paper "*Improving Named Entity Extraction Accuracy using Unlabeled Data and Several Extractors*" shows how several methods can be combined to improve the quality of identifying named entities in natural language documents. A named entity is a sequence of words that refer to a single concept, such as *Ministry of Foreign Affairs* or *John Smith*.

The last two papers of the special section address **spelling correction** in information retrieval context and **lexical resources acquisition**.

The paper "*Revised N-Gram based Automatic Spelling Correction Tool to Improve Retrieval Effectiveness*" presents an improved method for correcting spelling errors in words via statistics of letter combinations. Besides the usual application of the spelling correction in improving human writing, the paper shows that spelling correction is useful for information retrieval by correcting spelling errors in the query and in the documents being searched and thus improving chances for correct matching between the query and the documents.

Finally, the paper "*Bilingual Lexical Data Contributed by Language Teachers via a Web Service: Quality vs. Quantity*" discussed the issue of lexical resources acquisition. Many of the natural language techniques depend crucially on dictionaries and other sources of knowledge about language. Constructing such dictionaries is a major concern for the research community. The paper discusses the authors' experience in involving language teachers in a collaborative project aimed to develop a bilingual dictionary.

This paper concludes the special section on information retrieval and natural language processing. The last five papers are **regular papers**.

The paper "*Tecnología RFID Aplicada al Control de Accesos*" presents a short introduction to the modern radio frequency identification technology that permits remotely identify objects or resources via special tags attached to them. The paper also shows a practical application of this technology in access control.

The paper "*An Extended Payment Model for M-Commerce with Fair Non-Repudiation Protocols*" suggests an improved model for mobile commerce using fair non-repudiation protocols widely used in the context of electronic commerce but not studied in enough detail in the context of mobile commerce. Non-repudiation protocols ensure that the buyer or

seller will not be able to refute the validity of the transaction after it has been concluded.

The paper "*Análisis Numérico de Pérdidas de Inserción de Conmutadores Diseñados con Diodos p-i-n*" gives a detailed introduction in the theory and applications of the *p-i-n* diodes and microwave switches used in high-frequency circuits. The paper then presents numeric analysis of the insertion loss in microwave switches designed with the *p-i-n* diodes.

The paper "*Restricción del Uso de Teléfonos Celulares en Ambientes Controlados*" presents a Bluetooth-based tool that can disable mobile phones at a certain place or block some of their functions such as transmitting of video signal. This allows to enforce the restrictions on use of mobile phones at some public places such as cultural events, banks, or airplanes.

Finally, the paper "*Evaluation of E-Learning Readiness: A Study of Informational Behavior of University Students*" studies the way students in Thailand use information sharing and search in the computer-assisted learning process. The results of this study can be used for elaboration of the part of e-learning curriculum that deals with personal information management and can be relevant for countries with similar cultural and economic situation, such as most of Latin American countries.

Alexander Gelbukh
Head of the Natural Language Processing Laboratory,
Center for Computing Research, National
Polytechnic Institute, Mexico

TrainQA: a Training Corpus for Corpus-Based Question Answering Systems

David Tomás, José L. Vicedo, Empar Bisbal, and Lidia Moreno

***Abstract*—This paper describes the development of an English corpus of factoid TREC-like question-answer pairs. The corpus obtained consists of more than 70,000 samples, containing each one the following information: a question, its question type, an exact answer to the question, the different contexts levels (sentence, paragraph and document) where the answer occurs inside a document, and a label indicating whether the answer is correct (a positive sample) or not (a negative sample). For instance, TrainQA can be used for training a binary classifier in order to decide if a given answer is correct (positive) to the question formulated or not (negative). To our knowledge, this is the first corpus aimed to train on every stage of a trainable Question Answering system: question classification, information retrieval, answer extraction and answer validation.**

***Index Terms*—Question answering, corpus-based systems.**

I. Introduction

EMPIRICIST approach to Natural Language Processing (NLP) suggests that we can learn the complicated and messy structure of language studying large amount of real-life language samples by means of different techniques such as statistical, pattern recognition or machine learning methods. This data-driven approach is based on large corpus, i.e., large body of language data: written texts, spoken discourse, samples of written or spoken language.

Many researchers agree that significant progress can be made in text understanding by attempting to automatically extract information about language from very large corpora. For this reason, many resources have been developed to assist the learning task. These text resources present different levels of annotation that determine the task they are useful for. There are plain corpus like Project Gutenberg[1] that present no extra information, but plain text. There are also corpus like Spanish EFE Press Agency news of 1994 and 1995 (see CLEF[2]), with formatting attributes that identify information about edition, authors, headlines or paragraphs. Finally, there are annotated corpus like Penn Treebank [1] with more elaborated information about part of speech or syntactic structure. All these are general corpora not intended for a concrete task.

Manuscript received November 23, 2008. Manuscript accepted for publication August 15, 2009.

David Tomás and José L. Vicedo are with the Department of Software and Computing Systems, University of Alicante, Spain (dtomas@dlsi.ua.es, vicedo@dlsi.ua.es)

Empar Bisbal and Lidia Moreno are with Department of Information Systems and Computation, Technical University of Valencia, Spain (ebisbal@dsic.upv.es, lmoreno@dsic.upv.es)

[1]http://www.gutenberg.org

[2]http://www.clef-campaign.org

In this paper we present a corpus developed to assist data-driven Question Answering (QA) systems. These systems try to obtain exact answers from large corpus to precise questions formulated in natural language. We have developed a corpus of English question-answer pairs suited to train on every stage of a machine learning based QA system: question classification, information retrieval, answer extraction and answer validation.

The corpus consists of more than 70,000 samples. Each of these samples contains information that relates a question with its answer in four different contexts: exact match, sentence, paragraph and document. Every sample is labelled as *positive* or *negative*, depending whether the answer given is correct or not. Negative instances are useful to provide the context in which an extracted answer is incorrect. We obtained a total of 7,598 positive samples and 64,384 negative ones. This way, the corpus can be used to train a binary classifier in order to decide if a given answer is correct (*positive*) to the question formulated or not (*negative*). Moreover, information about question type is also stored to assist the question classification process.

Other corpora have been previously employed to train isolated parts of a QA system. Nevertheless, to our knowledge this is the first corpus that can be used to train all the different components of a QA system and also, the only one that contains positive and negative instances.

This paper is organized as follows: in Section II we introduce the current research related to corpus development and trainable QA systems; Section III describes the samples that make up the corpus; Section IV presents the resources employed to build the corpus and details the generation process; Section V outlines corpus statistics and finally, in Section VI we discuss possible corpus applications and main challenges for future work.

II. Related Work

There are several QA systems that apply machine learning techniques based on corpus of question-answer pairs, covering different stages of the question answering process.

In [2], a corpus of question-answer pairs (called KM database) was developed. Each of the pairs in KM represents a trivia question and its corresponding answer, such as the ones used in the trivia card game. The question-answer pairs were filtered to retain only questions and answers that look

similar to the ones presented in the TREC task[3]. Finally, 16,228 pairs were obtained in all. Using this corpus as seed, they automatically collected a set of text patterns which are used for answer extraction purposes.

In [3], they built their QA system around a noisy-channel architecture which exploited both a language model for answers and a transformation model for answer/question terms. In order to apply the learning mechanisms, they first built a large training corpus consisting of question-answer pairs of a broad lexical coverage. They collected FAQ pages and obtained a total of roughly 1 million question-answer pairs. They applied this training corpus in the query analysis and answer extraction modules. This system was intended to be applied to non-factoid questions.

The system developed by [4] used a collection of approximately 30,000 question-answer pairs for training, obtained from more than 270 FAQ files on various subjects in the FAQFinder project [5]. They used this corpus to automatically learn phrase features for classifying questions into different types, to generate candidate query transformations, and to evaluate the candidate transforms on target information retrieval systems such as real-world general purpose search engines.

The approach in [6] is heavily inspired by machine learning. Starting from a large collection of answered questions, the algorithms described learn lexical correlations between questions and answers. To serve as a collection of answered questions, they assembled two types of data sets: 1,800 pairs from Usenet FAQs and 5,145 from Call-center dialogues.

All the corpora mentioned above present some of the following problems when employed to train a QA system:

- No question type is given, so that the corpus can not be employed in the question classification stage.
- Negative samples are not provided, which are useful for a classifier to determine the context of incorrect answers.
- The context where the answers occurs is inadequate for different QA stages: too constrained for information retrieval or too loose for answer extraction.

We have developed a corpus that overrides all these problems. It consists of question-answer pairs in XML format that have been obtained from TREC[4] resources (specifically TREC QA track questions and corpora). This way, we have gathered a corpus of factoid TREC-like questions and answers fully oriented to the QA task. Unlike the other approaches, every sample is tagged with a question type that makes them useful for question classification. The corpus presents four different context levels for every answer that make them suitable for every QA stage: document and paragraph context for information retrieval, sentence context for answer validation and exact match for answer extraction. Moreover, our corpus contains correct and incorrect answers, which means that we have pairs labelled as *positive* or *negative* that can be very useful to train binary classifiers. Finally, the number of samples obtained (over 70,000) makes the corpus appropriate for machine learning purpose.

[3]Questions with 10 words or less, and were not multiple choice.

[4]Text REtrieval Conference: http://trec.nist.gov

III. Corpus Description

The corpus developed consists of a set of English question-answer pairs samples including the following fields:

- The number of sample, used as identifier.
- The number of question in the TREC set.
- The question itself.
- A question type indicating the class of the question from a taxonomy of fifteen different classes (see [7]) such as LOCATION, PROPER_NAME, EVENT, ORGANIZATION, ACRONYM, ... This information is useful for the question classification task, where a class or category is assigned to the question proposed.
- The exact answer string. This information can assist the answer extraction process, which allows to obtain nothing but the exact answer to the question formulated.
- The textual context, with the size of a sentence, where the answer was found. This information along with the question, can be employed to train textual entailment systems [8] which can cope with answer validation processes.
- The textual context, with the size of a paragraph, where the answer was found. This information is useful to train a passage retrieval system in order to discriminate between relevant and non relevant paragraphs.
- The identifier of the document where the answer was found. Documents are useful to train good document retrieval or document re-ranking systems to reject non answer bearing documents.
- A label indicating whether the answer is correct (*positive* sample) or incorrect (*negative* sample). This way, binary classifiers can be trained with our corpus in order to determine if an exact answer, a sentence, or a paragraph fit the given question.

Figures 1, 2 and 3 show three different corpus samples for the question "*Who is Tom Cruise married to?*". The question type is PROPER_NAME, indicating that it expects the name of a person as answer. In Fig. 1, the answer given "*Nicole Kidman*" is correct, and the context (sentence and paragraph) justifies it. Consequently, the sample is classified as *positive*.

In the sample included in Fig. 2, "*Nicole Kidman*" is also given as response, but in this case, the context does not support the answer. This sample is therefore classified as *negative*.

In Fig. 3 "*Bill Harford*" is the answer, and besides this is not true, the context where it was extracted from does not justify it anyway. This sample is also classified as *negative*.

IV. Building the Corpus

First subsection describes the resources necessary to build the corpus. The next one describes the process carried out to obtain the set of samples that make up the corpus.

```
<SAMPLE id="26821" class="POSITIVE">
  <QID>
    1395
  </QID>
  <QUESTION>
    Who is Tom Cruise married to?
  </QUESTION>
  <QTYPE>
    PROPER_NAME
  </QTYPE>
  <ANSWER>
    Nicole Kidman
  </ANSWER>
  <SENTENCE>
    The drama is said to be about a pair of married psychiatrists (played by
    the married Tom Cruise and Nicole Kidman) and their sexual lives, but
    only a few Warner executives, Cruise and Kidman, and Pat Kingsley, a top
    public relations executive, have seen the film.
  </SENTENCE>
  <PARAGRAPH>
    Along the way, Kubrick's secretive methods generated a continual buzz. Actors
    had to sign agreements not to talk to the press, and shooting scripts were kept
    under strict security. The drama is said to be about a pair of married psychiatrists
    (played by the married Tom Cruise and Nicole Kidman) and their sexual lives, but
    only a few Warner executives, Cruise and Kidman, and Pat Kingsley, a top public
    relations executive, have seen the film.
  </PARAGRAPH>
  <DOCID>
    NYT19990326.0303
  </DOCID>
</SAMPLE>
```

Fig. 1. An example of a *positive* sample from the corpus. Information is separated in different tags: the identifier of the sample (attribute `id` in tag `SAMPLE`), the class indicating whether it is positive or not (attribute `class` in tag `SAMPLE`), the identifier of the question (tag `QID`), the question itself (tag `QUESTION`), the question type (tag `QTYPE`), the exact answer (tag `ANSWER`), the sentence context (tag `SENTENCE`), the paragraph context (tag `PARAGRAPH`) and the document identifier (tag `DOCID`).

```
<SAMPLE id="26824" class="NEGATIVE">
  <QID>
    1395
  </QID>
  <QUESTION>
    Who is Tom Cruise married to?
  </QUESTION>
  <QTYPE>
    PROPER_NAME
  </QTYPE>
  <ANSWER>
    Nicole Kidman
  </ANSWER>
  <SENTENCE>
    The film itself, starring Tom Cruise and Nicole Kidman as a married couple in
    New York on a sexual odyssey, received wildly mixed reviews.
  </SENTENCE>
  <PARAGRAPH>
    The film itself, starring Tom Cruise and Nicole Kidman as a married couple in
    New York on a sexual odyssey, received wildly mixed reviews. After strong box
    office sales in its first weekend, attendance has dropped sharply.
  </PARAGRAPH>
  <DOCID>
    NYT19990326.0303
  </DOCID>
</SAMPLE>
```

Fig. 2. A negative sample. Despite the answer is correct, the context does not justify it.

```
<SAMPLE id="26831" class="NEGATIVE">
    <QID>
      1395
    </QID>
    <QUESTION>
      Who is Tom Cruise married to?
    </QUESTION>
    <QTYPE>
      PROPER_NAME
    </QTYPE>
    <ANSWER>
      Bill Harford
    </ANSWER>
    <SENTENCE>
      The story follows the descent of Bill Harford (Cruise, toothy as ever), a successful young
      doctor on the Upper West Side of Manhattan, into a perilous, secretive netherworld.
    </SENTENCE>
    <PARAGRAPH>
      At the same time ''Eyes Wide Shut'' is a sternly anti-erotic movie that regards its sexual
      license with a cold puritanical hauteur. The movie is not a turn-on (it is really a horror
      film without gore), and the sexual chemistry between its married stars, Tom Cruise and
      Ms. Kidman, is tepid at best. The story follows the descent of Bill Harford (Cruise, toothy as
      ever), a successful young doctor on the Upper West Side of Manhattan, into a perilous, secretive
      netherworld. The catalyst is a confession by his wife, Alice (Ms. Kidman), about the fierce,
      unconsummated desire she once felt for a young naval officer. In black-and-white sequences that
      punctuate the movie, Bill torments himself with visions of Alice and her would-be lover in bed
      together, and these images drive him to examine his own wayward impulses.
    </PARAGRAPH>
    <DOCID>
      NYT19990719.0343
    </DOCID>
  </SAMPLE>
```

Fig. 3. A negative sample from the corpus with incorrect answer.

A. The Resources

The resources necessary to build this corpus were obtained from the Question Answering collections in TREC conferences.

In order to collect question-answer pairs, we wanted to focus on questions with exact answers, so only questions formulated from TREC 2002 to TREC 2005 competitions were taken into account. On previous QA tracks (TREC 1999 to TREC 2001), systems were asked for passages instead of exact answers, so we discarded them. For the same reason, only questions in "main" subtask were collected, avoiding "list" or "passage" queries. We finally gathered a collection of 1,505 typical factoid TREC-like questions. The TREC 2004 and 2005 questions sets had to be reviewed as their format slightly differs from the previos competitions. In this case, a *target* was given (i.e. "*Horus*") and questions referred to that target were formulated ("*What country is he associated with?*"), so that we had to manually reformulate them in other to obtain an homogeneous question set ("*What country is Horus associated with?*") with no anaphoric references.

We also used the AQUAINT[5] document collection, which is also part of the resources of the TREC QA track. This collection was used to obtain the contexts where answers to the selected TREC questions occurred. It consists of 1,033,461 documents in English with roughly 375 million words, drawn from three sources: the Xinhua News Service (People's Republic of China), the New York Times News Service, and the Associated Press Worldstream News Service. This document set was used in the last QA tracks, from TREC 2002 to TREC 2005.

Finally, we used the judgement set files from TREC 2002 to TREC 2005. These files contain information about all submissions to the track. A judgement consist of four fields:

- The question number.
- The identifier of the document on the AQUAINT collection that supports the answer.
- The judgement made by the assessors.
- The answer string.

The judgement made by assessors indicates whether the answer is correct, incorrect, inexact or unsupported."Unsupported" means that the string contains a correct response, but the document returned with that string does not allow one to recognize that it is a correct response. "Inexact" means that the answer string contains a correct answer and the document supports that answer, but the string contains more than just the answer or is missing bits of it. See [9] for detailed description on how answer strings were judged. Figure 4 shows a snippet of these files.

B. The Process

The corpus was semi-automatically obtained from the resources described above, by means of automatic extraction

[5]Linguistic Data Consortium (LDC) catalog number LDC2002T31 and ISBN1-58563-240-6.

TABLE I
CORPUS STATISTICS

Set	Questions	Judgements	Positive	Negative	Total Samples
TREC 2002	500	15,948	1,837	24,818	26,655
TREC 2003	413	9,841	2,359	15,070	17,429
TREC 2004	230	6,235	1,235	8,219	9,454
TREC 2005	362	11,967	2,167	16,277	18,444
TOTAL	**1,505**	**43,991**	**7,598**	**64,384**	**71,982**

```
1395  NYT19991220.0294  -1  Julia Roberts
1395  NYT19991101.0416   1  Nicole Kidman
1395  APW19990712.0006   3  actress Nicole Kidman
1395  NYT19991101.0416   3  actress Nicole Kidman
1395  APW19990712.0006   1  Nicole Kidman
1395  APW19990423.0019   2  Tom Cruise and Nicole Kidman
1395  NYT19990628.0254   2  Nicole Kidman
```

Fig. 4. Judgement file snippet. The third column indicates if the answer is incorrect (-1), correct (1), unsupported (2) or inexact (3).

of the samples and subsequent manual revision of the *positive* ones, as we will describe later.

First, all the factoid questions from TREC 2002 to TREC 2005 QA tracks were collected. These questions were then manually labelled with their respective question type according to the classification presented in [7].

For every question gathered, an automatic process looked into the judgement set files for all its related submissions. These submissions reflect what the systems participating in these QA tracks replied to the questions formulated during the competition. Each judgement was processed following these steps:

1) Read the answer given to the question.
2) Read the judgement made by the assessors.
3) Read the document identifier and try to match the answer in the document.
4) Retrieve and store all the different paragraphs in the document that contain the answer. As every paragraph is already tagged in the AQUAINT corpus, these tags[6] are employed to easily extract them. A sample is generated for every paragraph.
5) Extract from every paragraph the sentence where the answer occurs. The MXTERMINATOR software [10] was employed to detect sentence boundaries.

At this point of the automatic process, we had a set of samples connecting a question with its exact answer and with the different contexts (sentence, paragraph and document) where this answer was found. If the assessors judged the answer as "incorrect", the sample was labelled as *negative*. If the judgement was "correct", the sample was labelled as *positive*. In case the judgement was "unsupported", the sample was labelled as *negative*, since the answer is correct but the context does not justify it.

Judgements labelled as "inexact" demand a special treatment. In this case, the document justifies the answer but this answer does not perfectly fit the user needs as the string contains extra information or is missing bits of it. To solve this problem, the automatic process gathers all the "correct" answers given to the question in the judgement set, and tries to match them in the document where the "inexact" answer was found. For instance, let's suppose Fig. 4 shows all the judgments for question number "1395". When the third judgement of the file is processed, the answer given is "*actress Nicole Kidman*", that was judged as "inexact" for the TREC assessors. In that case, the automatic process looks for all the possible "correct" answers for question "1395" in the judgement set and tries to match them with document "APW19990712.0006", where the "inexact" answer occurred. In this example, only "*Nicole Kidman*" from the second and fifth judgment is correct, so this exact answer is searched in document "APW19990712.0006". Samples obtained this way are labelled as *positive* as inexact answers are now substituted with exact ones.

After finishing the automatic process, we had a large set of samples with the information shown in Figs. 1, 2 and 3. But the whole corpus development process is not completed as there was a problem with some pairs that had to be manually reviewed.

There are two different problems with the automatic extraction process. First, in some cases the answer obtained from the judgement set occurred in different paragraphs inside the same document. The automatic process extracted every matching paragraph and created a sample for each one, labelling all of them either as *positive* or *negative* according to the criterion described above. There is no problem if the label assigned is *negative* as we can assure for every sample that the answer is not correct or the paragraph does not support it. The problems arise when the samples are all labelled as *positive*, because we can not guarantee that all the paragraphs matching the answer in the document support it.

Secondly, in some cases there are anaphoric references between paragraphs in the documents so that the answer an its justification appear in different paragraphs. Thus, as we have established, these samples can not be considered *positive* as the context does not justify them.

This ways, all the *positive* samples were set apart for manual

[6]For some reason the paragraphs in the 1998 Associated Press Worldstream News Service corpus are not labelled, thus answers related to this collection were not taken into account.

review in order to decide whether they are correctly labelled as *positive* or must be changed to *negative*. This reviewing task was carried out by two assessors that decided separately if the *positive* label was correctly assigned. A total of 7,598 samples were reassessed with a kappa agreement of 0.94. The expected agreement was computed according to [11], taken as equal for the coders the distribution of proportions over the categories. In case there was no agreement, a third adjudicator made the final determination.

V. Corpus Statistics

The TrainQA corpus has 71,198 samples from 1,505 different questions (47.31 samples per question on average). The number of *positive* samples collected was 7,598, while the number of *negative* samples was 63,384. The amount of *negative* samples (89%) largely exceeds the amount of *positive* ones (11%). We decided to keep this proportion in the corpus since this are the results of real QA systems submissions.

Table I shows the final corpus statistics. For each TREC competition we include the partial results obtained. Last row shows total results. "Set" column indicates which conference provides textual resources. "Questions" indicates the number of questions used to extract question-answer pairs. "Judgements" indicates the number of judgments included in the judgement set files, that is, the total number of submissions made by participants. "Positive" shows the number of samples labelled as *positive*, while "Negative" shows *negative* ones. Finally, column "Total" summarizes the total number of samples gathered, *positive* and *negative*.

The results obtained reflect that the number of samples that we collected from each TREC competition differs. While TREC 2003 and TREC 2005 present similar results (17,429 and 18,444 samples each one), TREC 2002 set largely exceeds TREC 2004 (26,655 and 9,454 respectively). The main factor for this difference is the number of judgements included in the judgment set file. This number of judgments depends on three circumstances:

- The number of questions formulated to the systems. For instance, there are 500 questions in TREC 2002, while there are only 230 in TREC 2004.
- The number of competing systems and the number of runs submitted. In 2002 there were 67 runs, while 'only' 54 took part in 2003.
- The convergence of the systems: only different judgements are taken into account. If two systems found the same answer in the same paragraph in the same document, only one sample is obtained.

VI. Conclusions and Future Work

Many natural language applications try to automatically extract information from very large corpora in order to learn linguistic phenomena. Corpus-based approaches have demonstrated to ease the adaptation of systems to new languages and domains. In this paper, we have described the development of a corpus intended to assist every stage of corpus-based QA systems. The corpus was semi-automatically obtained, so that the human effort needed to develop it was minimal. We have focused on English resources as they are much more readily available than for other languages. As our data is fully based on TREC QA track resources, the samples obtained perfectly fit the needs of actual QA systems.

A data collection of 71,982 samples was obtained, which seems large enough to train a corpus-based QA system. Every sample relates a question with its question type, its exact answer, and also provides the sentence, paragraph and document context where this answer occurs. Thus, the corpus is suitable to train on every stage of the QA process, where different contexts are required: question types for question classification stage, exact answers for answer extraction stage, sentences for answer validation stage and documents or paragraphs for information retrieval stage.

Another benefit of our approach is that, unlike other similar corpora, we have not only positive samples but also negative ones, providing the context in which an extracted answer is incorrect. For instance, a binary classifier could be trained on this corpus in order to decide whether a possible answer matches the questions formulated or not.

As future work, we will investigate the use of this corpus together with machine learning techniques in order to build versatile and trainable low cost QA systems.

Acknowledgement

This work has been developed in the framework of the project CICYT R2D2 (TIC2003-07158-C04).

References

[1] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of english: The penn treebank," *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, 1994.

[2] D. Ravichandran, A. Ittycheriah, and S. Roukos, "Automatic derivation of surface text patterns for a maximum entropy based question answering system," in *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 85–87.

[3] R. Soricut and E. Brill, "Automatic question answering using the web: Beyond the factoid," *Information Retrieval*, vol. 9, no. 2, pp. 191–206, 2006.

[4] E. Agichtein, S. Lawrence, and L. Gravano, "Learning search engine specific query transformations for question answering," in *WWW '01: Proceedings of the 10th international conference on World Wide Web*. New York, NY, USA: ACM, 2001, pp. 169–178.

[5] R. D. Burke, K. J. Hammond, V. A. Kulyukin, S. L. Lytinen, N. Tomuro, and S. Schoenberg, "Question answering from frequently asked question files: Experiences with the faq finder system," Chicago, IL, USA, Tech. Rep., 1997.

[6] A. Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal, "Bridging the lexical chasm: statistical approaches to answer-finding," in *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2000, pp. 192–199.

[7] E. Bisbal, D. Tomás, L. Moreno, J. L. Vicedo, and A. Suárez, "A multilingual svm-based question classification system," in *MICAI 2005: Advances in Artificial Intelligence, 4th Mexican International Conference on Artificial Intelligence*, ser. Lecture Notes in Computer Science, A. F. Gelbukh, A. de Albornoz, and H. Terashima-Marín, Eds., vol. 3789. Springer, November 2005, pp. 806–815.

[8] I. Dagan, O. Glickman, and B. Magnini, "Recognizing textual entailment," in *PASCAL Proceedings of the First Challenge Workshop*, Southampton, UK, April 2005, pp. 1–8.

[9] E. M. Voorhees, "The trec-8 question answering track report," in *Eighth Text REtrieval Conference*, ser. NIST Special Publication, vol. 500-246. Gaithersburg, USA: National Institute of Standards and Technology, November 1999, pp. 77–82.

[10] J. C. Reynar and A. Ratnaparkhi, "A maximum entropy approach to identifying sentence boundaries," in *Proceedings of the fifth conference on Applied natural language processing*. Morristown, NJ, USA: Association for Computational Linguistics, 1997, pp. 16–19.

[11] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

Cross Language Information Retrieval using Multilingual Ontology as Translation and Query Expansion Base

Mustafa Abusalah, John Tait, and Michael Oakes

***Abstract*—This paper reports an experiment to evaluate a Cross Language Information Retrieval (CLIR) system that uses a multilingual ontology to improve query translation in the travel domain. The ontology-based approach significantly outperformed the Machine Readable Dictionary translation baseline using Mean Average Precision as a metric in a user-centered experiment.**

***Index terms*—Ontology, multilingual, cross language information retrieval.**

I. Introduction

THE growing requirement on the Internet for users to access information expressed in language other than their own has led to Cross Language Information Retrieval (CLIR) becoming established as a major topic in IR. One approach to CLIR uses different translation approaches to translate queries to documents and indexes in other languages. As queries submitted to search engines suffer lack of context, translation approaches have great problems with resolving query ambiguity. In our approach, we built a multilingual ontology to be used as a translation base for CLIR. In this paper we evaluate our proposed query translation methodology and compare it with a base line system that uses a Machine Readable Dictionary (MRD) as translation base in a user-centered experiment.

II. Background

CLIR approaches are decomposed into two research fields, the first is bilingual MRD and machine translation (MT), and the second is concept driven approaches.

The major problem in the bilingual dictionary approach is translation ambiguity in addition to problems of word inflection, problems of translating word compounds, phrases, proper names, spelling variants and special terms [8], [9], [10]. MT systems normally attempt to determine the correct word sense for translation by using context analysis [11]. However, a typical search engine query lacks context as it consists of a small number of keywords. MT is more efficient in document translation as the context is clearer.

Manuscript received October 14, 2008. Manuscript accepted for publication August 3, 2009.

The authors are with the School of Computing and Technology, University of Sunderland, Sunderland SR6 0DD, UK, (e-mail: mustafa.abusalah@sunderland.ac.uk, john.tait@sunderland.ac.uk, michael.oakes @sunderland.ac.uk).

Concept driven approaches such as thesauri and multilingual ontologies bridge the gap between the linguistic term and its meaning.

A Bilingual Thesaurus groups words with similar meanings in hierarchies (with several levels) of classes and sections and maps them according to their meanings. EuroWordNet is an example of a multilingual thesaurus that uses "is-a" relations (amongst other types of relations) between "synsets", or groups of synonymous words and maps them according to their meanings using a bilingual index. However, the thesaurus does not include the definition of words. In fact, words in a group are merely related, not synonymous. In addition, words under a common heading can be of different syntactic categories. EuroWordNet groups terms of synsets with basic semantic relations between them.

In our approach we considered developing a bilingual ontology rather than collecting a thesaurus, because we consider ontology as a generalized collection of knowledge that will be used to add a context to search queries by the query expansion, enabling word sense disambiguation. Ontology defines concepts, terms and vocabulary in a domain, and also the relationship among these concepts. Concepts are organized in a taxonomic structure, with subclasses inheriting properties and specializing from superclasses. Current semantic web technologies also have the added capability of inferring new facts from old facts already captured in the ontology. An ontology, together with a set of instances of the classes or concepts defined, constitutes a knowledge base about the domain being described [12].

III. Ontology versus MRD

The ontology was built to model the travel domain and decomposed into two ontologies (Arabic and English Ontologies). The ontology was developed manually with the help from a domain expert. Both ontologies are mapped using an English Arabic bilingual index. The manually created ontology consists of 100 English concepts mapped to their Arabic equivalents and it was updated with 100 English concepts mapped automatically to the equivalent Arabic concepts a total of 200 mapped concepts. The automatic ontology mapping process that applied WSD (Word Sense Disambiguation) scored a precision of 0.83 in a user based evaluation. In addition to concept relations, such as "is a" and "has a" relationships, ontology also includes "instance of" and many other relations. Those relationships are represented in ontology languages like owl and rdf constructs. Concept

relations are used to expand queries with semantically related concepts to improve the information retrieval system's monolingual and cross lingual effectiveness. For example "Hotel" is a sort of "Accommodation", so if "hotel" was a query keyword it will be expanded to hotel or accommodation to return more relevant results in monolingual retrieval and referred to its equivalent concepts in Arabic to return more relevant results in Cross Lingual retrieval. In the retrieval system the ontology is combined with an MRD so if the ontology did not succeed in translating concepts, the MRD will translate them, and the translated query will be a combined translation of the ontology and the MRD. The ontology was constructed prior to the experimental query set being identified. It was developed using Protégé as it allows the developers to create, browse and edit domain ontologies in a frame-based representation. In addition plug-ins to enhance ontology development such as the OWL plug-in, were used to develop the OWL ontology. Both ontologies, Arabic and English, are mapped at the semantic level; each concept in both ontologies is mapped to its equivalent concept using a bilingual index defined in the English Ontology. We have developed an automatic ontology mapping tool to define and execute semantic bridges to map both ontologies. Figure 1 demonstrates a simple ontology translation process.

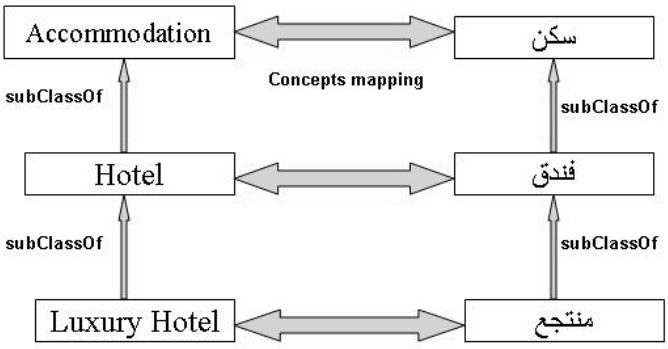


Fig. 1. Simple Concept Matching Task.

As a base for our Information Retrieval system we used the full text search technique. Full text search (also called free search text) refers to a technique for searching corpora; in a full text search, the search engine examines all of the words in every stored document as it tries to match search words supplied by the user. Some Web search engines, such as AltaVista, employ full text search techniques.

In our approach to employ full text search we generated a complete index for all of the searchable documents in the corpora. For each word (excepting stop words which are too common to be useful) an entry is made which lists the exact position of every occurrence of it within the database of documents. From such a list it is relatively simple to retrieve all the documents that match a query.

The MRD is Al-Mawred English Arabic dictionary [1] which has 100,000 English/Arabic entries and 67,000 Arabic/English entries. As noted above, it is used for MRD based CLIR as a baseline and to augment the ontology based translation. The Dictionary based IR system passes each query keyword to the Arabic/English Dictionary and the results are submitted to the search engine. In the dictionary model when a keyword is translated and has many synonyms the first matched synonym is selected. Figure 2 shows CLIR using MRD.

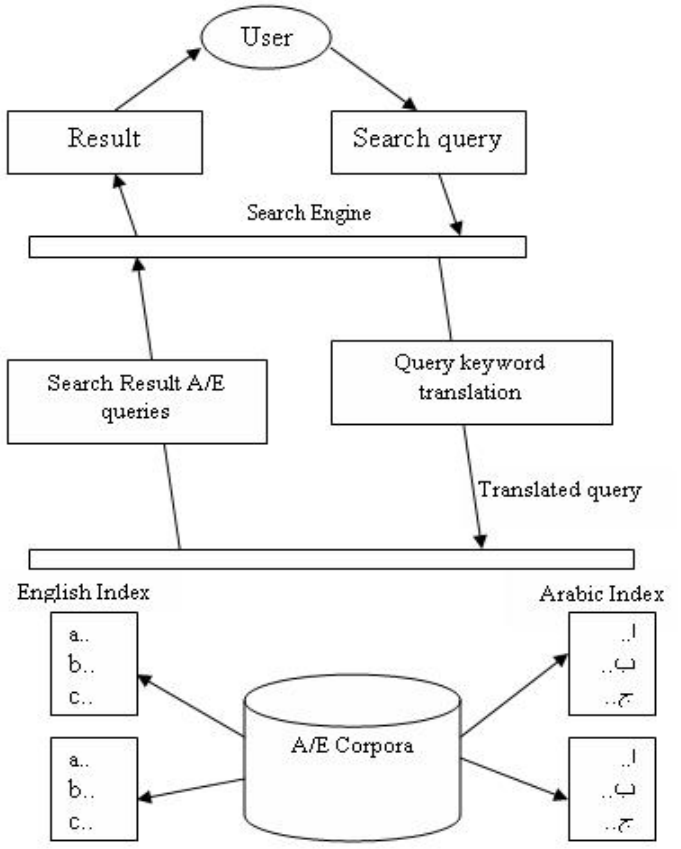


Fig. 2. Shows the CLIR process using MRD.

The Ontology based IR system submits the query keywords to XSL (Extensible style sheet Language) to query the ontologies, extracting related concepts and concept relations. Then concepts associated with semantic relations are studied by the ontology based CLIR system and identified for query expansion if synonyms were found, this is all done monolingual, then concepts are translated into their equivalent concepts in the other language using the ontology bilingual index. If the concept was not found in the ontology, the Dictionary is used to find the relevant translated concepts. Figure 3 shows the CLIR process using ontologies. In both dictionary and ontology based CLIR systems the final translated query terms are combined using the Boolean OR and then matched with the corpora documents. The results then are ranked depending on many factors such as the number of matching terms found in each document and the number of terms occurring in the document. We used the BM25 [13] (Best Match) weighting scheme to rank the found documents. TREC tests have shown BM25 to be the best of the known probabilistic weighting schemes [14].

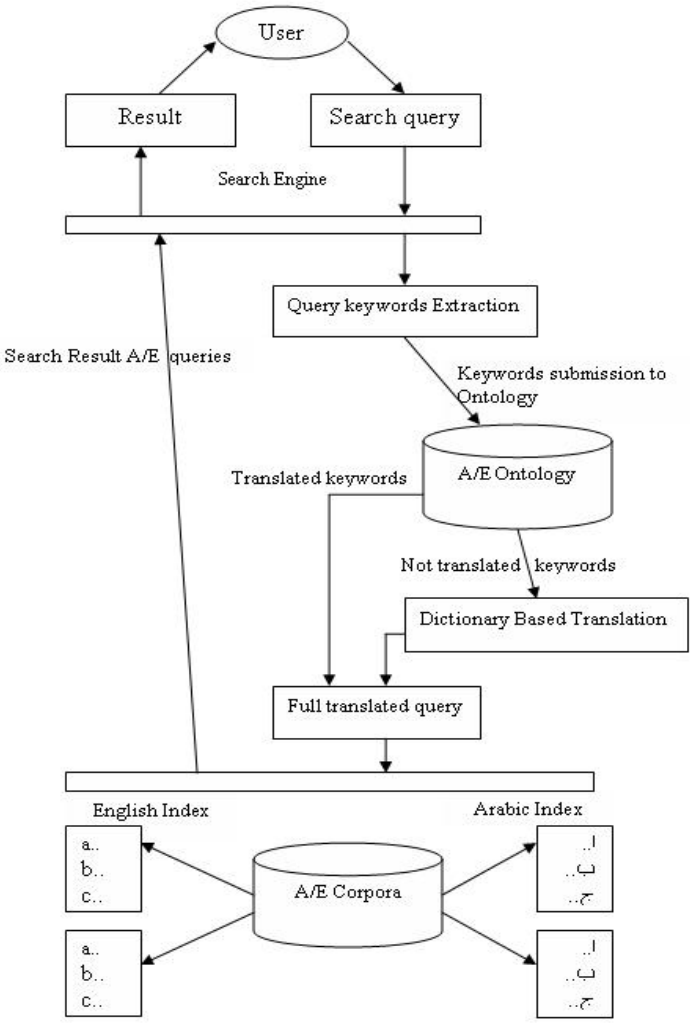


Fig. 3. Shows the CLIR process using ontology.

IV. Evaluations

The evaluation is based on human relevance assessments during experimental search sessions. 25 common queries were identified by discussion with experts in the travel and tourism field as being of interest to potential users of a travel search engine. They were expressed in the English language. The judges who evaluated both systems are not experts in the travel field but they have at least traveled abroad once. All judges are native Arabic speakers and have a very good knowledge of the English language. Twenty judges made a relevance judgment for each query submitted to each system with a total of 1000 judgments for the 25 queries for both systems. The judges access the system using a web-based interface, and submit the queries to both systems. We conducted two experimental runs.

Run 1: The Judge submits the query to the dictionary based system and evaluates the first 40 results appearing on the web browser with title and brief description.

Run 2: The same procedure applied in run 1 is applied in run 2 but using ontology based translation.

The judgment was binary as to whether result was relevant or not [2]. The judges were asked to score the quality of relevance match according to one of the following relevancy scale (not relevant, don't know, possibly relevant, relevant, critically relevant), as shown in figure 4.

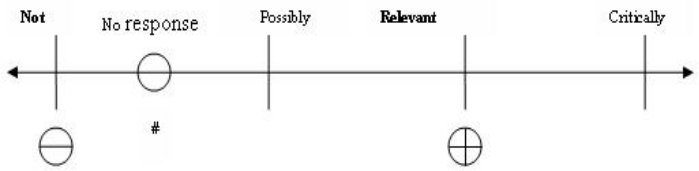


Fig. 4. Relevance Scale.

These responses were mapped onto a binary scale relevant if the document retrieved is at least possibly relevant, otherwise the document is not relevant. For example, critically relevant documents specifies to the user exactly what he is looking for, while possibly relevant might have some useful information, but doesn't specify exactly the user need. If the judge did not know whether the document is relevant or not his judgment is considered not relevant [7]. The document collection used in this experiment is about 8,000 documents in the Arabic language. The documents are all related to the travel domain and either published in Al-Nahar newspaper [3] from the year 1996-1999 or documents collected from the Palestinian ministry of tourism [4]. The scale of the collection, together with only two related systems being used in the experiment, meant no reliable assessment of recall could be made within the available time and resources.

V. Results

After the users' assessment the Mean Average Precision [5] was measured, where Mean Average Precision is the average of the precision after each relevant document is retrieved.

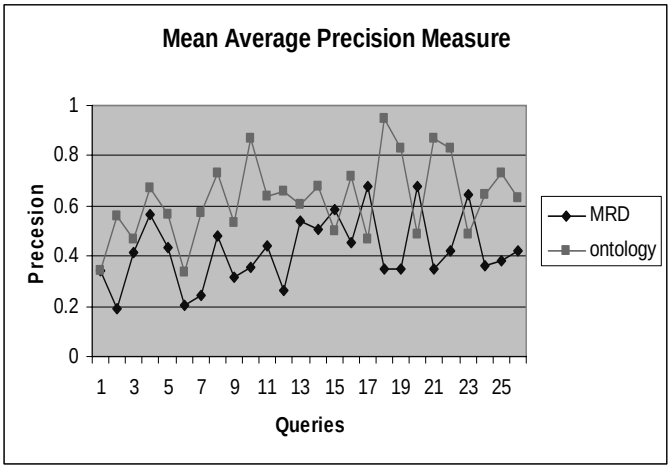


Fig. 5. Ontology versus Dictionary precision measure result.

$$MAP = \frac{\sum_{r=1}^{N}(pr(r) \times rel(r))}{Number\ of Relevant\ \ Documents} \qquad (1)$$

where r is the rank, N is the number retrieved, *rel()* is a binary function on the relevance of a given rank, and *Pr()* is precision at a given cut-off rank.

Figure 5 shows the measurement of mean average precision for both the Dictionary and ontology based CLIR systems. The first run that measured the dictionary based CLIR system scored average MAP result **0.42** while run two that measured the ontology based CLIR system scored average MAP result **0.63** which is much better than the Dictionary based system average MAP result.

VI. Conclusions and Discussion

In this experiment, the effectiveness of the ontology based CLIR was better than the Dictionary based one. The benefit of using ontology is not limited to normal word to word translation. These results are especially interesting because they contrast with early monolingual work (e.g. Voorhees [6]) in which this sort of query expansion degraded rather than improved retrieval effectiveness. It is difficult to determine at this stage whether the improvement is a product of operating in a narrow (and known) domain, the scale and variety of the document collection or some other cause.

After the evaluation of both the pure dictionary and the ontology systems, the ontology based system scored higher in terms of precision. In future development we will enhance and extend the ontology by using annotation tools to align new concepts to the ontology and then test it again with the dictionary system. Other areas for investigation include ease of use, the use of relevance feedback, the effect of more extensive use of concept relations and possibly experiments with larger data sets.

References

[1] D. A. Lelmalayin, http://www.malayin.com/index_e.asp 26-01-2006.

[2] F. C. Gey, "The TEC-2001: Cross Language Information Retrieval Track," 2001.

[3] Evaluations and Language Resources Distribution Agency, http://www.elda.org 26-01-2006.

[4] Palestinian ministry of tourism, http://www.visit-palestine.com/ 26-01-2006.

[5] J. Levelling, "Towards a better baseline for NLP methods in domain-specific information retrieval," in *Results of the CLEF 2005 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2005 Workshop*, Wien, Österreich: Centromedia, 2005.

[6] E. Voorhees, "Query expansion using lexical-semantic relations," in *Proc. of SIGIR*, Dublin, 1994.

[7] P. Buitelaar, D. Steffen, M. Volk, D. Widdows, B. Sacaleanu, Š. Vintar, S. Peters, and H. Uszkoreit, "Evaluation Resources for Concept-based Cross-Lingual Information Retrieval in the Medical Domain," in *Proc. of LREC*, Lissabon, 2004.

[8] L. Ballesteros and B. Croft, "Phrasal Translation and Query Expansion Techniques for Cross-language Information Retrieval," in *Proc. of SIGIR*, 1997, pp. 84-91.

[9] L. Ballesteros and B. Croft, "Resolving Ambiguity for Cross-Language Retrieval," in *Proc. of SIGIR*, 1998, pp. 64-71.

[10] T. Hedlund, E. Airio, H. Keskustalo, R. Lehtokangas, A. Pirkola, and J. Kalervo, "ARVELIN: Dictionary-Based Cross-Language Information Retrieval: Learning Experiences from CLEF 2000–2002," *Information Retrieval*, 7, 99–119, 2004.

[11] M. Braschler, "Combination Approaches for Multilingual Text Retrieval Eurospider Information Technology," *Information Retrieval*, 7, 183–204, 2004.

[12] N. F. Noy and D. L. McGuiness, "Ontology Development 101: A Guide to Creating Your First Ontology," SMI Technical Report SMI-2001-0880, 2001.

[13] S. E. Robertson, S. Walker, M. M. Beaulieu, M. Gatford, and A. Payne, "Okapi at TREC-4," in *NIST Special Publication 500-236, the Fourth Text Retrieval Conference (TREC-4)*, 1995, pp 73-96.

[14] TREC http://trec.nist.gov/ 26-12-2006

English-to-Japanese Cross-Language Question-Answering System using Weighted Adding with Multiple Answers

Masaki Murata, Masao Utiyama, Toshiyuki Kanamaru, and Hitoshi Isahara

***Abstract*—We describe a method of using multiple documents with decreasing weights as evidence to improve the performance of a question-answering system. We also describe how it was used in cross-language question answering (CLQA) tasks. Sometimes, the answer to a question may be found in multiple documents. In such cases, using multiple documents for prediction generates better answers than using a single document. Therefore, our method uses information from multiple documents by adding the scores of candidate answers extracted from the various documents. Because simply adding scores degrades the performance of question-answering systems, we add scores with decreasing weights to reduce the negative effect of simply adding. We used this method in the CLQA part of NTCIR-5. It was incorporated into a commercially available translation system that carries out cross-language question-answering tasks. Our method obtained relatively good CLQA results.**

***Index Terms*—Machine translation, cross-language question-answering, decreased adding, multiple documents, NTCIR.**

I. Introduction

A question-answering system is an application designed to produce the correct answer to a question given as input. For example, when "What is the capital of Japan?" is given as input, a question-answering system may retrieve a document containing a sentence, like "Tokyo is Japan's capital and the country's largest and most important city. Tokyo is also one of Japan's 47 prefectures." from an online text, such as a website, a newspaper article, or an encyclopedia. The system can then output "Tokyo" as the correct answer. We expect question-answering systems to become increasingly important as a more convenient alternative to systems designed for information retrieval and as a basic component of future artificial intelligence systems. Recently, many researchers have been attracted to this important topic. These researchers have produced many interesting studies on question-answering systems [1], [2], [3], [4], [5], [6]. Evaluation conferences or contests on question-answering systems have been held in both the U. S. A. and Japan. In the U. S. A., one evaluation conference was called the Text REtrieval Conference (TREC) [7], while in Japan, another conference was called the Question-Answering Challenge (QAC) [8]. These evaluation conferences aim to improve question-answering systems by having researchers use their question-answering systems to solve the same questions, and then examining each system's performance to glean possible methods of improvement. We investigated the potential of question-answering systems [9] and studied their construction by participating in the QAC [8] at NTCIR workshop [10].

Manuscript received October 19, 2008. Manuscript accepted for publication August 3, 2009.

M. Murata, M. Utiyama, T. Kanamaru, and H. Isahara are with the National Institute of Information and Communications Technology 3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan (e-mail: murata@nict.go.jp).

We proposed a new method that uses multiple documents as evidence but decreases adding to improve performance. Sometimes, the answer to a question may be found in multiple documents. In such cases, question answering systems that use multiple documents for prediction generate better answers than those that use only one document [3], [4], [5], [11]. In our method, information from multiple documents is used by adding the scores for the candidate answers extracted from the various documents [4], [11]. Because simply adding the scores degrades the performance of a question-answering system, our method adds the scores with decreasing weights to overcome the problems of simple adding. More concretely, our method multiplies the score of the i-th candidate answer by a factor of $k^{(i-1)}$ before adding the score to the running total. The final answer is then determined based on the total score. For example, suppose that "Tokyo" is extracted as a candidate answer from three documents and has scores of "26", "21", and "20", and that k is 0.3. In this case, the total score for "Tokyo" is "34.1" (= 26 + 21 $\times$ 0.3 + 20 $\times$ 0.3^2). Thus, we calculate the score in the same way for each candidate and take the answer with the highest score as the correct answer. When this method was used at CLQA (NTCIR-5), it scored higher than most participants' methods.

II. Use of Multiple Documents as Evidence with Decreased Adding

Suppose that the question, "What is the capital of Japan?", is input to a question-answering system, with the goal of obtaining the correct answer, "Tokyo". A typical question-answering system would output the candidate answers and scores listed in Table I. These systems also output a document ID indicating the document from which each candidate answer was extracted.

TABLE I
CANDIDATE ANSWERS WITH ORIGINAL SCORES, WHERE "TOKYO" IS THE CORRECT ANSWER

Rank	Candidate answer	Score	Document ID
1	Kyoto	3.3	926324
2	Tokyo	3.2	259312
3	Tokyo	2.8	451245
4	Tokyo	2.5	371922
5	Tokyo	2.4	221328
6	Beijing	2.3	113127
...	...	...	...

TABLE II
CANDIDATE ANSWERS CHOSEN BY SIMPLE ADDITION WHERE "TOKYO" IS THE CORRECT ANSWER

Rank	Cand. ans.	Score	Document ID
1	Tokyo	10.9	259312, 451245, ...
2	Kyoto	3.3	926324
3	Beijing	2.3	113127
...	...	...	...

TABLE III
CANDIDATE ANSWERS WITH ORIGINAL SCORES, WHERE "KYOTO" IS THE CORRECT ANSWER

Rank	Cand. ans.	Score	Document ID
1	Kyoto	5.4	926324
2	Tokyo	2.1	259312
3	Tokyo	1.8	451245
4	Tokyo	1.5	371922
5	Tokyo	1.4	221328
6	Beijing	1.3	113127
...	...	...	...

TABLE IV
CANDIDATE ANSWERS CHOSEN SIMPLY BY ADDING SCORES WHERE "KYOTO" IS THE CORRECT ANSWER

Rank	Cand. ans.	Score	Document ID
1	Tokyo	6.8	259312, 451245, ...
2	Kyoto	5.4	926324
3	Beijing	1.3	113127
...	...	...	...

For the example shown in Table I, the system outputs an incorrect answer, "Kyoto", as the first answer.

A method based on simple addition of the scores of candidate answers was used previously [4], [11]. For our current example question, this produces the results shown in Table II. In this case, the system outputs the correct answer, "Tokyo", as the first answer. The method can thus obtain correct answers using multiple documents as evidence.

The problem with this method, however, is that it is likely to select candidate answers with high frequencies. This is a serious problem from a performance standpoint. In the case of a system with good inherent performance, the original scores that it outputs are often more reliable than the simple addition scores, so using this method often degrades system performance.

To overcome this problem, we developed our new method of using multiple documents with decreased adding as evidence. Instead of simply adding the scores of the candidate answers, the method adds the scores with decreasing weights. This approach reduces the likelihood that a question-answering system will select candidate answers with high frequencies, while still improving the accuracy of the system by adding the scores.

We can demonstrate the effect of our proposed method with an example. Suppose that a question-answering system outputs Table III in response to the question, "What was the capital of Japan in A.D. 1000?". The correct answer is "Kyoto", and the system outputs the correct answer as the first answer.

When we use a method that simply adds scores in this system, however, we obtain the results shown in Table IV. In this case, the incorrect answer, "Tokyo", scores the highest.

To overcome this problem, we can try to apply our proposed method of adding candidate scores with decreasing weights. Suppose that we implement our method by multiplying the score of the i-th candidate by a factor of $0.3^{(i-1)}$ before adding scores. In this case, the score for "Tokyo" is 2.8 (= $2.1 + 1.8 \times 0.3 + 1.5 \times 0.3^2 + 1.4 \times 0.3^3$) and we obtain the results shown in Table V. The correct answer, "Kyoto", achieves the highest score, while the score for "Tokyo" is notably lower.

We can also apply our method to the first example question, "What is the capital of Japan?". When we use our method, the score for "Tokyo" is 4.3 (= $3.2 + 2.8 \times 0.3 + 2.5 \times 0.3^2 + 2.4 \times 0.3^3$), and we obtain the results shown in Table VI. As expected, "Tokyo" scores the highest.

As shown here, our method of adding scores for candidate answers with decreasing weights successfully obtained the correct answers to each of the example questions. This suggests that the method reduces the likelihood that a question-answering system will select candidate answers with high frequencies, while at the same time improving the system's accuracy.

III. QUESTION-ANSWERING SYSTEMS USED IN THIS STUDY

The system has three basic components:

1) Prediction of answer type
   The system predicts the answer to be a particular type of expression, based on whether the input question is indicated by an interrogative pronoun, an adjective, or an adverb. For example, if the input question is "Who is the prime minister of Japan?", the expression "Who" suggests that the answer will be a person's name.
2) Document retrieval
   The system extracts terms from the input question and retrieves documents using these terms. The retrieval process thus gathers documents that are likely to contain the correct answer. For example, for the input question "Who is the prime minister of Japan?", the system extracts "prime", "minister", and "Japan" as terms and retrieves documents accordingly.
3) Answer detection

TABLE V
CANDIDATE ANSWERS OBTAINED BY DECREASED ADDING, WHERE "KYOTO" IS THE CORRECT ANSWER

Rank	Cand. ans.	Score	Document ID
1	Kyoto	5.4	926324
2	Tokyo	2.8	259312, 451245, ...
3	Beijing	1.3	113127
...	...	...	...

TABLE VI
CANDIDATE ANSWERS OBTAINED BY DECREASED ADDING, WHERE "TOKYO" IS THE CORRECT ANSWER

Rank	Cand. ans.	Score	Document ID
1	Tokyo	4.3	259312, 451245, ...
2	Kyoto	3.3	926324
3	Beijing	2.3	113127
...	...	...	...

The system extracts linguistic expressions that match the predicted expression type, as described above, from the retrieved documents. It then outputs the extracted expressions as candidate answers. For example, for the question "Who is the prime minister of Japan?", the system extracts people's names as candidate answers from documents containing the terms "prime", "minister", and "Japan".

A. Prediction of Answer Type

1) Heuristic rules: The system we used applies manually defined heuristic rules to predict the answer type. There are 39 of these rules. Some of them are listed here:

1) When *dare* "who" occurs in a question, a person's name is given as the answer type.
2) When *itsu* "when" occurs in a question, a time expression is given as the answer type.
3) When *doko* "where" is in a question sentence and the focus word in a question is not *chiiki* (area), *basho* (location), or so on, an organizational expression is given as the answer type.
4) When *doko* "where" is in a question sentence and the focus word in a question is not *kaisha* (company), *soshiki* (organization), or so on, a location expression is given as the answer type.
5) When *doko no kuni* "what country" is in a question sentence, a country expression is given as the answer type.
6) When '*nani* (what) + suffix' is in a question sentence, the suffix is extracted as a unit expression.
7) When *donokurai* "how many" occurs in a question, a numerical expression is given as the answer type. The unit expression is estimated using the following method.

Our system uses a new method, which we call *unit estimation*, to obtain a correct unit expression answer. With this method, we gather sentences containing expressions like "UNIT-FOCUS + *wa* (be) + 'numerical expressions' + 'unit expressions"' and extract the unit expressions. We then eliminate unnecessary unit expressions by applying a statistical test based on a binomial distribution. Eliminated expressions are as follows:

TABLE VII
AN EXAMPLE OF USING UNIT ESTIMATION

e	k	n	P(e)
meetoru (meter)	50	128175	1.000000
senchi (centimeter)	28	47050	1.000000
miri (millimeter)	11	25897	1.000000
kiro (kilometer)	11	99618	0.999996
kounen (light-year)	2	538	1.000000
hun (minute)	2	955808	0.000000
yaado (yard)	1	2744	0.998205
inchi (inch)	1	1865	0.999160
hon (piece)	1	1625073	0.000000
shaku (*shaku*)	1	2146	0.998892

$$\text{Unnecessary expressions} = \{t | P(e) \leq k_p\}, \tag{1}$$

where $P(e)$ is calculated by the following equation and k_p is a constant identified based on experimental results.

$$P(e) = \sum_{r=0}^{k} C(n, r) p(u)^r (1 - p(u))^{n-r}, \tag{2}$$

where $C(x, y)$ is the number of combinations when we select y items from x items, n is the number of times expression e occurs in the corpus, k is the number of times the unit expression e occurs in the pattern of "UNIT-FOCUS + *wa* (be) + 'numerical expressions' + 'unit expressions"' in the corpus, and $p(u)$ is calculated by

$$p(u) = \frac{\text{freq}(u)}{N}, \tag{3}$$

where $\text{freq}(u)$ is the frequency of the UNIT-FOCUS u appearing in the corpus and N is the number of all characters in the corpus. In this study, we used articles from newspapers issued over a 10-year period [12] as the corpus for the unit estimation calculation.

An example of using unit estimation is as follows. Consider the question sentence 'X *no nagasa wa dono kurai desuka?*' (What is the length of X?). In this case, we extract a noun *nagasa* (length) as the UNIT-FOCUS and gather candidate unit expressions using "*nagasa* + *wa* + 'numerical expressions' + 'unit expressions"'. We obtain *meetoru* (meter), *senchi* (centimeter), *miri* (millimeter), *kiro* (kilometer), *kounen*

(light-year), *hun* (minute), *yaado* (yard), *inchi* (inch), *hon* (piece), and *shaku* (a measure unit for a length, equal to about 30.3 cm) as candidates. We calculate $P(e)$ for each candidate and obtain the results shown in Table VII. In this case, N is 533,366,720, the frequency of *nagasa* is 11,887, and $p(u) = \frac{11,887}{533,366,720} = 0.000022289$. As shown in Table VII, our method can correctly eliminate *hun* (minute) and *hon* (piece). When using our unit estimation, we do not need a dictionary for unit expressions. Another valuable feature of unit estimation is that it presents various expressions that appear in the corpus. Unit estimation can also be used to construct a dictionary of unit expressions. Thus, our unit estimation method offers various benefits.

B. Document Retrieval

Our system extracts terms from a question using a morphological analyzer called ChaSen [13]. The analyzer first eliminates terms that are prepositions or similar parts of speech and then retrieves using the extracted terms.

The document retrieval method operates as follows:

We first retrieve the top k_{dr1} documents with the highest scores calculated from the equation

$$Score(d) = \sum_{\substack{\text{term} \\ t}} \left(\frac{tf(d,t)}{tf(d,t) + k_t \frac{length(d) + k_+}{\Delta + k_+}} \times log \frac{N}{df(t)} \right) \tag{4}$$

where d is a document, t is a term extracted from a question, $tf(d,t)$ is the occurrence frequency of t in document d, $df(t)$ is the number of documents in which t appears, N is the total number of documents, $length(d)$ is the length of d, and Δ is the average length of all documents. k_t and k_+ are constants identified based on experimental results. We based this equation on Robertson's equation [14], [15]. This approach is very effective, and we have used it extensively for information retrieval [16], [17], [18]. In the question answering system, we use a large number for k_t.

Next, we re-rank the extracted documents according to the following equation and extract the top k_{dr2} documents, which are used in the ensuing answer extraction phase.

$$\begin{aligned} & Score(d) \\ = \; & -min_{t1 \in T} log \prod_{t2 \in T3} (2dist(t1,t2) \frac{df(t2)}{N})^{w_{dr2}(t2)} \\ = \; & max_{t1 \in T} \sum_{t2 \in T3} w_{dr2}(t2) log \frac{N}{2dist(t1,t2) * df(t2)} \end{aligned} \tag{5}$$

$$T3 = \{t | t \in T, 2dist(t1,t) \frac{df(t)}{N} \leq 1\}, \tag{6}$$

where d is a document, T is the set of terms in the question, and $dist(t1,t2)$ is the distance between $t1$ and $t2$ (defined as the number of characters between them) with $dist(t1,t2)$ = 0.5 when $t1 = t2$. $w_{dr2}(t2)$ is a function of $t2$ that is adjusted based on experimental results.

Because our question-answering system can determine whether terms occur near each other by re-ranking them according to Eq. 5, it can use full-size documents for retrieval. In this study, we extracted 20 documents for retrieval. The following procedure for answer detection is thus applied to the 20 extracted documents.

C. Answer Detection

To detect answers, our system first generates expressions as candidates for the answer from the extracted documents. We initially used morpheme n-grams as candidate expressions, but this approach generated too many candidates. We now use only candidates consisting exclusively of nouns, unknown words, and symbols. Also, we use the ChaSen analyzer to determine morphemes and what parts of speech they are.

Our approach to judging whether each candidate is a correct answer is to add the score ($Score_{near}(c)$) for the candidate, under the condition that it is near an extracted term, and the score ($Score_{sem}(c)$) based on heuristic rules according to the answer type. The system then selects the candidates with the highest total points as correct answers.

We used the following method to calculate the score for a candidate c with the condition that it must be near the extracted terms.

$$\begin{aligned} Score_{near}(c) &= -log \prod_{t2 \in T3} (2dist(c,t2) \frac{df(t2)}{N})^{w_{dr2}(t2)} \\ &= \sum_{t2 \in T3} w_{dr2}(t2) log \frac{N}{2dist(c,t2) * df(t2)} \end{aligned} \tag{7}$$

$$T3 = \{t | t \in T, 2dist(c,t) \frac{df(t)}{N} \leq 1\},$$

where c is a candidate for the correct answer, and $w_{dr2}(t2)$ is a function of $t2$ that is adjusted based on experimental results.

Next, we describe how the score ($Score_{sem}(c)$) is calculated based on heuristic rules for the predicted answer type. We used 45 heuristic rules to award points to candidates and used total points as the score. Some of the heuristic rules are listed below:

1) Add 1000 to candidates when they match one of the predicted answer types (a person's name, a time expression, or a numerical expression). We use named entity extraction techniques based on the support-vector machine method to judge whether a candidate matches

a predicted answer type [19]. We used only five named entities as in our previous system [10].

2) When a country name is one of the predicted answer types, add 1000 to candidates found in our dictionary of countries, which includes the names of almost every country (636 expressions).
3) When the question contains *nani* Noun X "what Noun X", add 1000 to candidates having the Noun X.

Our system has an additional function that is used after answers are selected based on the scores. Our system compiles answers that are part of other answers and whose score is less than 90% of the best score. The system compiles answers by retaining the longest one and eliminating the others. We call this method *rate-based answer compiling*.

IV. How We Handle Cross-Language Question-Answering

We used commercially available translation software to translate questions and documents. We translated the questions into Japanese, to carry out the English-to-Japanese question-answering tasks. In the English-to-Japanese tasks, the questions were written in English and the documents were written in Japanese. We output Japanese answers in response to English queries.

V. Experiments

In this section, we show the experimental results for CLQA of NTCIR-5. Tables VIII to IX show these results. We did one official run (NICT-E-J-01) and two unofficial runs (NICT-E-J-u-01, NICT-E-J-u-02). After the formal run, we made two additional runs (NICT-J-J--01, NICT-J-J--02). We used the decreasing weights method with $k = 0.3$ in NICT-E-J-01, NICT-E-J-u-01, and NICT-J-J--01. We did not use it in NICT-E-J-u-02, and NICT-J-J--02. 200 questions were given for each run. In the tables, "top 1" in the leftmost column indicates that only one answer was evaluated for each question, while "5 ans." indicates that five answers were evaluated for each question, in which case we used the top five answers. "Acc", "MRR", and "Top5" are evaluation metrics. "Acc" indicates the accuracy rate of the first answer. "MRR" indicates a score of $1/r$ when the r-th submitted answer is correct. "Top5" indicates the ratio when one of the top five answers was correct. "*+U" indicates answers that were not supported by a relevant document and were judged to be correct. No "*+U" indicates only the answers that were supported and were judged to be correct. Tables VIII shows the results for the English-to-Japanese question answering tasks. Table IX shows the results for the Japanese-to-Japanese task. The Japanese-to-Japanese task is not relevant to NTCIR-5. We did the experiments with NTCIR-5 to compare the results for English-to-Japanese and for Japanese-to-Japanese tasks.

The experimental results indicate the following.

– The method of weighted adding was effective (compare "NICT-E-J-u-01" and "NICT-E-J-u-02", or "NICT-J-J-u-01" and "NICT-J-J-u-02").). In every case, the accuracy of the method that uses weighted adding was higher than that of methods that do not use weighted adding.
– The Japanese monolingual question-answering tasks were easier than the Japanese-to-English or English-to-Japanese question-answering tasks (compare "NICT-E-J-u-01" and "NICT-J-J--01", and "NICT-E-J-u-02", and "NICT-J-J--02").
– Our cross-language (English-to-Japanese) question-answering obtained about half the accuracy of single-language (Japanese-to-Japanese) question-answering (0.09/0.170 or 0.120/0.265). We found that use of commercial translation software answering in cross-language question answering obtained about half the accuracy of single-language question.

VI. Conclusion

We described a new method of using decreasingly weighted multiple documents as evidence to improve the performance of question-answering systems. Our decreased adding method multiplies the score of the i-th candidate by $k^{(i-1)}$ before adding the score to the running total. We found experimentally that 0.3 were good values for k. Our proposed method is simple and easy to use, and scored much better than methods that did not use decreased adding. These results demonstrate the effectiveness and utility of our method. We used this method for the CLQA part of NTCIR-5. We incorporated it into a commercially available translation system that carries out cross-language question-answering tasks. Our method obtained relatively good results at CLQA.

References

[1] J. Kupiec, "MURAX: A robust linguistic approach for question answering using an on-line encyclopedia," in *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1993.
[2] A. Ittycheriah, M. Franz, W.-J. Zhu, and A. Ratnaparkhi, "IBM's Statistical Question Answering System," in *TREC-9 Proceedings*, 2001.
[3] C. L. A. Clarke, G. V. Cormack, and T. R. Lynam, "Exploiting redundancy in question answering," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2001.
[4] S. Dumis, M. Banko, E. Brill, J. Lin, and A. Ng, "Web question answering: Is more always better?" in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2002.
[5] B. Magnini, M. Negri, R. Prevete, and H. Tanev, "Is it the right answer? Exploiting web redundancy for answer validation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 2002.
[6] D. Moldovan, M. Pasca, S. Harabagiu, and M. Surdeanu, "Performance issues and error analysis in an open-domain question answering system," *ACM Transactions on Information Systems*, vol. 21, no. 2, pp. 133–154, 2003.
[7] TREC-10 committee, "The tenth text retrieval conference," 2001, http://trec.nist.gov/ pubs/trec10/t10_proceedings.html.
[8] National Institute of Informatics, *Proceedings of the Third NTCIR Workshop (QAC)*, 2002.

TABLE VIII
EVALUATION OF JAPANESE ANSWERS IN THE ENGLISH-TO-JAPANESE QUESTION-ANSWERING TASKS

System ID	Acc	MRR	Top5	Acc+U	MRR+U	Top5+U
NICT-E-J-01 (top 1)	0.090	0.090	0.090	0.120	0.120	0.120
NICT-E-J-u-01 (top 1)	0.090	0.090	0.090	0.120	0.120	0.120
NICT-E-J-u-01 (5 ans.)	0.090	0.095	0.105	0.120	0.155	0.210
NICT-E-J-u-02 (top 1)	0.075	0.075	0.075	0.100	0.100	0.100
NICT-E-J-u-02 (5 ans.)	0.075	0.086	0.105	0.100	0.128	0.175

TABLE IX
EVALUATION OF JAPANESE MONOLINGUAL QUESTION-ANSWERING TASKS

System ID	Acc	MRR	Top5	Acc+U	MRR+U	Top5+U
NICT-J-J--01 (top 1)	0.170	0.170	0.170	0.265	0.265	0.265
NICT-J-J--01 (5 ans.)	0.170	0.239	0.370	0.265	0.386	0.605
NICT-J-J--02 (top 1)	0.190	0.190	0.190	0.240	0.240	0.240
NICT-J-J--02 (5 ans.)	0.190	0.261	0.380	0.240	0.362	0.565

[9] M. Murata, M. Utiyama, and H. Isahara, "Question answering system using syntactic information," 1999, http://xxx.lanl.gov/abs/cs.CL/9911006.
[10] ——, "A question-answering system using unit estimation and probabilistic near-terms IR," 2002.
[11] T. Takaki and Y. Eriguchi, "NTT DATA question-answering experiment at the NTCIR-3 QAC," 2002.
[12] Mainichi Publishing, "Mainichi Newspaper 1991-2000," 2000.
[13] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, and M. Asahara, "Japanese morphological analysis system ChaSen version 2.0 manual 2nd edition," 1999.
[14] S. E. Robertson and S. Walker, "Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval," in *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994.
[15] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3," in *TREC-3*, 1994.
[16] M. Murata, K. Uchimoto, H. Ozaku, Q. Ma, M. Utiyama, and H. Isahara, "Japanese probabilistic information retrieval using location and category information," 2000, pp. 81–88.
[17] M. Murata, M. Utiyama, Q. Ma, H. Ozaku, and H. Isahara, "CRL at NTCIR2," 2001, pp. 5–21–5–31.
[18] M. Murata, Q. Ma, and H. Isahara, "High performance information retrieval using many characteristics and many techniques," 2002.
[19] H. Yamada, T. Kudoh, and Y. Matsumoto, "Japanese named entity extraction using support vector machine," *Transactions of Information Processing Society of Japan*, vol. 43, no. 1, pp. 44–53, 2002.

Using Sense Clustering for the Disambiguation of Words

Henry Anaya-Sánchez, Aurora Pons-Porrata, and Rafael Berlanga-Llavori

***Abstract*—Clustering methods have been extensively used in the solution of many Information Processing tasks in order to capture unknown object categories. This paper presents an approach to Word Sense Disambiguation based on clustering. The underlying idea is that the clustering of word senses provides a useful way to discover semantically related senses. We evaluate our proposal regarding both fine- and coarse-grained disambiguation. Experimental results over Senseval-3 all-words, SemCor 2.0 and SemEval-2007 corpora are presented. Promising values of precision and recall are obtained.**

***Index Terms*—Word sense disambiguation, clustering.**

I. Introduction

THE task of Word Sense Disambiguation (WSD) consists of selecting the appropriate sense for a particular contextual occurrence of a polysemous word. This task can be specialized according to the sense definitions. For instance, word sense induction refers to the process of discovering different senses of an ambiguous word without prior information about the inventory of senses [21]. On the other hand, there are two major approaches for the disambiguation when predetermined sense definitions are provided: data-driven (or corpus-based) and knowledge-driven WSD. Data-driven methods are supervised because they require a learning model built from hand-tagged samples to disambiguate words. Instead, knowledge-driven methods exploit word relationships provided by a background knowledge source, avoiding thus the use of samples. Currently, lexical resources like WordNet [14] constitute the referred source in most cases.

WSD can be seen as a categorization problem consisting of assigning a category label (predefined sense) to each word. In this way, data-driven approaches can be regarded as supervised categorization methods, whereas knowledge-driven ones as unsupervised.

Clustering is one of the most accepted unsupervised categorization methods. It has been explicitly used in WSD for two main purposes. The first one consists of clustering textual contexts to represent different senses in corpus-driven WSD (e.g. [17]) and to induce word senses (e.g. [18], [3]). The other purpose has been the clustering of fine-grained word senses into coarse-grained ones for reducing the polysemy degree of words (e.g. [13], [1]). However, clustering has not been used as categorization method for WSD, that is, as a way to identify sets of word senses that are semantically related.

In this paper, we present a knowledge-driven approach to WSD based on sense clustering. Basically, our proposal uses sense clustering to capture the reflected cohesion among the words of a textual unit. More specifically, starting from an initial clustering of all the possible senses for a textual unit, clusters of senses with a high cohesion w.r.t the textual context are selected. The senses belonging to the selected clusters are grouped and selected again until all words are disambiguated.

The rest of the paper is organized as follows. First, Section II presents our proposal for the disambiguation of words. Section III describes some experiments carried out over Senseval-3 all-words, SemCor 2.0 and SemEval coarse-grained corpora. Finally, Section IV is devoted to offer some considerations and future work as conclusions.

II. Word sense clustering

In this section we address the problem of disambiguating a finite set of words $W = \{w_1, \ldots, w_n\}$ w.r.t its textual context T. The underlying idea of sense clustering is that meaningful word senses must be associated by means of a certain complex relation, which is non-relevant for our purposes because we are only interested in the senses it links. Hence, we propose to identify cohesive groups of senses which are assumed to represent different meanings for the set of words W. Finally, those clusters that fit in with the context T contain the suitable senses.

Algorithm 1 shows the general steps of our proposal. In the algorithm, *clustering* represents the basic clustering algorithm which groups word senses and, *filter* denotes the filtering process which selects the clusters that allow the disambiguation of words in W. The filtering process is described in Algorithm 2. Next paragraphs describe in detail the whole process.

a) Topic signatures: In our approach word senses are represented as topic signatures [12]. Thus, for each word sense s we define a vector $\langle t_1 : \sigma_1, \ldots, t_m : \sigma_m \rangle$, where each t_i is a WordNet term highly correlated to s with an association weight σ_i. The set of signature terms for a word sense includes all its WordNet hyponyms, its directly related terms (including coordinated terms) and their filtered and lemmatized glosses.

Manuscript received November 4, 2008. Manuscript accepted for publication August 28, 2009.

Henry Anaya-Sánchez and Aurora Pons-Porrata are with Center for Pattern Recognition and Data Mining, Universidad de Oriente, Santiago de Cuba, Cuba (henry@cepramid.co.cu, aurora@cepramid.co.cu).

Rafael Berlanga-Llavori is with Department of Langauges and Computer Systems, Universitat Jaume I, Castelló, Spain (berlanga@lsi.uji.es).

Algorithm 1 Clustering-based approach for the disambiguation of the set of words W in the textual context T

Input: The finite set of words W and the textual context T.
Output: The disambiguated word senses.
Let S be the set of all senses of words in W, and $i = 0$;
repeat
$i = i + 1$
G = *clustering*$(S, \beta_0(i))$
G' = *filter*(G, W, T)
$S = \bigcup_{g \in G'} \{s | s \in g\}$
until $|S| = |W|$ or $\beta_0(i+1) = 1$
return S

Algorithm 2 Definition of the filtering process

Input: The set of clusters G, the finite set of words W and the textual context T.
Output: The set of selected clusters G'.
for all g in G **do**
$scores(g) = compare(g, T)$
end for
Sort all groups in G by using the lexicographic order of its scores
Let Q be an empty queue, and G' an empty set
for all g in G **do**
if $\exists(s \in g) \forall(g' \in G')[words(\{s\}) \cap words(g') = \emptyset \quad \wedge$
$\forall(s' \in g)[words(\{s'\}) \subseteq words(g') \implies s' \in \bigcup_{g'' \in G'} g'']]$
then
G'=$G' \cup \{g\}$
else if $\neg\exists(s \in g) \forall(g' \in G')[words(\{s\}) \cap words(g') = \emptyset]$
then
Discard g
else
$Q.insert(g)$
end if
end for
while $words(\bigcup_{g' \in G'} g') \neq W$ **do**
g=$Q.front\_element$
G'=$G' \cup \{g\}$
$Q.remove\_front\_element()$
end while
return G'

To weight signature terms, the tf-idf statistics is used, considering each word as a collection and its senses as its documents. Notice that topic signatures form a Vector Space Model similar to those defined in Information Retrieval Systems. In this way, topic signatures can be compared with usual Information Retrieval measures such as cosine, Dice and Jaccard [19].

b) Clustering algorithm: Clustering is carried out by using the Extended Star Clustering Algorithm [7], which builds star-shaped and overlapped clusters. Each cluster consists of a star and its satellites, where the star is the sense with the highest connectivity of the cluster, and the satellites are those senses connected with the star. The connectivity is defined in terms of the β_0-similarity graph, which is obtained using the cosine similarity measure between topic signatures and the minimum similarity threshold β_0. The way this clustering algorithm relates word senses resembles the manner in which syntactic and discourse relations link textual elements.

c) Cluster filtering: Once clustering is performed over all possible word senses from W, a set of sense clusters is obtained. As some clusters can be more appropriate to describe the semantics of W than others, they are ranked according to a measure w.r.t the intended textual context T. This process can be seen as a context-driven filtering of word senses.

As we represent the context T in the same vector space that the topic signatures of senses, the following function can be used to score a cluster of senses g regarding T:

$$compare(g, T) = \left(|words(g)|, \frac{\sum_i \min(\bar{g}_i, T_i)}{\min(\sum_i \bar{g}_i, \sum_i T_i)}, -\sum_{s \in g} \text{nth}(s) \right)$$

where $words(g)$ denotes the set of words having senses in g, $\bar{g}$ is the centroid of g (computed as the barycenter of the cluster), and $\text{nth}(s)$ is the WordNet number of the sense s according to its corresponding word.

This function scores each cluster considering three measures: the number of words it has associated, its overlapping w.r.t the context and the WordNet sense frequency of its senses respectively. Therefore, we rank all clusters by using the lexicographic order of their scores w.r.t. this function.

Once the clusters have been ranked, they are orderly processed to select clusters for covering the words in W. A cluster g is selected if it contains at least one sense of an uncovered word and other senses corresponding to covered words are included in the current selected clusters. If g does not contain any sense of uncovered words it is discarded. Otherwise, g is inserted into a queue Q. Finally, if the selected clusters do not cover W, clusters in Q adding senses of uncovered words are chosen until all words are covered.

d) Disambiguation process: As a result of the filtering process, a set of senses for all the words in W is obtained (i.e. the union of all the selected clusters). Each word in W that only has a sense in such a set is considered disambiguated. If some word still remains ambiguous, we must refine the clustering process to get stronger cohesive clusters of senses. In this case, all the senses obtained in the previous step must be clustered again but raising the β_0 threshold. Notice that this process must be done iteratively until either all words are disambiguated or when it is not possible to raise β_0 no more. The following equation states how β_0 is set up at each iteration (i-th iteration):

$$\beta_0(i) = \begin{cases} \text{pth}(90, sim(S)) & \text{if } i = 1, \\ \min\limits_{q \in \{90,95,100\}} \{\beta = \text{pth}(q, sim(S)) | \beta > \beta_0(i-1)\} & \text{otherwise.} \end{cases}$$

In this equation, S is the set of current senses, and $\text{pth}(p, sim(S))$ represents the p-th percentile value of the pairwise similarities between senses (i.e. $sim(S) = \{cos(s_i, s_j) | s_i, s_j \in S, i \neq j\} \cup \{1\}$).

runner # 1 = {<criminal,1.056>, <outlaw,1.055>, <ilegal,1.006>, <contrabandist,1.006>, ...}
runner # 2 = {<travel,1.056>, <carrier,0.930>, <arrive,0.930>, <distant,0.772>, <tourist,0.772>, ...}
runner # 3 = {<deliver,1.037>, <boy,1.006>, <announce,0.936>, <dispatch,0.772>, <message,0.718>, ...}
runner # 4 = {<bat,1.055>, <pitcher,1.037>, <base_runner,1.006>, <hit,0.930>, <manager,0.772>, ...}
runner # 5 = {<plant,1.056>, <fungus,1.005>, <structure,1.054>, <branch,1.037>, <foliage,0.930>, ...}
runner # 6 = {<race,1.056>, <olympic,1.049>, <trained,1.037>, <marathon,0.930>, <gold,0.772>, ...}
runner # 7 = {<carpet,1.056>, <covering,1.055>, <include,0.930>, <color,0.930>, <thick,0.930>, ...}
runner # 8 = {<device,1.056>, <light,1.055>, <instrument,1.055>, <metal,1.055>, <machine,1.037>, ...}
runner # 9 = {<atlantic,1.049>, <western,1.049>, <cape,1.006>, <vertebrate,1.006>, <tropical,1.006>, ...}

win # 1 = {<contest,0.654>, <gold,0.587>, <medal,0.587>, <contend,0.487>, <contestant,0.487>, ...}
win # 2 = {<acquire,0.66>, <receive,0.665>, <earn,0.662>, <possession,0.662>, <get,0.635>, ...}
win # 3 = {<score,0.587>, <advance,0.587>, <gain_ground,0.587>, <get_ahead,0.587>, ...}
win # 4 = {<goal,0.662>, <attempt,0.654>, <achieve,0.635>, <attain,0.635>, <reach,0.635>, ...}

marathon # 1 = {<task,0.518>, <endurance_contest,0.503>, <arduous,0.503>, <labor,0.465>, ...}
marathon # 2 = {<race,0.528>, <footrace,0.528>, <mile,0.503>, <yard,0.503>, <steeplechase,0.386>, ...}
marathon # 3 = {<battle,0.528>, <defeat,0.528>, <force,0.528>, <army,0.528>, <troop,0.528>, ...}

Fig. 1. Portion of the representation of senses.

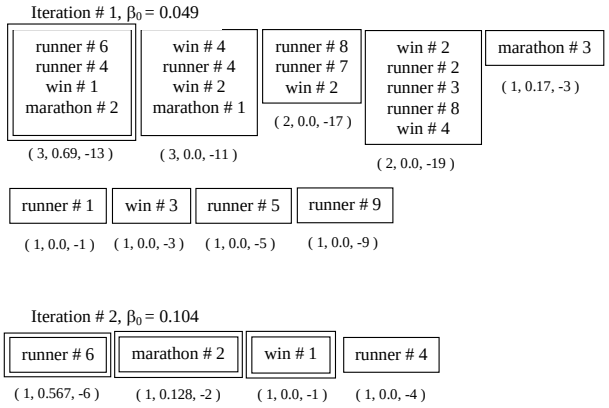


Fig. 2. Disambiguation of words in "*The runner won the marathon*".

A. An example

In this subsection we illustrate the use of our proposal in the disambiguation of the content words appearing in the sentence "*The runner won the marathon*". In this example, the set of disambiguating words W includes the nouns *runner* and *marathon*, and the verb *win* (lemma of the verbal form *won*). Also, in this case we consider that the context T is defined as the vector representation of the filtered and lemmatized sentence, i.e. $T = \langle runner : 1, win : 1, marathon : 1\rangle$. The rest of words are not considered because they are meaningless. As we use WordNet 2.0, we regard that the correct senses for the context are $runner\#6$, $win\#1$ and $marathon\#2$. In Figure 1, an extract of the representation of all word senses is shown.

Figure 2 graphically depicts the disambiguation process carried out by our method in the disambiguation of word senses. The boxes in the figure represent the obtained clusters, which are sorted regarding the lexicographic order given by the function $compare$ (scores are under the boxes).

Initially, the set of all word senses is clustered using the initial β_0=0.0498 (the 90th-percentile of the pairwise similarities between the senses). It can be seen that the first cluster comprises the sense $runner\#6$ (the star), which is the sense refering to a trained athlete who competes in foot races, and $runner\#4$, which is the other sense of *runner* related with the sports. Also, it includes the sense $win\#1$ that concerns the victory in a race or competition, and $marathon\#2$ that refers to a footrace. It can be easily appreciated that this first cluster includes senses that cover the set of disambiguating words. Hence, it is selected by the filter and all other clusters are discarded. After this step, S is updated with the set $\{runner\#6, runner\#4, win\#1, marathon\#2\}$. [1]

In this point of the process, the senses of S do not disambiguate W because the noun *runner* has two senses in S. Also, the next value for the threshold is $\beta_0(2) = 0.1043$. Therefore, the disambiguation of words does not hold because neither $|S| = |W|$ nor $\beta_0(i+1) = 1$. Consequently, a new cluster distribution must be obtained using the current set S.

[1] In the figure, doubly-boxed clusters depict the selected ones by the filter.

The set of boxes in the bottom of Figure 2 represents the new clusters. In this case, all clusters are singles. Obviously, the cluster containing the sense $runner\#4$ is discarded because the cluster that includes the sense $runner\#6$ overlaps better with the context T, and therefore precedes him in the order.

Then, the set of current senses becomes $S = \{ runner\#6, win\#1, marathon\#2\}$, which includes only one sense for each word in W, and thereby the disambiguation holds and the process is stopped. Finally, the current set S is returned as the set of senses that disambiguates the verb *win*, and the nouns *runner* and *marathon*.

III. Experimental results

In order to evaluate our approach, we consider the disambiguation at two different levels of sense granularity. A fine-grained disambiguation was evaluated by using both a subset of SemCor 2.0 composed by all the documents of *brown1* and *brown2*, and a version of Senseval-3 all-words corpus (annotated with WordNet 2.0). In contrast, we use the corpus provided by Task 7 of SemEval-2007 [16] to evaluate the performance of our approach in a coarse-grained WSD.

As evaluation measures, we use the well-known *Precision*, *Recall* and *Coverage*. In the fine-grained case we use their respective "Without U" versions (defined as in Senseval-3 [20]), because there are some word senses in the corpora that are not covered by WordNet 2.0.

In both cases, the disambiguation is performed at the sentence level, i.e., we assume that there is just one correct meaning per word in each sentence. Also, each context T is defined as the vector representation (regarding all lemmatized words) of the sentence.

A. Fine-grained WSD

In this case, we carry out two kinds of experiments. In the first one, we disambiguate all words of each sentence (i.e., W is the set of all meaningful words of the sentence), whereas in the second one we only disambiguate nouns (the set W only

TABLE I
WSD PERFORMANCE OVER THE SENSEVAL-3 ALL-WORDS CORPUS.

Experiment	Category	Instances	Untagged	Precision	Recall	Coverage
	Noun	951	25	0.475	0.462	97.3 %
	Verb	751	3	0.285	0.284	99.6 %
All-words	Adjective	364	11	0.610	0.592	96.9 %
	Adverb	15	0	0.933	0.933	100 %
	All	2081	39	0.432	0.424	98.1 %
Only nouns	Noun	951	25	0.490	0.477	97.3 %

TABLE II
WSD PERFORMANCE OVER THE SEMCOR 2.0 CORPUS.

Experiment	Category	Instances	Untagged	Precision	Recall	Coverage
	Noun	88058	105	0.536	0.535	99.8 %
	Verb	48328	154	0.291	0.290	99.6 %
All-words	Adjective	35664	408	0.626	0.619	98.8 %
	Adverb	20589	837	0.623	0.598	95.9 %
	All	192639	1504	0.500	0.496	99.2 %
Only nouns	Noun	88058	105	0.542	0.541	99.8 %

contains the nouns of the sentence). We will refer to these kind of experiments as "All-words" and "Only nouns" respectively.

Table I summarizes the results obtained over the Senseval-3 all-words corpus. The third column contains the total number of disambiguating word occurrences, and the fourth column shows the number of untagged word occurrences in the corpus, i.e. word occurrences that do not have a WordNet 2.0 sense.

It is worth mentioning that the official Senseval-3 results (reported in [20]) are obtained using a version of Senseval-3 all-words corpus that has been annotated with WordNet 1.7.1. Therefore, our results can not be directly compared with them. However, unlike most participants in Senseval-3 contest, our method obtains a 100 % of coverage if untagged words are ignored.

As we can see, the best performance is obtained in the disambiguation of adverbs and adjectives, while the worst is achieved by the verbs. It can be explained by the high polysemy degree of verbs and its relatively small number of relations in WordNet. Also, it can be appreciated that disambiguating only nouns produces slightly better results than disambiguating nouns together with other words.

The results obtained by our method over the SemCor 2.0 corpus are summarized in Table II. As we can see, they are in agreement with those obtained for the Senseval-3 corpus.

In order to have a better understanding of the behaviour of the algorithm over different knowledge domains, Table III summarizes the overall precision, recall and coverage split according to the SemCor categories.

As shown in Table III, our algorithm performs the best in *Press: reportage* category. In all other categories the recall values are similar. Thus, it seems that the performance is not affected with different knowledge domains.

Finally, we compare our method with four knowledge-driven WSD algorithms: Conceptual density [2], UNED method [6], the Lesk method [11] and the Specification marks with voting heuristics [15]. Table IV includes the recall values obtained over the whole SemCor corpus considering only polysemous nouns.

TABLE III
"ALL WORDS" WSD PERFORMANCE OVER THE SEMCOR CATEGORIES.

Categories	Precision	Recall	Coverage
A. Press: reportage	0.554	0.551	99.4 %
B. Press: editorial	0.520	0.518	99.5 %
C. Press: reportage	0.508	0.505	99.3 %
D. Religion	0.492	0.491	99.7 %
E. Skill & Hobbies	0.499	0.496	99.4 %
F. Popular lore	0.510	0.507	99.3 %
G. Belles letters, biography, essays	0.489	0.487	99.6 %
H. Miscellaneous	0.528	0.525	99.4 %
J. Learned	0.513	0.511	99.6 %
K. General fiction	0.472	0.468	99.0 %
L. Mystery & detective fiction	0.498	0.489	98.1 %
M. Science fiction	0.500	0.495	98.9 %
N. Adventure & western fiction	0.470	0.462	98.3 %
P. Romance & love story	0.461	0.451	97.8 %
R. Humor	0.497	0.490	98.5 %
Brown 1	0.502	0.499	99.3 %
Brown 2	0.497	0.493	99.0 %
Whole SemCor	0.500	0.496	99.2 %

TABLE IV
COMPARISON WITH OTHER METHODS OVER SEMCOR CORPUS.

WSD method	Recall
Conceptual density	0.220
Lesk	0.274
UNED method	0.313
Specification marks	0.391
Our method using SemCor 1.6	**0.472**
Our method using SemCor 2.0	**0.426**

In this case, we experiment with two versions of the SemCor corpus: SemCor 1.6 and SemCor 2.0, and obviously with their corresponding versions of WordNet. It is due to two reasons. The first one is that the results of the other algorithms are obtained using SemCor 1.6. The other reason consists of showing the impact in the disambiguation of the higher polysemy degree of WordNet 2.0 w.r.t. WordNet 1.6. As it can be appreciated, our approach improves all other methods considering both versions of WordNet.

B. Coarse-grained WSD

As the sense inventory corresponding to the coarse-grained English all-words task of SemEval-2007 consists of clusters of WordNet 2.1 senses, we proceed in the same way as with the fine-grained case. That is, we disambiguate each set of words from a sentence w.r.t. WordNet 2.1. However, we use the coarse-grained score provided by the task organizers to evaluate our approach.

In Table V, we show the performance of our method in the coarse-grained English all-word tast of SemEval-2007. In this table we have ommited the values of *Precision* and *Coverage* because all words are disambiguated by the algorithm, i.e. *Precision* values coincide with *Recall* and a 100% of *Coverage* is achieved.

TABLE V
WSD PERFORMANCE IN TASK 7 OF SEMEVAL-2007.

Word Category	Instances	Recall
Noun	1108	0.708
Verb	591	0.626
Adjective	362	0.787
Adverb	208	0.740
All	2269	0.702

TABLE VI
OVERALL COARSE-GRAINED PERFORMANCE.

System	F1
UPV-WSD [4]	0.786
Our method	0.702
RACAI-SYNWSD [9]	0.657
SUSSX-FR [10]	0.604
UOFL [5]	0.506
SUSSX-C-WD [10]	0.459
SUSSX-CR [10]	0.457
MFS *baseline*	0.788

As it can be appreciated, like in the fine-grained experiments the category of verbs significantly perform the worst. Also, the other word categories increase their scores w.r.t the fine grained case because of the relaxation of this new task.

In order to contextualize our results in the current State-of-the-Art, we show in Table VI a comparison between our results and those obtained by other unsupervised systems that participated in SemEval-2007 along with the Most Frequent Sense (MFS) baseline. Systems are ranked according to their F1 score (harmonic mean between *Precision* and *Recall*).

As it can be appreciated, our method obtains the second highest score, which constitutes a good result. It is worth mentioning that unlike most other methods, our proposal does not use any external resource except WordNet, neither the coarse-grained sense inventory provided by the task organizers. Also, it is not used the MFS backoff strategy.

IV. CONCLUSION

In this paper a new approach for the disambiguation of words has been proposed. Its novelty relies on the use of clustering as a natural way to connect semantically related word senses.

Most existing approaches attempt to disambiguate a target word in the context of its surrounding words using a particular taxonomical relation. Instead, we disambiguate a set of related words at once using a given textual context. Besides, we use a sense representation that overcomes the sparseness of WordNet relations, and that relates semantically word senses.

Our proposal relies on both topic signatures built from WordNet and the Extended Star clustering algorithm. The way this clustering algorithm relates sense representations resembles the manner in which syntactic or discourse relations link textual components.

We evaluate the proposed method according to both fine- and coarse-grained disambiguation. In the experiments carried out over Senseval-3 all-words, Semcor 2.0, and SemEval-2007 coarse-grained corpora, promising results were obtained. Our proposal achieves better recall values than other knowledge-driven disambiguation methods over the whole SemCor corpus in the disambiguation of nouns, and performs very well in the SemEval-2007 coarse-grained disambiguation task.

As further work, we plan to experiment with other levels of disambiguation such as phrases and simple sentences to explore its impact in the disambiguation task.

REFERENCES

[1] E. Agirre and O. López, "Clustering wordnet word senses," in *Proceedings of the Conference on Recent Advances on Natural Language Processing*, Bulgary, 2003, pp. 121–130.

[2] E. Agirre and G. Rigau, "Word Sense Disambiguation Using Conceptual Density," in *Proceedings of the 16th Conference on Computational Linguistic*, Vol. 1, Denmark, 1996, pp. 16–22.

[3] S. Bordag, "Word Sense Induction: Triplet-Based Clustering and Automatic Evaluation," in *11st Conference of the European Chapter of the Association for Computational Linguistic*, Italy, 2006.

[4] D. Buscaldi and P. Rosso, "UPV-WSD: Combining different WSD Methods by means of Fuzzy Borda Voting," in *Proceedings of the 4th International Workshop on Semantic Evaluations*, Association for Computational Linguistic, Prage, 2007, pp. 434–437.

[5] Y. Chali and S. R. Joty, "UofL: Word Sense Disambiguation Using Lexical Cohesion," in *Proceedings of the 4th International Workshop on Semantic Evaluations*, Association for Computational Linguistic, Prage, 2007, pp. 476–479.

[6] D. Fernández-Amorós, J. Gonzalo and F. Verdejo, "The Role of Conceptual Relations in Word Sense Disambiguation," in *Proceedings of the 6th International Workshop on Applications of Natural Language for Information Systems*, Spain, 2001, pp. 87–98.

[7] R. Gil-García, J. M. Badia-Contelles and A. Pons-Porrata, "Extended Star Clustering Algorithm," *Progress in Pattern Recognition, Speech and Image Analysis*, Lecture Notes on Computer Sciences, Vol. 2905, Springer-Verlag, 2003, pp. 480-487.

[8] N. Ide and J. Veronis, "Word Sense Disambiguation: The State of the Art," *Computational Linguistics* 24:1, 1998, pp. 1–40.

[9] R. Ion and D. Tufis, "RACAI: Meaning Affinity Models," in *Proceedings of the 4th International Workshop on Semantic Evaluations*, Association for Computational Linguistic, Prage, 2007, pp. 277–281.

[10] R. Koeling and D. McCarthy, "Sussx: WSD using Automatically Acquired Predominant Senses," in *Proceedings of the 4th International Workshop on Semantic Evaluations*, Association for Computational Linguistic, Prage, 2007, pp. 314–317.

[11] M. Lesk, "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone," in *Proceedings of the 5th Annual International Conference on Systems Documentation*, Canada, 1986, pp. 24–26.

[12] C.-Y. Lin and E. Hovy, "The Automated Acquisition of Topic Signatures for Text Summarization," in *Proceedings of the COLING Conference*, France, 2000, pp. 495–501.

[13] R. Mihalcea and D.I. Moldovan, "EZ. WordNet: Principles for Automatic Generation of a Coarse Grained WordNet," in *Proceedings of the FLAIRS Conference*, Florida, 2001, pp. 454–458.

[14] G. Miller, "WordNet: A Lexical Database for English," *Communications of the ACM* 38:11, 1995, pp. 39–41.

[15] A. Montoyo, A. Suárez, G. Rigau and M. Palomar, "Combining Knowledge- and Corpus-based Word-Sense-Disambiguation Methods," *Journal of Artificial Intelligence Research* 23, 2005, pp. 299–330.

[16] R. Navigli, K.C. Litkowski and O. Hargraves, "SemEval-2007 Task 07: Coarse-Grained English All-Words Task," in *Proceedings of the 4th International Workshop on Semantic Evaluations*, Association for Computational Linguistic, Prage, 2007, pp. 30–35.

[17] C. Niu, W. Li, R. K. Srihari, H. Li and L. Crist, "Context Clustering for Word Sense Disambiguation Based on Modeling Pairwise Context Similarities," in *SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Spain, 2004, pp. 187–190.

[18] T. Pedersen, A. Purandare and A. Kulkarni, "Name Discrimination by Clustering Similar Contexts," in *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing*, Mexico, 2005, pp. 226–237.

[19] G. Salton, A. Wong and C. S. Yang, "A Vector Space Model for Information Retrieval," *Journal of the American Society for Information Science* 18:11, 1975, pp. 613–620.

[20] B. Snyder and M. Palmer, "The English all-words task," in *Proceedings of the third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Spain, 2004, pp. 41–43.

[21] G. Udani, S. Dave, A. Davis and T. Sibley, "Noun Sense Induction Using Web Search Results," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Brazil, 2005, pp. 657–658.

Improving Named Entity Extraction Accuracy using Unlabeled Data and Several Extractors

Tomoya Iwakura and Seishi Okamoto

***Abstract*—This paper proposes feature augmentation methods using unlabeled data and several Named Entity (NE) extractors. We collect NE-related information of each word (which we call NE-related labels) from unlabeled data by using NE extractors. NE-related labels which we collect include candidate NE class labels of each word and NE class labels of co-occurring words. To accurately collect the NE-related labels from unlabeled data, we consider methods to collect NE-related labels by using outputs of several NE extractors. We use NE-related labels as additional features for creating new NE extractors. We apply our NE extraction methods using the NE-related labels to IREX Japanese NE extraction task. The experimental results show better accuracy than the previous results obtained with NE extractors using handcrafted resources.**

***Index Terms*—Named entity recognition, unlabeled data, combination of extractors.**

I. Introduction

NAMED Entity (NE) extraction is one of the basic technologies used in text processing like information extraction and question answering. NE extraction aims to extract proper nouns and numerical expressions in text, such as persons, locations, organizations, dates, times, and so on.

To implement NE extractors, the following approaches are mainly used. The first approach is handcrafted-rule based NE extractions [1]. The others are machine learning ones, such as unsupervised learning, semi-supervised learning and supervised learning.

In those machine leaning based methods, supervised learning is widely researched recently [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12]. Furthermore, supervised learning based NE extractors have shown state-of-the-art performance [5], [6], [7], [9].

In the supervised learning based approaches, information obtained with handcrafted lexical resources are often used as features for creating more accurate NE extractors. The handcrafted lexical resources include gazetteers, proper noun dictionaries, thesaurus, and so on. Experimental results obtained with NE extractors using such handcrafted resources have shown better performances than results obtained with NE extractors without using them [3], [5], [6], [7], [8], [9]. However, the construction of the lexical resources is very time-consuming.

Manuscript received November 4, 2008. Manuscript accepted for publication August 25, 2009.

Tomoya Iwakura and Seishi Okamoto are with Fujitsu Laboratories Ltd., 1-1, Kamikodanaka 4-chome, Nakahara-ku, Kawasaki 211-8588, Japan ({iwakura.tomoya,seishi}@jp.fujitsu.com)

To avoid or alleviate the construction of training data and external lexical resources, a number of approaches that exploit unlabeled data have been proposed. For example, various bootstrapping methods for finding rules and dictionaries from unlabeled data have been proposed [13], [2], [14].

Another approach is semi-supervised learning, which uses labeled and unlabeled data for training. For example, the following approaches have been proposed and shown to improve performance; co-training [15] automatically bootstraps labels, Expectation Maximization (EM) which has a theoretical base of likelihood maximization of incomplete data, and structural learning [12] which seeks shared predictive structures jointly learning multiple classification problems.

Feature augmentation methods using unlabeled data have been also applied to Natural Language Processing (NLP) tasks [16], [10], [11]. These techniques use word clusters created from unlabeled data by clustering algorithms as features. These experimental results by using word clusters as features have shown improving performance of supervised learning based NE extractors [10], [11].

This paper proposes feature augmentation methods for NE extractions. Compared to the previous works, our methods use candidate NE classes of words and candidate class labels of co-occurring words (NE-related labels which we call) as features. We collect these NE-related labels from unlabeled data. Furthermore, to collect NE-related labels accurately, our methods use outputs of several NE extractors. We use NE-related labels as new features for creating NE extractors.

This paper is organized as follows. In section II, we describe our Japanese NE extraction method. We describe our collection methods of NE-related labels in section III and report the experimental results on the IREX [17] Japanese NE task in section IV. Finally, we conclude the paper in section V.

II. Japanese Named Entity Extraction

A problem of Japanese NE extraction is that each Japanese NE consists of one or more words, or includes a substring of a word. In this section, we present the chunk representations for word chunks that become NEs at the first. Next, we describe our Japanese NE extraction methods by combining the following methods; NE extraction by word-unit chunking for extracting each NE that consists one or more words, and NE extraction by character-unit chunking for extracting each NE that includes a substring of a word. Finally, we describe

TABLE I
NE EXAMPLES DEFINED BY IREX COMMITTEE

ARTIFACT	LOCATION	ORGANIZATION	PERSON
Nobel Prize in Chemistry	Japan	the Ministry of Foreign Affairs	Tarou Yamada

DATE	MONEY	PERCENT	TIME
May	100 JPY	100%	10:00 a.m.

	田中 *(Tanaka)*	使節 *(mission)*	団 *(party)*	は *(particle)*	日 *(Japan)*	米 *(U.S.A)*	間 *(between)*
IOB1	I-ORG	I-ORG	I-ORG	O	I-LOC	B-LOC	O
IOB2	B-ORG	I-ORG	I-ORG	O	B-LOC	B-LOC	O
IOE1	I-ORG	I-ORG	I-ORG	O	E-LOC	I-LOC	O
IOE2	I-ORG	I-ORG	E-ORG	O	E-LOC	E-LOC	O
SE	B-ORG	I-ORG	E-ORG	O	S-LOC	S-LOC	O

Tanaka mission party dose ... between Japan and U.S.A.

Fig. 1. Examples of NE labels

Support Vector Machines (SVMs) that is the machine learning algorithm for crating NE extractors in our experiments.

A. Chunk Representation for Named Entity Extraction

We use IREX Japanese NE extraction task [17] to evaluate our methods. Table I lists the eight NE classes defined by IREX committee. One of the problems for extracting NEs is that each NE consists of one or more words. To extract NEs, we have to identify word chunks with their NE classes.

To identify word chunks, we use methods to annotate chunk tags to words. We use the following five chunk tag sets; IOB1 [18], IOB2, IOE1, IOE2 [19] and Start/End (SE) [3].

- IOB1: This representation uses three tags which are I, O and B, to represent the inside, outside and beginning of a chunk. B is only used at the beginning of a chunk which immediately follows another chunk. O is used for the outside of any chunk in all the chunk representations.
- IOB2: This representation uses the same three tags outlined for IOB1. However, B tag has a different meaning. B tag is given for the word at the beginning of a chunk.
- IOE1: This representation uses three tags which I, O and E, to represent the inside, outside and end of a chunk. E tag is used to mark the last word of a chunk immediately preceding another chunk.
- IOE2: This representation uses the same three tags outlined for IOE1. However, E tag has a different meaning. E tag is given for the word at the end of a chunk.
- SE: This representation uses five tags which are S, B, I, E and O, for representing chunks. S means that the current word is a chunk consisting of only one word. B means the start of a chunk consisting of more than one word. E means the end of a chunk consisting of more than one word. I means the inside of a chunk consisting of more than two words. O means the outside of any chunk.

We defined five NE label sets for IREX NE extraction tasks by using these five chunk representations. We use "O" to designate words to which none of the NE categories. Each label set in IOB1, IOB2, IOE1 and IOE2 based label sets has $(8 \times 2) + 1 = 17$ labels, and the SE based NE label set has $(8 \times 4) + 1 = 33$ labels. Figure 1 shows examples for these five representations.

B. Named Entity extraction by word-unit chunking

NE extraction by word-unit chunking identifies word chunks consisting of NEs and classifies word chunks into NE categories. We consider methods to assign one of the NE labels to each word for extracting NEs.

We follow the previous Japanese NE extraction methods that have shown better performance [3], [6], [8], [9] for feature selection. Our NE extractors use the following features of the current word, the preceding two words and the two succeeding words as features (5-word window). This setting has shown the best performance in several Japanese NE experiments [3], [6], [8], [9]. In the following explanations, we assume that a sentence consists of m words $\{w_1, ..., w_m\}$ $(0 < m)$.

- Word: We use words tokenized with a morphological analyzer because Japanese has no word boundary marker. We use ChaSen[1] as the morphological analyzer. When we classify i-th word w_i $(1 \leq i \leq m)$ to one of the NE labels, we use w_{i-2},w_{i-1}, w_i, w_{i+1} and w_{i+2} as features.
- Part of speech (POS) tags: We use POS tags of words assigned by ChaSen. Let $POS(w)$ be the POS tag of a word w. We use $POS(w_{i-2})$, $POS(w_{i-1})$, $POS(w_i)$, $POS(w_{i+1})$ and $POS(w_{i+2})$ as features.

[1] http://chasen.naist.jp/hiki/ChaSen/

TABLE II
BASIC CHARACTER TYPE AND NUMBER TYPE

Character type	Hiragana (Japanese syllabary characters), Katakana, Kanji (Chinese letter) Capital alphabet, Lower alphabet, Others.
Number type	$n \leq 12$, $25 \leq n$, $n \leq 100$, $n \leq 2000$, $n < 2000$

Word	POS	CharType	NE label
田中	Noun-Surname	Kanji+	B-ORG
使節	Noun-General	Kanji+	I-ORG
団	Noun-Suffix-General	Kanji	E-ORG
は	Particle-Case-General	Hiragana	O
訪米	Noun-Verb-Connection	Kanji+	S-LOC

↓

Char	POS	CharType	NE label of word	Word	Word CharType	NE label
田	B-Noun-Surname	Kanji	B-B-ORG	B-田中	Kanji+	B-ORG
中	E-Noun-Surname	Kanji	E-B-ORG	E-田中	Kanji+	I-ORG
使	B-Noun-General	Kanji	B-I-ORG	B-使節	Kanji+	I-ORG
節	E-Noun-General	Kanji	E-I-ORG	E-使節	Kanji+	I-ORG
団	S-Noun-Suffix-General	Kanji	S-E-ORG	S-団	Kanji	I-ORG
は	S-Particle-Case-General	Hiragana	S-O	S-は	Hiragana	O
訪	B-Noun-Verb-Connection	Kanji	B-S-LOC	B-訪米	Kanji	O
米	E-Noun-Verb-Connection	Kanji	E-S-LOC	E-訪米	Kanji	B-LOC

Fig. 2. Feature expression for training: The top table is an example of word unit chunking one and the bottom table is character unit chunking one.

– Character types: If a word consists of only one character, the character type is expressed by using the corresponding character types listed in Table II. If a word consisted of more than one character, the character type is expressed by using a combination of the basic character types listed in Table II, such as Kanji-Hiragana[2]. If a word is number, the character type is expressed by the number-type determined by the number type range listed in Table II.
  Let $CT(w)$ be the character type of a word w. We use $CT(w_{i-2})$, $CT(w_{i-1})$, $CT(w_i)$, $CT(w_{i+1})$ and $CT(w_{i+2})$ as features.
– NE labels of preceding words: We use NE labels assigned to the preceding two words in the extraction direction as features. Kudo and Matsumoto proposed to combine English base phrase parsers by voting [20]. To create several English base phrase parsers from a training corpus, they use distinct chunk representation with two parsing directions (Forward/Backward). We also consider the two parsing directions to create NE extractors. Let be $NEL(w)$ the NE label assigned to a word w. If the parsing direction is the end to the begin of a sentence, we use $NEL(w_{i-2})$ and $NEL(w_{i-1})$ as features. If the parsing direction is the begin to the end of a sentence, we use $NEL(w_{i+1})$ and $NEL(w_{i+2})$ as features. In addition to the two parsing direction, we create NE extractors without using the predicted labels as features.

[2]The same character type sequence expressed more than one time is denoted by the character type and "+". For example, character types of "Yamada" are "Capital-alphabet&Small-alphabet+".

To distinguish the same features appeared within 5-word window, features of each word are expressed with the position from the current word. For example, features of two preceding word from the current position i is expressed like "-2-th-word=w_{i-2}" and "-2-th-POS=$POS(w_{i-2})$".

For each NE label set, the following NE extractors are created: NE extractors start parsing from either the beginning or end of sentence direction, and NE extractors parse sentences without using the predicted labels as features. Finally, we implement $(5 \times 3) = 15$ NE extractors.

C. Named Entity extraction by character-unit chunking

Japanese NEs sometimes include a part of a word becoming beginning or end of an NE. To extract Japanese NEs including a part of a word, we apply a character-unit-chunking method.

For example, the "訪米 (visit U.S.A)" in "田中(Tanaka)使節(mission)団(party)は(particle)訪米(visited U.S.A) (Tanaka mission party visit U.S.A.). " dose not match with LOCATION "米(U.S.A)" because this sentence is tokenized as "田中(Tanaka)/使節(mission) /団(party)/は (particle)/訪米(visited U.S.A)", where "/" indicates a word boundary.

To solve this problem, we use a character-unit-chunking-based NE extraction algorithm [8], [9]. Figure 2 shows the examples of a word-unit-chunk representation in the top and a character-unit-chunk representation in the bottom. 2.

We follow the best model of Asahara and Matsumoto [8] for selecting features and chunk representation of character-unit

TABLE III
EXAMPLES OF EXTRACTION RESULTS (TOP) AND EXAMPLES OF NE-RELATED LABELS COLLECTED FROM THE EXTRACTION RESULTS (BOTTOM)

田中/B-ORG	株式会社/E-ORG	上場/O
(Tanaka)	(Co.Ltd.)	(go public)
田中/S-PERSON	社長/O	
(Tanaka)	(president)	
田中/S-PERSON	さん/O	
(Tanaka)	(Mr.)	

↓

Word	Position from the current word	Candidate NE labels in each position	Freq. of NE labels	Ranking in each position
田中 (Tanaka)	current	S-PERSON	2	1
		B-ORG	1	2
	next	O	2	1
		E-ORG	1	2
	two back	O	1	1
さん (Mr.)	current	O	2	1
	previous	S-PERSON	2	1

chunking. In addition to the current position, we use the two preceding and two succeeding characters (5-character window) in character-unit chunking. In the sequel, we assume that a sentence consists of n characters $\{c_1, ..., c_n\}$ ($0 < n$).

- Characters: We use each character as a feature. When we classify i-th character c_i ($1 \leq i \leq n$), we use c_{i-2},c_{i-1},c_i, c_{i+1} and c_{i+2} as features.
- Character types: We assign one of the character types listed in Table II to each character and use it as a feature. Let $CT(c)$ be the character type of a character c, then we use $CT(c_{i-2})$, $CT(c_{i-1})$,$CT(c_i)$, $CT(c_{i+1})$ and $CT(c_{i+2})$ as features.
- Words including characters: We use the words including characters within a widow size as features. Let be $W(c_i)$ the word including i-th character c_i and $P(c_i)$ be the identifier that indicates the position where c_i appears in $W(c_i)$. We combine $W(c_i)$ and $P(c_i)$ for creating a feature. $P(c_i)$ is one of the followings; begging of a word (B), inside of a word (I), end of a word (E) and a character is a word (S).
  For example, if "外務省*(the Ministry of Foreign Affairs)* は*(particle)*" is segmented as "外務省/は", then words including characters are follows; "W(外) = 外務省", "W(務) = 外務省, "W(省) = 外務省" and "W(は)=は". The identifiers that indicate positions where characters appear are follows; "P(外) =B", "P(務) = I", "P(省) = E" and "P(は)=S".
- POS tags of words including characters: Let be $POS(W(c_i))$ the POS tag of the word $W(c_i)$ including i-th character c_i. We use the POS tags of words including characters within window size as features. We express these features with the position identifier $P(c_i)$.
- Character types of words including characters: Let $CT(W(c_i))$ be the character type of the word including i-th character c_i. We use $CT(W(c_{i-2}))$, $CT(W(c_{i-1}))$, $CT(W(c_i))$, $CT(W(c_{i+1}))$ and $CT(W(c_{i+2}))$ as features.
- NE labels of words assigned by a word-unit NE extractor: Let $NEL(W(c_i))$ be the NE label of the word including i-th character c_i. We express these features with the identifier $P(c_i)$ and NE label $NEL(W(c_i))$. In this experiment, we use "$P(c_{i-2}) - NEL(W(c_{i-2}))$", "$P(c_{i-1}) - NEL(W(c_{i-1}))$", "$P(c_i) - NEL(W(c_i))$", "$P(c_{i+1}) - NEL(W(c_{i+1}))$" and "$P(c_{i+2}) - NEL(W(c_{i+2}))$" as features.
- NE labels of preceding extraction results: The NE labels of two preceding extraction results are used as features in the direction of the end to begin of a sentence. The setting have shown good performance in past experiments [8], [9]. Let $NEL(c)$ be the NE label assigned to character c by an NE extractor. In this experiment, we use $NEL(c_{i+1})$ and $NEL(c_{i+2})$ as features.

To distinguish the same features appeared within 5-character window, features of each character are expressed with the position of character from current character. For example, two preceding character from current position i is expressed like "-2-th-character=c_{i-2}".

Each character is classified into one of the $(8 \times 2 + 1) = 17$ NE labels represented by the IOB2 representation. To use the same character-unit-chunking-based NE extractor, we convert IOB1, IOB2, IOE1 and IOE2 based NE extractors output into the SE representation.

D. Support Vector Machines-based NE extractor

We used the chunker YamCha [21], which is based on Support Vector Machines (SVMs) [22], to implement NE extractors. Below we briefly describe SVMs based NE extraction. Suppose we have a set of training data for a binary class problem: $(x_1, y_1), ..., (x_N, y_N)$, where $x_i \in R^n$ is an n dimension feature vector of the i-th sample in the training data and $y_i \in \{+1, -1\}$ is the label of the sample.

TABLE IV
TRAINING AND EVALUATION DATA FOR JAPANESE NE EXTRACTION IN THIS EXPERIMENT

	Training	Evaluation					
NE / Data	CRL data	formal-run GENERAL	formal-run ARREST	domain-specific training	dry-run	dry-run training	Total
ARTIFACT	871	49	13	11	42	67	182
DATE	3654	277	72	69	110	137	665
LOCATION	5660	416	106	165	192	255	1134
MONEY	390	15	8	19	33	32	107
ORGANIZATION	3813	389	74	80	214	270	1027
PERCENT	500	21	0	3	6	19	49
PERSON	3870	355	97	94	169	138	853
TIME	503	59	19	18	24	8	128
Total	19261	1581	389	459	790	926	4145

TABLE V
NE EXTRACTION RESULT WITHOUT UNLABELED DATA ($F_{\beta=1}$): "-F", "-B", AND "-N" INDICATE FORWARD DIRECTION NE EXTRACTION, BACKWARD DIRECTION NE EXTRACTION, AND NE EXTRACTION WITHOUT USING PRECEDING NE LABELS, RESPECTIVELY. AV.($F_{\beta=1}$) AND AV.RANK ARE FOR FIVE PIECES DATA.

Chunk Rep. / Data	CRL Cross	formal-run GENERAL	formal-run ARREST	domain-specific training	dry-run	dry-run training	Av. ($F_{\beta=1}$)	Av. rank
IOB1-F	82.97	84.17	86.02	88.94	81.32	82.10	83.89	8.6
IOB1-B	83.85	83.88	84.69	88.84	81.85	**84.06**	84.18	8
IOB1-N	81.67	80.50	82.64	86.93	79.43	79.64	81.04	17.2
IOB2-F	85.11	82.92	85.56	87.95	80.27	81.49	82.92	14
IOB2-B	**86.07**	84.03	84.84	89.23	81.98	83.74	84.25	6.8
IOB2-N	81.60	83.53	84.44	88.72	81.11	82.65	83.55	13.4
IOE1-F	82.01	83.79	85.87	89.26	81.37	81.87	83.73	10.4
IOE1-B	82.64	83.80	84.65	88.69	81.74	83.77	84.04	10
IOE1-N	81.65	80.52	82.79	86.46	79.27	79.38	80.92	17.6
IOE2-F	82.70	84.65	85.90	89.94	82.68	81.78	84.36	4.8
IOE2-B	83.23	84.00	84.76	89.31	81.50	83.56	84.11	8.2
IOE2-N	81.60	84.11	**87.25**	89.35	82.77	82.09	84.30	4.8
SE-F	85.15	83.95	84.46	89.66	82.15	80.87	83.62	10
SE-B	85.90	83.67	85.52	89.46	81.85	83.11	84.03	8.4
SE-N	85.85	**85.49**	86.13	**90.47**	**83.27**	83.80	**85.83**	1.6

TABLE VI
EXPERIMENTAL RESULTS OF OUR PROPOSED METHODS AND SE-N OF THE BASE LINE ($F_{\beta=1}$): AV.($F_{\beta=1}$) AND AV.RANK ARE FOR FIVE PIECES DATA.

	CRL Cross	formal-run GENERAL	formal-run ARREST	domain-specific training	dry-run	dry-run training	Av. $F_{\beta=1}$	Av. rank
base line	85.85	85.49	86.13	90.47	83.27	83.80	85.83	5
Method 1	87.54(+1.69)	**87.20**	90.31	92.05	84.37	85.68	87.92 (+2.09)	2.8
Method 2	88.04(+2.19)	86.06	89.84	92.51	84.95	85.47	87.77 (+1.94)	3.4
Method 3	88.26(+2.41)	86.41	**91.85**	**92.61**	**85.49**	85.94	**88.46** (+2.63)	**1.6**
Method 4	**88.50**(+2.65)	87.09	90.20	91.78	**85.49**	**86.13**	88.14 (+2.31)	2

The goal is to find a decision function that predicts y for an unseen $x \in R^n$. A SVMs classifier gives the decision function $f(x) = sign(g(x))$ for an input vector x where

$$g(x) = \sum_{z_i \in SV} \alpha_i y_i K(x, z_i) + b.$$

$f(x) = +1$ means that x is a positive member, and $f(x) = -1$ means that is a negative member. The vectors z_i are called support vectors. $y_i \in \{-1, +1\}$ is the label of z_i and $0 < \alpha_i$ is the weight of z_i. Support vectors and other constants are determined by solving a quadratic programming problem. The $K(x, z)$ is a kernel function that maps vectors into a higher dimensional space. We use the polynomial kernel of degree 2 given by $K(x, z) = (1+x\cdot z)^2$ because NE extraction results with the polynomial kernel of degree 2 have shown the best performance in [6], [8], [9]. We used the soft-margin parameter with a value of 1, which is the same value used in past experiments with SVMs [6], [9].

We convert each set of features described in section II-B and II-C into a binary vector as the input of SVMs. We associate each feature with one of the elements of a vector space. Thus, the dimension of the vector space is equal to the number of the types of features. When converting a set of features into a binary vector, we set each element in a vector to 1 if the corresponding feature of the element is included in the set of features, and we set the other elements to 0.

We used the "one-versus-rest method" for extending binary classifiers to N-class classifiers, and prepared N binary classifiers, between a class and the remaining the classes.

This method may involve two or more classifiers which give $+1$ or no classifier give $+1$. In order to solve the problems, we used the Viterbi search. Since SVMs outputs are not probabilities, we use the sigmoid function $s(x) = 1/(1 + exp(-\beta x))$ with $\beta = 1$ to map $g(x)$ to a probability-like value [23].

TABLE VII
PERFORMANCE ON THE EIGHT IREX NE CLASS OF EACH EVALUATION DATA: THE TOP TABLE SHOWS THE RESULTS OF THE BASE LINE. THE BOTTOM TABLE SHOWS THE RESULTS OBTAINED WITH METHOD 3. BETTER PRECISION, RECALL AND F-MEASURE SCORES THAN BASE LINE ONES ARE IN BOLDFACE.

Base line										
	formal-run general		formal-run arrest		domain-specific training		dry-run		dry-run training	
	Rec.	Pre.	Rec.	Pre.	Rec.	Pre.	Rec.	Pre.	Rec.	Pre.
ORGANIZATION	73.78	84.41	66.22	90.74	77.50	95.38	70.09	85.23	71.11	85.71
PERSON	88.45	89.46	85.57	89.25	97.87	94.85	92.31	92.86	90.58	88.03
LOCATION	80.77	87.96	86.79	87.62	87.27	87.80	86.98	88.83	89.02	86.97
ARTIFACT	36.73	48.65	53.85	43.75	45.45	62.50	9.52	25.00	28.36	70.37
DATE	93.86	95.94	95.83	93.24	98.55	97.14	82.73	85.85	90.51	96.88
TIME	94.92	98.25	94.74	100.00	94.44	100.00	87.50	91.30	87.50	100.00
MONEY	100.00	100.00	100.00	100.00	94.74	94.74	96.97	96.97	81.25	100.00
PERCENT	90.48	100.00	-	-	66.67	66.67	100.00	100.00	89.47	94.44
Total	82.54	88.65	83.80	88.59	88.89	92.10	79.37	87.57	79.59	88.48
F-measure	85.49		86.13		90.47		83.27		83.80	
Method 3										
ORGANIZATION	**76.61**	81.64	**83.78**	88.57	77.50	93.94	**74.77**	**86.49**	**72.59**	85.59
PERSON	**89.86**	**90.88**	**92.78**	**90.91**	**98.94**	**95.88**	**95.27**	**93.60**	**92.75**	87.67
LOCATION	**84.62**	**88.00**	**90.57**	**96.97**	**91.52**	**91.52**	**90.10**	88.27	**92.16**	**91.44**
ARTIFACT	**38.78**	46.34	**61.54**	**53.33**	**63.64**	**63.64**	**19.05**	**38.10**	**43.28**	69.05
DATE	**94.95**	95.64	**100.00**	**97.30**	**100.00**	**100.00**	**88.18**	85.09	**93.43**	**96.97**
TIME	**96.61**	**98.28**	**100.00**	**100.00**	94.44	**100.00**	**95.83**	85.19	**100.00**	88.89
MONEY	100.00	100.00	100.00	100.00	94.74	94.74	96.97	96.97	81.25	100.00
PERCENT	90.48	95.00	-	-	**100.00**	**75.00**	100.00	100.00	89.47	94.44
Total	**84.88**	88.00	**91.26**	**92.45**	**91.50**	**93.75**	**83.54**	87.53	**82.83**	**89.29**
F-measure	**86.41**(+0.92)		**91.85**(+5.72)		**92.61**(+2.14)		**85.49**(+2.22)		**85.94**(+2.14)	

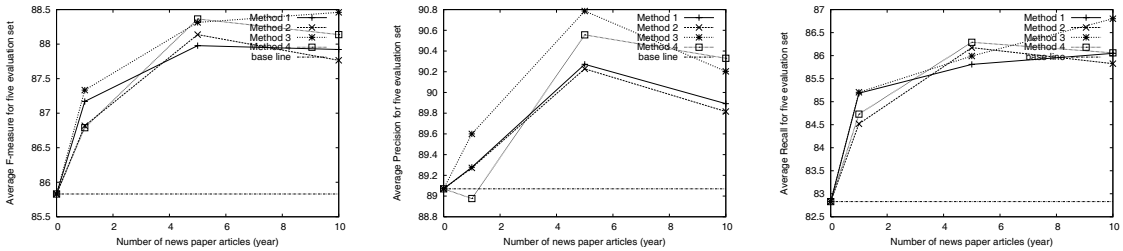


Fig. 3. Average F-measure, Precision and Recall obtained with NE extractors using NE-related labels of words collected from 1, 5 and 10 year news articles

III. FEATURE AUGMENTATION WITH UNLABELED DATA

In this section we describe NE-related labels collected from unlabeled data and the collection methods of the NE-related labels.

A. NE-related labels of words

We use NE-related information of words, which we call NE-related labels, as features. We collect the following information as NE-related labels from unlabeled data parsed with NE extractors.

- **Candidate NE class labels of words**: We collect the 33 types of NE labels defined by the SE representation as candidate NE labels of words. We also collect NE class labels of co-occurring words of each word by the same representation. We collect candidate NE labels of co-occurring words within two preceding and two succeeding words. We collect the information from the parsed results of unlabeled data with NE extractors.
  For example, as the candidate NE class labels of "田中*(Tanaka)*", B-ORG and S-PERSON are collected from the examples listed in Table III; as the candidate NE class labels of preceding words of "田中*(Tanaka)*", E-ORG and O for words appearing the next of "田中*(Tanaka)*", and O for the word two ahead position of "田中*(Tanaka)*" are collected.
- **Frequencies of candidate NE class labels**: These are the frequencies of the NE candidate labels of each word, which are counted from the parsed results. For example, as the frequencies for the candidate NE class labels of "田中*(Tanaka)*", "B-ORG" is counted once, "S-PERSON" is counted twice from the examples listed in Table III. For the candidate NE class labels of surrounding words of "田中*(Tanaka)*", "E-ORG" is collected once and "O" is collected twice as the candidate NE class labels of the next words position, and "O" is collected once as the candidate NE class labels of two ahead position. To express the frequencies of NE-related labels as binary features, we categorize the frequencies of these NE-related labels by n of each frequency; $n \leq 10$, $10 < n \leq 100$, and $100 < n$.

– **Ranking of NE-related labels**: This information is the ranking of candidate NE class labels for each word. Each ranking is decided with their frequencies counted from the parsed results. For the ranking in the candidate NE class labels of "田中*(Tanaka)*", "S-PERSON" is ranked as the first, and "B-ORG" is ranked as the second. As the ranking of the candidate NE class labels of words appeared in the next of "田中*(Tanaka)*", "O" is ranked as the first, and "E-ORG" is ranked as the second.

In this experimental setting, up to $495\ (= 3 \times 33 \times 5)$ features are collected for each word because we collect words with the candidate NE class labels of the two preceding and two succeeding word positions in addition to the current word ones.

We used the NE-related labels in the word-unit chunking as features. As a result, up to $2475\ (= 495 \times 5)$ features are added when training and extraction because we used features of words within 5-word window in addition to the current word features.

B. Collection methods of NE-related labels by using several NE extractors

We collect NE-related labels of words as follows:

– Step 1: Create NE extractors from a given training data by using SVMs. We create 15 NE extractors as described in Section II-A.
– Step 2: Extract NEs from unlabeled data with the NE extractors created in Step 1.
– Step 3: Collect NE-related labels from the automatically labeled data in Step 2.
– Step 4: Add the collected NE-related labels in Step 3 to the same training data used in Step 1 as new features.
– Step 5: Construct a new NE extractor from the training corpus created in Step 4 by using SVMs. As for the model of the new NE extractor, we use the same model of the best NE extractor model in the 15 extractor created at Step 1. When the new NE extractor extracts NEs, the collected NE-related labels are also used as features.

Some problems of using parsed results for collecting NE-related labels of words are noise and low coverage caused by wrong extraction. To avoid these problems as much as possible, we consider the following methods used at Step 3 and 4.

– **Method 1**: Collect NE-related labels of words from a parsed result with using an NE extractor. We use the NE extractor which shows the best accuracy of all 15 NE extractors and we use the F-measure ($F_{\beta=1}$) as a measure of accuracy. We think that the collected NE-related labels of words by this method archive moderately good for recall and precision.
– **Method 2**: Collect NE-related labels of words from each word chunk which all the NE extractors extract the word chunk as the same NE class. Even if one of the outputs of NE extractors to a word chunk is different from the outputs of the others, the word chunk is just ignored. By using this method, the collected NE-related labels of words have better precision than Method 1, but the recall is worse.
– **Method 3**: Collect NE-related labels from all the NE extractor outputs. By using this method, the collected NE-related labels of words are better recall than Methods 1 and 2, but the precision is worse than those two particular methods. When we use this method, we use the average frequencies of extracted results obtained from 15 NE extractor outputs as the frequencies of NE-related labels.
– **Method 4**: Collect NE-related labels of words by using Methods 1, 2 and 3, and all the NE-related labels are used as features differently. By using all the NE-related labels from Methods 1, 2 and 3, an NE extractor uses all their characteristics.

For collecting NE-related labels by using NE extractors based on distinct chunk representations as the SE representation, we convert the NE extractor outputs into the SE representation.

A problem of parsing large data by SVMs based NE extractors is the processing speed [6], [21]. To improve the processing speed of SVMs based NE extractors, we use Polynomial Kernel Expanded (PKE) method [21], which converts a kernel-based classifier into a simple linear classifier by expanding all feature combinations.

IV. EXPERIMENTS

A. Experimental Settings

We used the following data.

– Training and Evaluations data: We used the six pieces of Japanese NE data prepared by IREX [17]. We used CRL data for training and evaluation with cross-validation. We used the five data for evaluation, consisting of formal-run GENERAL, formal-run ARREST, domain-specific training data, dry-run data and dry-run training data. Table IV lists the statistics of NE types for each data set.
– Unlabeled data: We used 10-year period news articles: The news articles are the Mainichi Shinbun over a 10-year period between 1991 and 1993, 1995 and 1998, and 2000 to 2002. To keep the five pieces of evaluation data fully unseen in the training phases, we excluded the 1994 and 1999 news articles because they include the evaluation data.

We did the following experiments.

– Five-fold cross-validation on CRL data: We split CRL data into five pieces data. Each piece data consists of about $1/5$ articles included in CRL data. We separately collected NE-related labels for each classifier created from $4/5$ size of CRL data. We totally parsed $(75 * 10) = 750$ years amount of newspaper articles because we created $(15 * 5) = 75$ extractors for cross-validation. We

TABLE VIII
COMPARISON WITH RELATED WORKS: GE AND AR INDICATE GENERAL AND ARREST.

Method	GE	AR	CRL-DATA	NE extraction algorithm	Lexical resources
Uchimoto et al.[3]	80.17	85.75	-	Chunking by word with ME and transformation rules for word unit problems	Handcrafted NE dictionaries
Takemoto et al. [1]	83.86	-	-	Handcrafted Rules and Compound Lexicon for word unit problems	-
Utsuro et al. [24]	84.07	-	-	Combining three ME based NE extractor outputs by decision list based stacking	-
Yamada et al.[4]	-	-	83.2	Chunking by word with SVMs and examples in training data are segmented	-
Isozaki and Kazawa [6]	85.77	-	86.77	Chunking by word with SVMs and template rules for word unit problems	NTT Goi Taikei
Asahara and Matsumoto [8]	-	-	87.21	Chunking by character with SVMs using n-best results of morphological analysis	NTT Goi Taikei
Nakano and Hirai [9]	-	-	89.03	Chunking by character with SVMs using Japanese base phrase information	NTT Goi Taikei
Sasano and Kurohashi [25]	87.72	-	89.40	Chinking by word with SVMs using structural information	NTT Goi Taikei
Kazama and Torisawa [26]	-	-	88.93	Chunking by character with CRFs	Web documents and Wikipedia
our base line	85.49	86.13	85.85	Character-based chunking using outputs of a word-based NE extractor by stacking with SVMs	-
Method 1	**87.20**	**90.31**	87.54		unlabeled data
Method 2	**86.06**	**89.84**	88.04		unlabeled data
Method 3	**86.41**	**91.85**	88.26		unlabeled data
Method 4	**87.09**	**90.20**	88.50		unlabeled data

created 5 NE extractors based on character-unit chunking for each cross-validation.

– Evaluation on five pieces data: We created NE extractors by using CRL data, and we applied the NE extractors to the five pieces data.

We compared performance of NE extractors by using F-measure, which is defined as follows.
Recall = NUM / (the number of correct NEs),
Precision = NUM / (the number of NEs extracted by an NE extractor),
F-measure = 2 × Recall × Precision / (Recall + Precision), where NUM is the number of NEs correctly identified by an NE extractor.

B. Experimental Results

Table V lists the experimental results obtained with the 15 NE extractors. These NE extractors did not use NE-related labels presented in Section III-A as features. Of all 15 NE extractors, the SE-N based extractor, which is based on the SE representation without using NE labels assigned to preceding words as features, showed the best performance in the evaluation one the five pieces of data and the third best performance in the evaluation on cross-validation of CRL data. We used the SE-N based extractor results as the base line and the NE extractor for the Method 1. We also used SE-N model for implementing new NE extractors.

Table VI lists the experimental results obtained with NE extractors using NE-related labels of words collected from 10 year news articles by Methods 1 to 4 and the results of our base line.

All our methods exceeded the base line. These results show that NE-related labels of words collected from unlabeled data contributed to improved accuracy. Methods 2 to 4, which used NE-related labels collected with several extractors, showed better performance on cross-validation of CRL data than Method 1. Methods 3 and 4 showed better performance on five pieces data than Method 1. These results show that NE-related labels collected with several NE extractors contributed to improved accuracy.

Method 4 showed the best performance on five-fold cross validation, and the average F-measure was 2.65 points higher than the base line. Method 3 showed the best performance on five pieces data, and the average F-measure was 2.63 points higher than the base line.

Table VII lists that performance on eight NEs of each evaluation data. The results show that recalls of eight NEs are improved by using the NE-related labels while keeping higher precision in many cases.

Figure 3 shows the average F-measure, precision and recall for the five evaluation data with different amount of newspaper articles for collecting NE-related labels. Methods 1 to 4 based NE extractors with 1, 5 and 10 year news articles showed higher accuracy than the base line one. Furthermore, these results show that the higher recall than the base line while keeping higher precision except for Method 4 with 1 year news articles.

However, the precision scores of Methods 1 to 4 with 10 year news articles were worse than the precisions of Methods

1 to 4 with 5 year news articles. The F-measure scores of Methods 1, 2 and 3 were also worse than the corresponding Methods with 5 year news articles. A reason seems to be noise given by much number of NE-related labels of words.

One of the methods to solve the problem what we think is utilization of richer information as features, such as larger context information, phrase information [9] and dependency information, and so on, may be effective. The other is pruning of NE-related labels by using a threshold value like their frequencies and their ranking.

C. Comparison with Previous Work

Table VIII lists the results of the previous works using IREX Japanese NE extraction tasks. Compared to previous works, our base line showed relatively higher performance without using additional lexical resources. The result showed that the combination of a word-unit NE extraction and a character-unit NE-extraction is effective for Japanese Named Entity Extraction.

All the results obtained with our Methods 1 to 4, which use CRL data and unlabeled data of 10 year news articles for training, showed higher performance than the previous best results on IREX GENERAL and ARREST tasks. Uchimoto et al. [3] used a handcrafted NE dictionary and training data consisting of CRL data, dry-run data, dry-run training data, and domain-specific training data, for creating a maximum entropy (ME) model based NE extractor. Isozaki and Kazawa [6] used NTT GOI Taikei [27], which is a handcrafted thesaurus and CRL data for training. These results showed that NE-related labels of words collected from unlabeled data are useful just as well as handcrafted resources.

Our results showed higher performance than the handcrafted-rule based NE extractor [1] and the NE extraction combining outputs of three NE extractors by stacking [24].

Our methods showed better performance than a character-unit chunking using n-best morphological analysis results and NTT GOI Taikei as features [8]. However, our methods showed slightly worse performance than methods using Japanese base phrase information and NTT GOI Taikei [9], [25] and gazetteers induced from web documents and Wikipedia [26]. The reason seems to be the difference of features. We think that we can improve performance of our NE extractors by using features used in their character-unit chunking algorithms.

V. Conclusion

This paper proposes feature augmentation methods using unlabeled data and several Named Entity (NE) extractors. Our feature augmentation techniques collect candidate NE class labels of words and NE class labels of co-occurring words from unlabeled data parsed with several NE extractors. The experimental results with IREX GENERAL and ARREST tasks showed that NE extractors with our proposal methods showed higher performance than the previous best results using handcrafted lexical resources.

References

[1] Y. Takemoto, T. Fukushima, and H. Yamada, "A Japanese named entity extraction system based on building a large-scale and high quality dictionary and pattern-matching rules (in Japanese)," in *IPSJ Journal, 42(6)*, 2001, pp. 1580–1591.

[2] M. Collins and Y. Singer, "Unsupervised models for named entity classification," in *Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999. [Online]. Available: citeseer.ist.psu.edu/collins99unsupervised.html

[3] K. Uchimoto, Q. Ma, M. Murata, H. Ozaku, M. Utiyama, and H. Isahara, "Named entity extraction based on a maximum entropy model and transformati on rules." in *Proc. of the ACL 2000*, 2000, pp. 326–335.

[4] H. Yamada, T. Kudoh, and Y. Matsumoto, "Japanese named entity extraction using Support Vector Machine (in Japanese)," in *IPSJ Journal, 43(1)*, 2002, pp. 44–53.

[5] X. Carreras, L. Màrques, and L. Padró, "Named entity extraction using adaboost," in *Proc. of CoNLL-2002*. Taipei, Taiwan, 2002, pp. 167–170.

[6] H. Isozaki and H. Kazawa, "Speeding up named entity recognition based on Support Vector Machines (in Japanese)," in *IPSJ SIG notes NL-149-1*, 2002, pp. 1–8.

[7] R. Florian, A. Ittycheriah, H. Jing, and T. Zhang, "Named entity recognition through classifier combination," in *Proc. of CoNLL-2003*, 2003, pp. 168–171.

[8] M. Asahara and Y. Matsumoto, "Japanese named entity extraction with redundant morphological analysis," in *Proc. of HLT-NAACL 2003*, 2003, pp. 8–15.

[9] K. Nakano and Y. Hirai, "Japanese named entity extraction with bunsetsu features (in Japanese)," in *IPSJ Journal, 45(3)*, 2004, pp. 934–941.

[10] S. Miller, J. Guinness, and A. Zamanian, "Name tagging with word clusters and discriminative training." in *HLT-NAACL*, 2004, pp. 337–342.

[11] D. Freitag, "Trained named entity recognition using distributional clusters," in *Proc. of EMNLP 2004*. Association for Computational Linguistics, July 2004, pp. 262–269.

[12] R. Ando and T. Zhang, "A high-performance semi-supervised learning method for text chunking," in *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics*. Ann Arbor, Michigan: Association for Computational Linguistics, June 2005, pp. 1–9. [Online]. Available: http://www.aclweb.org/anthology/P/P05/P05-1001

[13] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proc. of ACL-1995*, 1995, pp. 189–196.

[14] E. Riloff and R. Jones, "Learning dictionaries for information extraction by multi-level bootstrapping," in *AAAI/IAAI*, 1999, pp. 474–479. [Online]. Available: citeseer.ist.psu.edu/article/riloff99learning.html

[15] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. of the 11th COLT*, 1998, pp. 92–100.

[16] R. K. Ando, "Semantic lexicon construction: Learning from unlabeled data via spectral analysis," in *Proc. of CoNLL-2004*. Boston, MA, USA, 2004, pp. 9–16.

[17] C. IREX, *Proc. of the IREX workshop*, 1999.

[18] L. Ramshaw and M. Marcus, "Text chunking using transformation-based learning," in *Proc. of the Third Workshop on Very Large Corpora*. Association for Computational Linguistics, 1995, pp. 82–94. [Online]. Available: citeseer.ist.psu.edu/article/ramshaw95text.html

[19] E. Tjong Kim Sang and J. Veenstra, "Representing text chunks." in *Proc. of EACL '99*, Bergen, Norway, 1999. [Online]. Available: http://www.cnts.ua.ac.be/Publications/1999/TV99

[20] T. Kudo and Y. Matsumoto, "Chunking with Support Vector Machines," in *Proc. of NAACL 2001*, 2001.

[21] ——, "Fast methods for kernel-based text analysis," in *Proc. of ACL-2003*, 2003, pp. 24–31.

[22] V. Vapnik, *Statistical Learning Theory*. John Wiley & Sons, 1998.

[23] J. C. Platt, *Probabilities for SV machines*, A. J. Smola, P. L. Bartlett, B. Sch ¨ olkopf, and D. Schuurmans, Eds. MIT Press, 2000.

[24] T. Utsuro, M. Sassano, and K. Uchimoto, "Combining outputs of multiple Japanese named entity chunkers by stacking," in *Proc. of EMNLP 2002*, 2002, pp. 281–288.

[25] R. Sasano and S. Kurohashi, “Japanese named entity recognition using structural natural language processing,” in *Proc. of IJCNLP’08*, 2008, pp. 607–612.
[26] J. Kazama and K. Torisawa, “Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations,” in *Proc. of ACL-08: HLT*, 2008, pp. 407–415.
[27] S. Ikehara, M. Miyazaki, S. Shirai, A. Yokoo, H. Nakaiwa, K. Ogura, Y. Ooyama, and Y. Hayashi, *Goi-Taikei -A Japanese Lexicon CDROM*. Iwanami Shoten, 1999.

Revised N-Gram based Automatic Spelling Correction Tool to Improve Retrieval Effectiveness

Farag Ahmed, Ernesto William De Luca, and Andreas Nürnberger

***Abstract*—We present a language-independent spell-checker that is based on an enhancement of the n-gram model. The spell checker is proposing correction suggestions by selecting the most promising candidates from a ranked list of correction candidates that is derived based on n-gram statistics and lexical resources. Besides motivating and describing the developed techniques, we briefly discuss the use of the proposed approach in an application for keyword- and semantic-based search support. In addition, the proposed tool was compared with state-of-the-art spelling correction approaches. The evaluation showed that it outperforms the other methods.**

***Index terms*—Spelling correction, n-gram, information retrieval effectiveness.**

I. Introduction

THE problem of devising algorithms and techniques to automatically correct words in texts has become a perennial research challenge. Work began as early as the 1960s on computer techniques for automatic spelling correction and automatic text recognition, and it has continued up to the present. There are good reasons for the continuing research efforts in this area in order to improve quality and performance and to broaden the spectrum of possible applications [1]. For example, even though system programs (language processors, operating systems, etc.) have become increasingly powerful and sophisticated, they do not assist the user (with very few exceptions) in correcting many of the obvious spelling errors in the source input. There are two types of word errors, the real-word error and the non-word error. Real-word errors are misspelled words that have a meaning and can be found in a dictionary. Non-word errors are words that have no meaning and are thus not included in a dictionary. We concentrate on the correction of the non-word error with the proposed algorithm. Damerau (1964) found that 80% of misspelled words that are non-word errors are the result of a single insertion, deletion, substitution or transposition of letters [2]. Therefore, it seems reasonable to base correction algorithms on measures that consider these simple operations. However, approaches based on pure n-gram statistics (which account for these operations implicitly) have also proven to provide good performance [1, 15].

In this paper, we propose an approach that is based on an enhancement of the n-gram model. Therefore, we first discuss briefly, related work on spelling correction in Section 2. Afterwards, we describe, in detail, in Section 3 our spell checking approach MultiSpell. In Section 4, we present an evaluation based on benchmark data sets in the English and Portuguese language and conclude with a brief discussion.

II. Approaches of Some Spell Checkers

Algorithmic techniques for detecting and correcting spelling errors in text have a long and robust history in computer science [1]. Many approaches have been applied since people started to deal with this problem. Different techniques like edit distance [4], rule-based techniques [10], n-grams [20], probabilistic techniques [14], neural nets [15], similarity key techniques [16, 17] and noisy channel model [18, 19] have been proposed. All of these are based on the idea of calculating the similarity between the misspelled word and the words contained in a dictionary. In the following, we describe briefly one of the most popular approaches (Aspell) and one recently proposed approach for the Portuguese language (TST) [13] that we used for comparison.

GNU Aspell, usually called just Aspell, is a standard spell-check software for the GNU software system. There are dictionaries for about 70 languages available. GNU Aspell is a Free and Open Source and can be downloaded under http://aspell.sourceforge.net/. In contrast to Ispell, which suggests words with small edit-distance, Aspell in addition compares sounds-like equivalents (computed for English words using the metaphone algorithm [21]) up to a given edit distance.

The Ternary Search Trees [13] approach (TST) is a dictionary data structure working with string-keys. It can find, remove and add these keys quickly and also easily search the tree for partial matches. Additionally, near-match functions can be implemented. These give the possibility to suggest alternatives for misspelled words.

For a more conclusive overview of spell-check approaches see [1, 15].

Manuscript received October 23, 2008. Manuscript accepted for publication August 22, 2009.

Farag Ahmed and Andreas Nürnberger are with Data and Knowledge Engineering Group, Institute for Knowledge and Language Engineering, Otto-von-Guericke University of Magdeburg, Germany.

Ernesto William De Luca is with Compentence Center Information Retrieval & Machine Learning Distributed Artificial Intelligence Laboratory, Technical University of Berlin, Germany.

III. An Algorithm Based on N-Gram Statistics: MultiSpell

The algorithm we propose, in the following, is a language-independent spell-checker that is based on an enhancement of the n-gram model. It is able to detect the correction suggestions by assigning weights to a list of possible correction candidates, based on n-gram statistics and lexical resources, in order to detect the non-word errors and to derive correction candidates. In the following, we describe first of all the lexical re-source we used (MultiWordNet) and then in detail the proposed MultiSpell algorithm.

A. Lexical Resources

Lexical resources provide linguistic information about words of natural languages. This information can be represented in very diverse data structures, from simple lists to complex resources with many types of linguistic information and relations associated with the entries stored in the resource.

These resources are used for preparing, processing and managing linguistic information and knowledge needed for the computational processing of natural language [3]. An example of such large scale lexical resources is given by linguistic ontologies that cover many words of a language and have a hierarchical structure based on the relationship between concepts.

We propose to use these dictionaries, and especially MultiWordNet [6], the most important lexical resource available. It covers nouns, verbs, adjectives and adverbs. For our purpose, we use the words provided (~80000 entries for the English language) from this resource to correct the misspelled word. Therefore, we extracted all words contained in it with all its linguistic relationships.

B. Computing Similarity Scores Based on N-Grams

The idea of using n-grams in language processing was discussed first by Shannon [8]. After this initial work, the idea of using n-grams has been applied to many problems such as word prediction, spelling correction, speech recognition, translated word correction and string searching. One main advantage of the n-gram method is that it is language independent.

In a spelling correction task, an n-gram is a sequence of n letters in a word or a string. The n-gram model can be used to compute the similarity between two strings, by counting the number of similar n-grams they share. The more similar n-grams between two strings exist the more similar they are. Based on this idea the similarity coefficient [9] can be derived. The similarity coefficient δ is defined by the following equation:

$$\delta_n(a,b) = \frac{|\alpha \cap \beta|}{|\alpha \cup \beta|} \qquad (1)$$

where α and β are the n-gram sets for two words *a* and *b* to be compared. $|\alpha \cap \beta|$ denotes the number of similar n-grams in α and β, and $|\alpha \cup \beta|$ denotes the number of unique n-grams in the union of α and β. Table I shows an example for the calculation of the similarity coefficient for the misspelled word "secceeded" and the correct word "succeeded" using an n-gram with n=2 (bigram).

TABLE I
CALCULATING THE BIGRAMS SIMILARITY COEFFICIENT BETWEEN TWO STRINGS.

bi-grams union	*succeeded*	*secceeded*
su	1	-
uc	1	-
cc	1	1
ce	1	1
ee	1	1
ed	1	1
de	1	1
ed	1	1
se	-	1
ec	-	1
Similarity coefficient	6/10 = 0.6	

C. Revised N-Gram Based Approach

Yannakoudakis and Fawthrop [10] found that in most cases the first letter in the misspelled word is almost always correct and also the misspelled and real word will be either the same length or the length differs just by one. For some examples, we like to refer the reader to the list of commonly misspelled words in English published in [12]. Furthermore, the pure n-gram based approach to compute the similarity coefficient as described above, does not consider the order of the n-grams [22]. This might, however, be important since typing or misspelling errors usually affect only a specific part of the word. Therefore, we revised the computation of a similarity between words to take these two aspects into account.

In the following, we describe our algorithm for $n = 2$ (bigrams) for simplicity. However, the approach can be applied for trigrams and n-grams with $n > 3$ as well. We define bigrams of words by their respective position in the word $w_{i,i+(n-1)}$ where i defines the position of the first letter and $i + (n-1)$ the position of the last letter of the considered n-gram. Thus, the last possible position of an n-gram in a word is defined by $j = |w| - n + 1$, where $|w|$ defines the length of the word.

In order to consider the findings of Yannakoudakis and Fawthrop as mentioned above, we replace the first and the last n-gram by the first and the last letter of the respective words. Thus, when computing the similarity score these elements are compared directly, independent of the remaining n-grams between them.

In order to deal with the second aspect mentioned above, we define a window of n-grams of the correction candidate words that should be compared, i.e. while in Eq. (1) all n-grams are compared with each other, we only compare n-grams that are in close proximity to the position of the n-gram in the word to be corrected when computing the similarity score. An example is given in Fig. 1, where w' defines the misspelled word and w a correction candidate. Here, the n-gram $w'_{4,5}$ of w' will only be compared to the n-grams $w_{3,4}$,

$w_{4,5}$ and $w_{5,6}$ of the correction candidate w, i.e. even if the n-gram $w'_{4,5}$ is similar to $w_{2,3}$ this would not count towards the similarity score of the words w' and w.

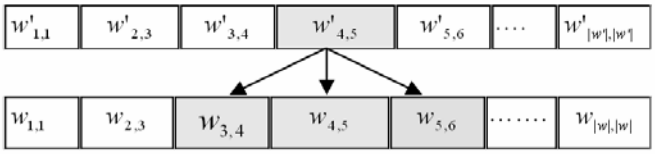


Fig. 1. Bigram comparison for misspelled word w' and a correction candidate w using a comparison window of size 3. Notice that the first and last n-gram represent the first and the last letters only and are therefore always of size one.

Overall, the computation of the similarity score S for a given n-gram size n and a given odd-numbered window size m can be defined as follows, assuming that u is the longer word (if v is longer than u and v can simply be exchanged):

$$S_{n,m}(u,v) = \frac{g(u_{1,1},v_{1,1}) + g(u_{|u|,|u|},v_{|v|,|v|}) + \sum_{i=2}^{|u|-n+1}\sum_{j=-\frac{m-1}{2}}^{\frac{m-1}{2}} g(u_{i,i+(n-1)},v_{i+j,i+j+(n-1)})}{N} \quad (2)$$

$$\text{where } g(a,b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{otherwise} \end{cases} \quad \text{and}$$

$$u_{i,j} = \begin{cases} substring(u,i,j) & \text{if } i <= j \\ "" & \text{otherwise.} \end{cases}$$

Here, $g(u_{1,1},v_{1,1})$ compares the first and $g(u_{|u|,|u|},v_{|v|,|v|})$ the last characters of the words u and v and the nested sum counts the number of n-grams in v that are similar to n-grams in a window of size m around the same position in word v. N is computed similarly as in Eq. (1). In Fig. 2 the specific cases that have to be considered when computing the similarity score S are summarized.

D. The MultiSpell Algorithm

The first stage of the MultiSpell algorithm is to compare the keywords given from the user with the correct words contained in the dictionary. First of all, we check based on the used dictionary (here, based on the words extracted from MultiWordNet) if the word is misspelled. If this is the case, the algorithm builds n-grams for the misspelled word. Then we select correction candidates from the dictionary. In order to keep the number of correction candidates as small as possible, we select only words as candidates that are two charters shorter or longer than the misspelled word. This is motivated by the work of Turba [11], who has shown that most misspelled words differ in length only by one character from the correct word.

For the selected words the n-grams are computed and the similarity score is computed according to Eq. (2). The correction candidates can then be simply sorted by the obtained similarity score and the word with the highest score is proposed as the best correction candidate.

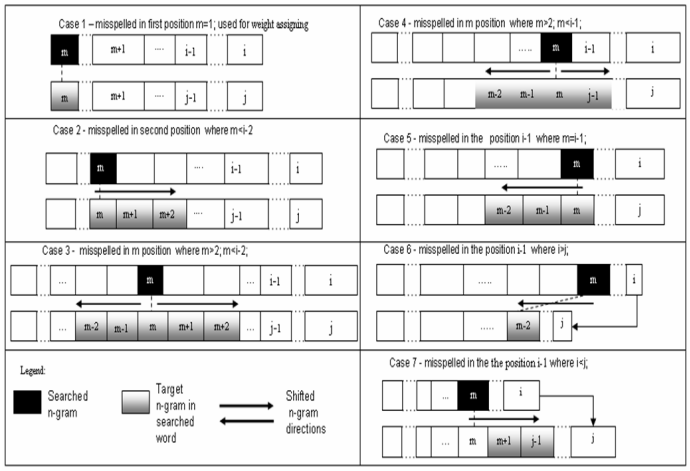


Fig. 2. Comparing n-grams based on the MultiSpell algorithm.

E. Spelling Correction for Keyword- and Semantic-based Search Support

MultiSpell has been also integrated as a pre-processing approach in the Sense Folder Framework [25]. It can be applied to queries and documents, in order to support users during keyword-based and semantic-based search. The first is an important task for retrieving the relevant documents related to the query identifying the misspelled words and correct them for a correct interpretation [23] (see also Fig. 3). The second is specifically trying to improve the semantic search process [24]; therefore several problems have to be addressed, before the semantic classification of documents is started. When users mistype the query in writing, the system has to be able to give correction alternatives to continue the semantic-based search.

The semantic-based search differs from the "normal" search, because users are "redirected" to semantic concepts that could describe their query. This semantic support is provided in the user interface. On the left side of the user interface (see Fig. 4) suggestions are generated by MultiSpell and presented to the user for starting the semantic-based search.

In this case, the use of Multispell is mostly helpful, not only because it performs an efficient correction (as shown in Fig. 3), but also because it can "redirect" the user to a semantic search (see Fig. 4). Thus, if the user types a word that is not contained in the lexical resource used, the system can suggest other "similar" words according to the words found in the resource. Then, a semantic classification is started using the words selected by the user.

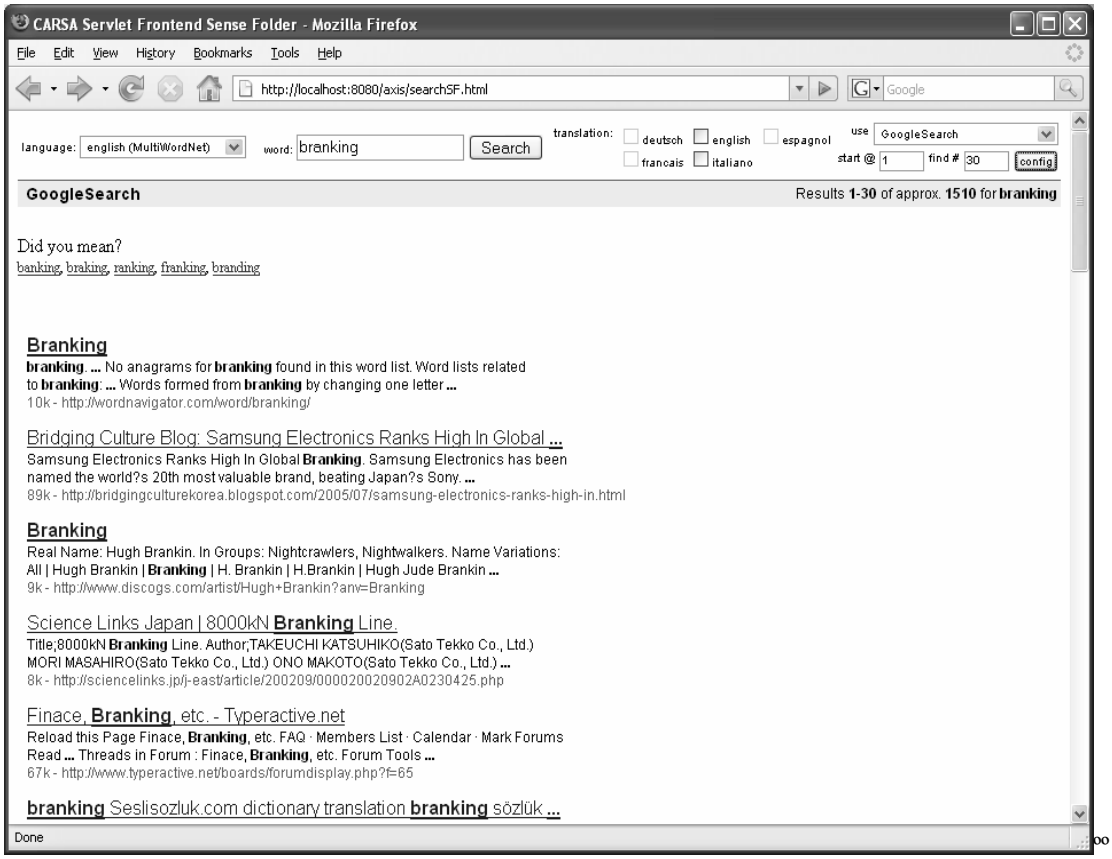


Fig. 3. Corrections for a misspelled word (MultiSpell) in the Sense Folder Framework .

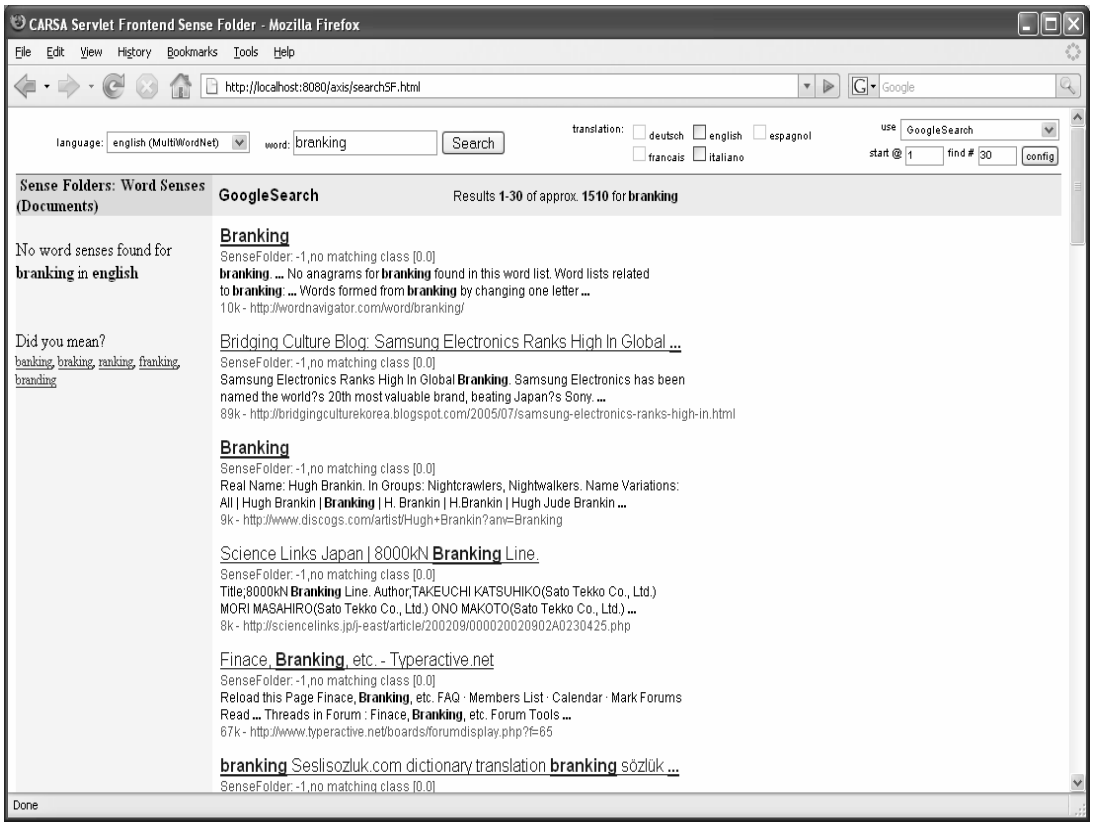


Fig. 4. Using MultiSpell in the Sense Folder Framework for Semantic Search Support.

IV. Comparison and Evaluation of Results

In the following, we show results of some experiments done for the English and Portuguese language. The first evaluation was done on the whole English commonly misspelled word list provided in [12]. Afterwards, we compared the results of our spell checker MultiSpell with the results of the TST approach (in one experiment, for the Portuguese language) and of the Aspell approach (in two experiments, for the Portuguese and the English language), showing that the proposed approach always achieved the best results.

For the first evaluation, we used the whole list of commonly misspelled words in English consisting of 3975 words as published in [12]. This list of common spelling mistakes is represented by a table consisting of two columns. The first one shows the misspelled word, the second the correct spelling. For the evaluations, we only considered the correction words that were ranked as best correction word, i.e., even if the second word would have been the correct candidate, this was counted as a wrong correction. We first used all misspelled words of the list, using the bigram case and just the first candidate correction. MultiSpell corrected 3334 misspelled words (84%) and failed for 641 misspelled words (16%) although it provided similar corrections in many cases. For example the word advice was suggested instead of advised for the misspelled word advised. Another example is the provided correction algebraically instead of algebraic for the misspelled word algebraical (see Table V in the Appendix). These suggestions were classified as wrong in our approach, even though they belong to the same word sense. Second, we used trigrams. This showed lower performance and efficiency. MultiSpell corrected 2900 words (73%) and failed for 1075 (27%) as shown in Table II.

A. Evaluation of English Spelling Correction

For the second evaluation, we randomly selected a set of only 120 misspelled words obtained from Wikipedia [12] and not the whole list. All error types and starting letters of the words were taken into account. We compared MultiSpell with Aspell, MicrosoftWord, and Google. Since Aspell provides a list of candidate corrections we took just the first candidate from the list assuming that the first candidate is the most likely one proposed by the algorithm. MicrosoftWord and Google provided only one correction candidate. Table III and Table V (in the Appendix) show that MultiSpell finds the correct spelling for 109 words (90%). In comparison, Google can correct 106 (88%) words, while Aspell and MicrosoftWord 105 words (87.5%). MultiSpell detected 6 of 16 of the multiple correction words (which have more than one possible correction), but it doesn't fail to provide at least one correct suggestion. Aspell detected just two of the multiple corrections and it failed just one time to provide a suggestion for one of the multiple corrections.

TABLE II
Comparison between bigram and trigram in whole English data set (3975 words).

	bigram	trigram
correct	3334 (84%)	2900 (73%)
wrong	641 (16%)	1075 (27%)

TABLE III
Comparison of MultiSpell, Aspell, Microsoft Word and Google for English.

	MultiSpell	Aspell	Microsoft Word	Google
correct	109 (90%)	105 (87.5%)	105 (87.5%)	106 (88%)
wrong	11 (10%)	15 (12.5%)	15 (12.5%)	14 (12%)

TABLE IV
Comparison of MultiSpell, Aspell and TST for the Portuguese language.

	MultiSpell	TST	Aspell
correct	97 (80%)	78 (65%)	65 (54%)
wrong	23 (20%)	42 (35%)	55 (46%)

B. Evaluation of Portuguese Spelling Correction

The last evaluation was done for the Portuguese language. Bruno and Mário [13] implemented an algorithm using Ternary Search Trees (TST). The authors show experiments in correcting a list of some Portuguese words and comparing their results with Aspell. Here we compared MultiSpell on the whole list (120 Portuguese words) available from their experiments explained in [13], applying our algorithm and comparing it with the Aspell and TST algorithm. Given that MultiWordNet does not provide any Portuguese word senses, we used the dictionary made available from [13] comparing the approaches. Our algorithm succeeded to correct 97 misspelled words (80%), TST succeeded to correct 78 misspelled words (65%) and Aspell succeeded to correct 65 misspelled words (54%) as shown in Table IV and Table VI (in the Appendix).

IV. Conclusions

In this paper we proposed a language-independent spell-checker that is based on an enhancement of a pure n-gram based model. Furthermore, we presented evaluations on English and Portuguese benchmark data sets of misspelled words. The obtained results outperformed other state-of-the-art methods. In future work, we plan to further optimize the algorithm and data structure used to compute the similarity scores. Furthermore, the algorithm should be tested on data sets for other languages.

APPENDIX: EVALUATION TABLES FOR ENGLISH AND PORTUGUESE

Table V contains results of word corrections in English, while Table VI contains results of word corrections in Portuguese.

TABLE V
RESULTS OF WORD CORRECTIONS IN ENGLISH.

Misspelling	Correct Spelling	Aspell	Microsoft word	Google	MultiSpell
Abberration	aberration	aberration	aberration	aberration	aberration
accomodation	accommodation	accommodation	accommodation	accommodation	accommodation
acheive	achieve	Achieve	achieve	achieve	achieve
abortificant	abortifacient	**aficionados**	-	abortifacient	abortifacient
absorbsion	absorption	**absorbs ion**	absorpsion	absorption	absorption
ackward	(awkward, backward)	awkward	(awkward, backward)	awkward	(awkward, backward)
additinally	additionally	additionally	additionally	additionally	additionally
adminstration	administration	administration	administration	administration	administration
admissability	admissibility	admissibility	admissibility	admissibility	admissibility
advertisments	advertisements	advertisements	advertisements	advertisements	advertisements
adviced	advised	advised	advised	**advice**	**advice**
afficionados	aficionados	aficionados	aficionados	aficionados	aficionados
affort	(effort ,afford)	effort	afford	afford	afford
agains	against	**agings**	**agings**	against	against
aggreement	agreement	agreement	agreement	agreement	agreement
agressively	aggressively	aggressively	aggressively	aggressively	aggressively
agriculturalist	agriculturist	-	-	-	agriculturist
alcoholical	alcoholic	alcoholically	**alcoholically**	alcoholic	alcoholic
algebraical	algebraic	algebraic	**algebraically**	algebraic	**algebraically**
algoritms	algorithms	algorithms	algorithms	algorithms	algorithms
alterior	(ulterior , anterior)	ulterior	(anterior, ulterior)	ulterior	(anterior, ulterior)
anihilation	annihilation	annihilation	annihilation	annihilation	annihilation
anthromorphization	anthropomorphization	**anthropomorphizing**	-	-	anthropomorphization
bankrupcy	bankruptcy	bankruptcy	bankruptcy	bankruptcy	bankruptcy
baout	(about,bout)	bout	(about,bout)	about	bout
basicly	basically	basically	basically	basically	basically
breakthough	breakthrough	**break though**	breakthrough	breakthrough	breakthrough
carachter	character	**crocheter**	character	character	character
cannotation	connotation	connotation	(connotation ,annotation)	connotation	(connotation ,annotation)
carismatic	charismatic	charismatic	charismatic	charismatic	charismatic
carmel	caramel	**Carmel**	-	-	caramel
cervial	(cervical, servile)	cervical	cervical	cervical	cervical
clasical	classical	classical	classical	classical	classical
cleareance	clearance	clearance	clearance	clearance	clearance
comissioning	commissioning	commissioning	commissioning	commissioning	commissioning
commemerative	commemorative	commemorative	commemorative	commemorative	commemorative
compatabilities	compatibilities	compatibilities	compatibilities	compatibilities	compatabilities
committment	commitment	commitment	commitment	commitment	commitment
debateable	debatable	debatable	debatable	debatable	debatable
determinining	determining	determinining	determinining	determinining	determining
childbird	childbirth	**child bird**	**child bird**	childbirth	childbirth
definately	definitely	definitely	definitely	definitely	definitely
decribe	describe	describe	describe	describe	describe
elphant	elephant	elephant	elephant	elephant	elephant
emmediately	immediately	immediately	immediately	immediately	immediately
emphysyma	emphysema	emphysema	emphysema	emphysema	emphysema
erally	(orally, really)	orally	really	really	orally
eyasr	(years, eyas)	**eyesore**	years	years	eyas
facist	fascist	fascist	fascist	fascist	fascist
fluoroscent	fluorescent	fluorescent	fluorescent	fluorescent	fluorescent
geneology	genealogy	genealogy	genealogy	genealogy	genealogy
gernade	grenade	grenade	grenade	grenade	grenade
girates	gyrates	**grates**	gyrates	**pirates**	gyrates

Misspelling	Correct Spelling	Aspell	Microsoft word	Google	MultiSpell
gouvener	governor	governor	**souvenir**	**gouverneur**	**convener**
gurantees	guarantee	guarantee	guarantee	guarantee	guarantee
guerrila	(guerilla, guerrilla)	guerrilla	guerrilla	guerrilla	(guerilla, guerrilla)
guerrilas	(guerillas, guerrillas)	guerrillas	guerrillas	guerrillas	(guerillas, guerrillas)
Guiseppe	Giuseppe	Giuseppe	Giuseppe	Giuseppe	Giuseppe
habaeus	(habeas, sabaeus)	habeas	**habitués**	habeas	**sabaeus**
hierarcical	hierarchical	hierarchical	hierarchical	hierarchical	hierarchical
heros	heroes	heroes	heroes	heroes	**herbs**
hypocracy	hypocrisy	hypocrisy	hypocrisy	hypocrisy	hypocrisy
independance	Independence	Independence	-	Independence	Independence
intergration	integration	integration	integration	integration	integration
intrest	interest	interest	interest	interest	interest
Johanine	Johannine	Johannes	Johannes	Johannes	Johannine
judisuary	judiciary	judiciary	judiciary	-	judiciary
kindergarden	kindergarten	kindergarten	kindergarten	kindergarten	kindergarten
knowlegeable	knowledgeable	knowledgeable	knowledgeable	knowledgeable	knowledgeable
labatory	(lavatory, laboratory)	(lavatory, laboratory)	(lavatory, laboratory)	laboratory	(lavatory, laboratory)
lonelyness	loneliness	loneliness	loneliness	loneliness	loneliness
legitamate	legitimate	legitimate	legitimate	legitimate	legitimate
libguistics	linguistics	linguistics	linguistics	linguistics	linguistics
lisence	(license, licence)	licence	**silence**	licence	licence
mathmatician	mathematician	mathematician	mathematician	mathematician	mathematician
ministery	ministry	ministry	ministry	ministry	ministry
mysogynist	misogynist	misogynist	misogynist	misogynist	misogynist
naturaly	naturally	naturally	naturally	naturally	naturally
ocuntries	countries	countries	countries	countries	countries
paraphenalia	paraphernalia	paraphernalia	paraphernalia	paraphernalia	paraphernalia
Palistian	Palestinian	**Alsatain**	**politian**	Palestinian	Palestinian
pamflet	pamphlet	pamphlet	pamphlet	pamphlet	pamphlet
psyhic	psychic	psychic	psychic	psychic	psychic
Peloponnes	Peloponnesus	Peloponnese	Peloponnese	Peloponnese	Peloponnesus
personell	personnel	personnel	personnel	personnel	personnel
posseses	possesses	possesses	possesses	possesses	possess
prairy	prairie	**priory**	prairie	prairie	**airy**
qutie	(quite, quiet)	quite	quite	**cutie**	**queue**
radify	(ratify,ramify)	ratify	ratify	ratify	ramify
reccommended	recommended	recommended	recommended	recommended	recommended
reciever	receiver	receiver	receiver	receiver	**reliever**
reconaissance	reconnaissance	reconnaissance	reconnaissance	reconnaissance	reconnaissance
restauration	restoration	restoration	restoration	restoration	**instauration**
rigeur	(rigueur, rigour, rigor)	**rigger**	rigueur	-	(rigueur, rigour)
Saterday	Saturday	Saturday	Saturday	Saturday	Saturday
scandanavia	Scandinavia	Scandinavia	Scandinavia	Scandinavia	Scandinavia
scaleable	scalable	scalable	-	scalable	scalable
secceeded	(seceded, succeeded)	succeeded	succeeded	seceded	succeeded
sepulchure	(sepulchre, sepulcher)	sepulcher	**sepulchered**	sepulcher	sepulchre
themselfs	themselves	themselves	themselves	themselves	themselves
throught	(thought, through, throughout)	(thought, through)	(thought ,through)	**throat**	(thought ,through, throughout)
troups	(troupes, troops)	(troupes, troops)	troupes	troops	troops
simultanous	smultaneous	simultaneous	simultaneous	simultaneous	simultaneous
sincerley	sincerely	sincerely	sincerely	sincerely	sincerely
sophicated	sophisticated	**suffocated**	**supplicated**	-	sophisticate
surrended	(surrounded, surrendered)	surrounded	surrender	surrender	surrounded
unforetunately	unfortunately	unfortunately	unfortunately	-	unfortunately
unnecesarily	unnecessarily	unnecessarily	unnecessarily	-	unnecessarily
usally	usually	usually	usually	usually	usually
useing	using	using	using	using	**seeing**
vaccum	vacuum	vacuum	vacuum	vacuum	vacuum

Misspelling	Correct Spelling	Aspell	Microsoft word	Google	MultiSpell
vegitables	vegetables	vegetables	vegetables	vegetables	vegetables
vetween	between	between	between	between	between
volcanoe	volcano	volcano	volcano	volcano	volcano
weaponary	weaponry	weaponry	weaponry	weaponry	weaponry
worstened	worsened	worsened	worsened	-	worsened
wupport	support	support	support	support	support
yeasr	years	years	years	years	yeast
Yementite	(Yemenite, Yemeni)	Yemenite	Yemenite	Yemenite	Yemenite
yuonger	younger	Younger	younger	younger	**sponger**

TABLE VI
RESULTS OF WORD CORRECTIONS IN PORTUGUESE.

Correct Form	Spelling Error	TST	Aspell	MultiSpell
acerca	àcerca	acerca	acerca	acerca
açoriano	açoreano	açoriano	**coreano**	açoriano
alcoolémia	alcoolemia	**alcoolÚmia**	-	alcoolémia
ameixial	ameixeal	ameixial	ameixial	ameixial
antárctico	antártico	catártico	antárctico	antárctico
antepor	antepôr	-	antepor	antepor
árctico	artico	**artigo**	**aórtico**	**aórtico**
artífice	artífece	artífice	artífice	artífice
bainha	baínha	bainha	bainha	bainha
bebé	bébé	bebé	**bebe**	bebé
bege	beje	bege	**beije**	**bejense**
bênção	benção	**bençao**	-	bênção
benefcência	benefciência	beneficência	beneficência	beneficência
biopsia	biópsia	**biópsiu**	-	biopsia
burburinho	borborinho	burburinho	burburinho	burburinho
caiem	caem	-	-	**cabem**
calvície	calvíce	calvície	calvície	calvície
camoniano	camoneano	camoniano	camoniano	camoniano
campeão	campião	campeão	campeão	campeão
chiita	xiita	chiita	xiitas	**xiitas**
comboio	combóio	comboio	comboio	comboio
compor	compôr	-	compor	compor
comummente	comumente	**comovente**	comummente	comummente
constituia	constituía	-	-	constituia
constituiu	constituíu	constituiu	constituiu	constituiu
cor	côr	-	cor	cor
crânio	crâneo	crânio	**cárneo**	crânio
definição	defenição	definição	definição	definição
definido	defenido	definido	-	defendido
definir	defenir	definir	definir	definir
desequilíbrio	desequilibrio	desequilíbrio	desequilíbrio	desequilíbrio
despretensioso	despretencioso	despretensioso	despretensioso	despretensioso
dignatários	dignitários	dignatários	**digitarias**	dignatários
dispender	despender	dispender	-	**despendes**
dispêndio	dispendio	**dispundio**	**dispundio**	dispendioso
ecrã	ecran	-	écran	écran
emirados	emiratos	**estratos**	**méritos**	emirados
esotérico	isotérico	-	-	esotérico
esquisito	esquesito	esquisito	esquisito	esquisito
estratego	estratega	estratego	-	estratego
feminino	femenino	feminino	feminino	feminino
feminismo	femininismo	-	feminismo	feminismo
fôr	for	-	-	**forcar**
gineceu	geneceu	gineceu	gineceu	gineceu
gorjeta	gorgeta	gorjeta	gorjeta	gorjeta
granjear	grangear	granjear	granjear	granjear
guisar	guizar	guisar	**gizar**	**guinar**
halariedade	hilaridade	hilariedade	-	**polaridade**
hectare	hectar	hectare	-	hectare
hiroshima	hiroxima	**aproxima**	**próxima**	hiroshima
ilacção	elação	ilação	ilação	**delação**
indispensável	indespensável	indispensável	indispensável	indispensável
inflacção	inflação	-	-	**inalação**
interveio	interviu	**intervir**	Inter viu	**intervim**
intervindo	intervido	intervindo	-	intervindo
invocar	evocar	invocar	-	**evocai**

Correct Form	Spelling Error	TST	Aspell	MultiSpell
ípsilon	ipslon	ípsilon	ípsilon	ípsilon
irisar	irizar	irisar	**razar**	irisar
irupção	irrupção	-	-	irupção
jeropiga	geropiga	jeropiga	**Georgia**	jeropiga
juiz	juíz	-	juiz	Juiz
lampião	lampeão	lampião	sarjeta	**campeão**
lêem	lêm	**lês**	**lema**	lêem
linguista	linguísta	-	linguista	linguista
lisonjear	lisongear	lisonjear	lisonjear	lisonjear
logótipo	logotipo	**logo tipo**	**logo tipo**	logótipo
maciço	massiço	mássico	mássico	**massudo**
majestade	magestade	majestade	majestade	majestade
manjerico	mangerico	manjerico	manjerico	manjerico
manjerona	mangerona	**tangerina**	**tangerina**	manjerona
meteorologia	metereologia	meteorologia	meteorologia	meteorologia
miscigenação	miscegenação	miscigenação	miscigenação	miscigenação
nonagésimo	nonagessimo	nonagésimo	nonagésimo	nonagésimo
oceânia	oceania	oceânia	**Oceania**	oceânia
oficina	ofecina	oficina	oficina	oficina
opróbrio	opróbio	**aeróbio**	**próbio**	opróbrio
organograma	organigrama	organograma	-	organograma
paralisar	paralizar	paralisar	paralisar	paralisar
perserverança	preseverança	perserverança	perserverança	perseverance
persuasão	persuação	persuasão	persuasão	persuasão
pirinéus	pirenéus	-	pirinéus	pirinéus
pretensioso	pretencioso	pretensioso	pretensioso	pretensioso
privilégio	previlégio	privilégios	privilégios	privilegios
quadricromia	quadricomia	quadricromia	**quadriculai**	quadricromia
quadruplicado	quadriplicado	quadruplicado	quadruplicado	quadruplicado
quasímodo	quasimodo	-	**quisido**	quasímodo
quilo	kilo	quilo	Nilo	**dilo**
quilograma	kilograma	**holograma**	**holograma**	**holograma**
quilómetro	kilómetro	**milímetro**	**milímetro**	quilómetro
quis	quiz	quis	**qui**	**juiz**
rainha	raínha	rainha	rainha	rainha
raiz	raíz	-	raiz	raiz
raul	raúl	raul	Raul	raul
rectaguarda	retaguarda	rectaguarda	-	rectaguarda
rédea	rédia	rédea	**radia**	**radia**
regurgitar	regurjitar	regurgitar	regurgitar	regurgitar
rejeitar	regeitar	rejeitar	**regatar**	**receitar**
requeiro	requero	requere	requeiro	requer
réstia	réstea	réstia	**resta**	réstia
rubrica	rúbrica	**rúbreca**	rubrica	rubrica
saem	saiem	**saiam**	saem	**caiem**
saloiice	saloice	**baloice**	saloiice	saloiice
sarjeta	sargeta	sarjeta	sarjeta	Sarjeta
semear	semiar	semear	semear	Semear
suíça	suiça	suíça	suíça	Suíça
supor	supôr	-	supor	Supôs
trânsfuga	transfuga	**transfira**	**transfira**	trânsfuga
transpôr	transpor	-	-	transportar
urano	úrano	-	-	**grano**
ventoinha	ventoínha	ventoinha	ventoinha	ventoinha
verosímil	verosímel	-	-	verosímil
vigilante	vegilante	vigilante	vigilante	vigilante
vôo	voo	-	-	**ovo**
vultuoso	vultoso	vultuoso	-	**vultosos**
xadrez	xadrês	xadrez	**ladres**	xadrez
xamã	chamã	chama	chama	chamá
xelindró	xilindró	**cilindro**	**cilindro**	xelindró
zângão	zangão	zangai	-	**mangão**
zepelin	zeppelin	**zepelim**	**zeplim**	zepelin
zoo	zoô	zoo	**coo**	zoo

REFERENCES

[1] K. Kukich, "Techniques for automatically correcting words in text," *ACM Computing Surveys*, 24(4), 377–439, 1992.

[2] F. J. Damerau, "A technique for computer detection and correction of spelling errors," *Communications of ACM*, 7(3):171-176.7, 1964.

[3] W. Peters, "Lexical Resources," NLP group, Dept. of Comp. Sc., Uni. of Sheffield, 2001.

[4] R. A. Wagner and M. J. Fisher, "The string to string correction problem," *Journal of Assoc. Comp. Mach*., 21(1):168-173, 1974.

[5] A. Stanier, "How accurate is Soundex matching?" *Comp. in Genealogy*, vol. 3:7, 1990.

[6] C. Fellbaum, "WordNet, an electronical lexical database," Cambridge, MIT Press, 1998.

[7] E. Pianta, L. Bentivogli, and C. Girardi, "MultiWordNet: developing an aligned multilingual database," in *Proc. of 1st Int. Conf. on Global WordNet*, 2002.
[8] C. E. Shannon, "Prediction and entropy of printed English," *Bell Sys. Tec. J.* (30):50–64, 1951.
[9] U. Pfeifer, "Retrieval Effectiveness of Proper Name Search Methods," *Information Processing and Management*, 32(6):667–679, 1996.
[10] E. J. Yannakoudakis and D. Fawthrop, "An intelligent spelling error corrector," *Information Processing and Management,* 19:1, 101-108, 1983.
[11] T. N. Turba, "Length-segmented lists," *Comm. of the ACM*, 25:8, pp 522-526, 1982.
[12] Wikipedia, list of Common Misspelling Word List, http://en.wikipedia.org/wiki/Wikipedia: List_of_common_misspellings, 05.10.2006.
[13] B. Martins, M. J. Silva, "Spelling Correction for Search Engine Queries," in *EsTAL - España for Natural Language Processing*, Alicante, Spain, 2004.
[14] K. Church and W. A. Gale, "Probability scoring for spelling correction," *Statistics and Computing*, Vol. 1, No. 1, pp. 93–103, 1991.
[15] V. J. Hodge and J. Austin, "A comparison of standard spell checking algorithms and novel binary neural approach," *IEEE Trans. Know. Dat. Eng.*, Vol. 15:5, pp. 1073-1081, 2003.
[16] J. J. Pollock and A. Zamora, "Collection and characterization of spelling errors in scientific and scholarly text," *Journal Amer. Soc. Inf. Sci*., Vol. 34, No. 1, pp. 51–58, 1983.
[17] ———, "Automatic spelling correction in scientific and scholarly text," *Comm. ACM*, Vol. 27, No. 4, pp. 358–368, 1984.
[18] E. Brill and R. C. Moore, "An improved error model for noisy channel spelling correction," in *Proc. 38th Annual Meet. of the Assoc. for Comp. Ling*., Hong Kong, 2000, pp. 286–293.
[19] K. Toutanova and R. C. Moore, "Pronunciation modeling for improved spelling correction," in *Proc. 40th Annual Meeting of the Assoc. for Comp. Ling*., Hong Kong, 2002, pp. 144–151.
[20] Jin-ming Zhan, Xiaolong Mou, Shuqing Li, Ditang Fang, "A Language Model in a Large-Vocabulary Speech Recognition System," in *Proc. of Int. Conf. ICSLP98*, Sydney, Australia, 1998.
[21] S. Deorowicz and M. G. Ciura, "Correcting Spelling Errors by Modelling Their Causes," *Int. Journal of Applied Mathematics and Computer Science,* 15(2):275–285, 2005.
[22] B. Khaltar, A. Fujii, and T. Ishikawa, "Extracting loanwords from Mongolian corpora and producing a Japanese-Mongolian bilingual dictionary," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL,* Sydney, Australia: ACL. Pages: 657 – 664, 2006.
[23] E. W. De Luca and A. Nürnberger, "Using Clustering Methods to Improve Ontology-Based Query Term Disambiguation," *International Journal of Intelligent Systems*, 21:693–709, 2006.
[24] E. W. De Luca and A. Nürnberger, "Rebuilding Lexical Resources for Information Retrieval using Sense Folder Detection and Merging Methods,"in: *Proc. of the 5th Int. Conf. on Language Resources and Evaluation (LREC 2006)*, 2006.
[25] E. W. De Luca, "Semantic Support in Multilingual Text Retrieval," Shaker Verlag, Aachen, Germany, 2008.

Bilingual Lexical Data Contributed by Language Teachers via a Web Service: Quality vs. Quantity

Valérie Bellynck, Christian Boitet, and John Kenwright

***Abstract*—IToldU is a light web service which, in its first year of use for teaching technical English in French engineering schools, has enabled the contribution of just over 17000 English terms in about twenty technical domains. These terms are associated with their French translations (95% of which are correct) and examples of use (about 85% correct). In the second year, emphasis has been on quality rather than on quality: about 6000 high-quality entries have been contributed by the same number of students and classes. Some desirable extensions are in progress, e.g. to add English when this language is not included in the original language pair, and to synchronize with off-line contributions prepared on a PDA or a hand-held calculator.**

***Index Terms*—Collaborative dictionary construction, examples of use, technical English teaching**.

I. Introduction

THE collaborative construction of free lexical resources has been hampered by the difficulty of obtaining many individual small and voluntary contributions. IToldU (Interactive Technical On-Line Dictionary for Universities) is a light web service which can be used for the collaborative construction of a bilingual lexicon by a small community (typically, a group of students) while learning a foreign language in technical or specific domains. Contributions are freely offered, but are also constrained in that part of the students' English grades are computed by IToldU itself.

For the first two authors, the initial objective in building this site was to collect the produced lexica in order to populate the multi-usage multilingual lexical database (MLDB) Papillon (see http://www.papillon-dictionary.org/). For the third author, an English teacher of ICTE (Information and Communication Techniques for Education) at INPG (Institut Polytechnique de Grenoble), the objective was to improve the teaching of technical English vocabulary to French engineering students.

Manuscript received November 25, 2008. Manuscript accepted for publication August 15, 2009.

Valérie Bellynck is with équipe STG, LGP2 461 rue de la Papeterie, BP 65, 38402 Saint-Martin-d'Hères, France (e-mail: Valerie.Bellynck@efpg.inpg.fr).

Christian Boitet is with équipe GETA, laboratoire CLIPS, 385 rue de la Bibliothèque, BP 53, 38041 Grenoble Cedex 9, France (e-mail: Christian.Boitet@imag.fr).

John Kenwright is with Cellule TICE, bureau 2.12, 701 rue de la Piscine, BP 81, 38402 Saint-Martin-d'Hères, France (e-mail: John.Kenwright@inpg.fr).

In its current state, IToldU addresses mainly the instructional objective rather than the lexicographical one. Moreover, its use has led to a third interesting possibility, that of teaching the structure of simple sentences of English through examples in use: it turns out that students are not satisfied with copying and pasting sentences containing the terms they translate, but prefer to create their own examples.

In the following sections, we will: present IToldU; evaluate its first two full years of use (describing its pedagogical impact on students and teachers and the quantitative and qualitative lexicographical results obtained when varying the desired quality level); and describe plans for increasing contributions, for extending collection to other languages and types of information, and for synchronization with the Papillon online multilingual lexical database.

II. The IToldU Web Service

A. Teaching Context and Goals

The teaching context is as follows:

- Acquiring and using technical English.
- The most important translation direction is English-French.
- Students don't yet know the technical terms in English and have only recently encountered them in French.
- There are probably 10,000-20,000 terms with which the teacher is not necessarily familiar (either in French or in English).
- The teaching goals of the English courses, over the three years spent in the schools by students, are twofold:
- The base technical vocabulary that is to be learned by all students represents about 10% (1000-2000 items) of the terms.
- Each student should choose and learn a small fraction of the remaining 90%.

Students know how to use between 150-300 specific English words or terms associated with their technical field (paper industry) by the time they leave in the third year. Of course, they know many more general terms, and terms in all other domains encountered during their courses (including other technical fields, work placements, themes and skills seen in traditional English classes, job hunting, etc.).

B. Initial Requirements

During the English courses, each student must collect or create the lexical data for his or her own dictionary, based on texts or other sources given by the teacher. Other words or findings encountered during pursuit of language acquisition can also be added. Students can choose from existing found examples and can correct or create their own. Contributing a translation or selecting an existing example generates a vote for the responsible student.

Teachers and students can restrict their views to the elements most useful to them: students and visitors can search for, create, and memorize translations of technical (or thematic) English expressions, and teachers can run quantitative statistics, control student contributions, and enliven the site using "word hunts," etc. The coordinating teacher is the only one allowed to manage the site (through lists of teachers, students, classes, etc.).

The objective of collecting lexical data is not mentioned to the students and teachers, who are only aware of the pedagogical objectives enunciated by the coordinator:

- Motivating the students to do "lexical" work outside of the class room,
- Minimizing the supplementary workload of the teachers.

C. Implementation

IToldU associates a MYSQL database with each group of students for their three years at the school. It contains the teachers, the students, and the groups of students, with their access rights. It also contains the current dictionary of the group, with students associated with created or adopted entries.

IToldU is programmed in HTML/SQL/PHP, and installed on a free Internet provider (laposte.net, then grenet.fr). It is easy to clone, to install on other sites, and to adapt to other languages, because all messages and menu items are contained in text resources, and can be edited without any special knowledge of programming.

Users have passwords and access rights. The global parameters can only be set by the coordinating teacher. Other teachers can consult students' accounts and direct them. Students can only capture data and consult their personal dictionaries and the dictionary of their group.

D. Usage by Students

Students must seek technical expressions in English and propose correct French translations. For each term, they must include (by citation or creation) an example in context and its source (e.g. from class, booklets, lab sessions, magazines, press, or web or bibliographic sources).

In the examples, the interface is in French, because the students are French speakers. But, as said above, IToldU is easily localizable to other interface languages.

When a student connects to his or her own digital dictionary, he or she finds a summary (Fig. 1) page providing access to the digital dictionary (to search for translations and add new expressions). Also on the page are useful teachers' tools ("Outils") for preparing CVs, application letters, or word hunts. A user can look at his own statistics, measure his knowledge against that of fellow classmates, or print the current dictionary (Fig. 5).

The current access form is minimal: one can only enter an expression or the first letters of an expression in the first input field. However, it has been designed to be easily replaced or combined with richer ones later.

If there is no entry for a word or expression, the student should enter a translation proposal, with an example of use, the context where it was found, and its bibliographical reference. Each voluntary contribution by a student counts toward its statistics and grades.

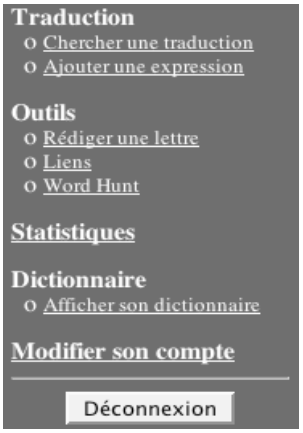


Fig. 1. Students summary.

The principle used for motivating the students and regulating their contributions is simple: the student begins by checking, before introducing a term of interest, whether it has already been handled by a groupmate.

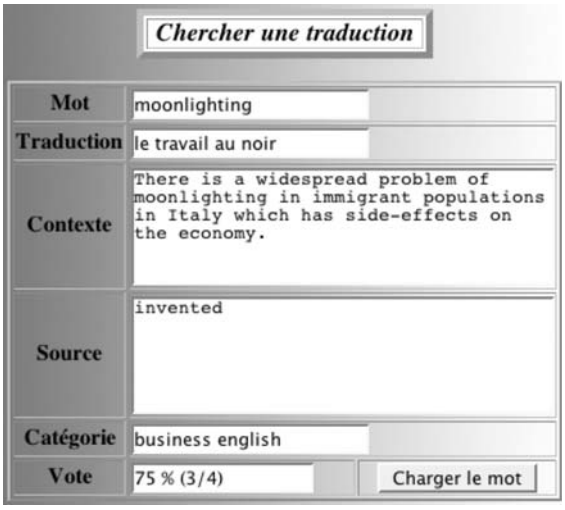


Fig. 2. Form for adding a term in IToldU.

If so, and if the translation and the example look acceptable, s/he can (but does not have to) "adopt" it by adding it to his/her personal dictionary. S/he can also create a new entry, Fig. 2.

Students receive a point for uploading an entry onto their dictionary (effectively "voting" for it). However, if the entry is wrong, the student will lose a point later. In both cases, IToldU motivates students via the possibility of gaining or losing points. This incentive instills in them a positive learning attitude. Moreover, the publication of the "top ten" best scores on the web site motivates them to participate more and more often, creating a healthy competitiveness among individuals and groups.

E. Teachers

IToldU offers teachers the possibility of supervising student groups, encouraging involvement through the use of bonus marks, and livening up vocabulary acquisition via playful "word hunts". Fig. 3 shows the summary of a teacher's session.

S/he can customize general properties (e.g. the title of the site, or its language), broadcast learning activities, contribute to the digital dictionary's construction (by searching for a translation, adding a new expression and creating new technical domains – called "categories"), manage student groups ("*Gestion des comptes*" – account management), and look at the contribution of each student or classroom, as shown in Fig. 4 and Fig. 5 ("*Statistiques*", "*Afficher un dictionnaire*" – display a dictionary).

Fig. 3. Teachers' summary.

A particular blessing is that teachers never have to look inside the source code of an HTML page or (even worse!) other program code. Another important point is that the time constraints of the teachers are taken into account: teachers have almost no time to follow students' work outside the classroom (perhaps 1-2 minutes per student). The use of IToldU should not increase their work time, but if possible reduce it.

That seems to be the case now, as the grading system has been designed to optimize the teacher's time. During the first few weeks of use by a new group, the teacher systematically goes online and deletes any incorrect words. This supervision encourages rigor at of the start of the program.

Fig. 4. Resource pooling statistics.

During the second year, evaluations of contributions are scheduled (every five months) in which teachers check a few dictionary samples from each student in their class. Students don't know which sample will be checked, and are hence motivated to check and improve their entire dictionary. Owing to lack of time and for pedagogical reasons, teachers do not correct mistakes, but simply mark that a translation or an example is wrong. IToldU supports such error marking on fields. Then students must make the corrections before a certain time elapses, or IToldU will subtract the corresponding points.

Fig. 6 shows an example of a "word hunt" screen. "Word hunt" is a challenging but enjoyable part of IToldU for both teachers and students. The first student to find a translation wins a point! Thus students log on as often as possible to see if there are words up for grabs!

III. Evaluation

A. Pedagogical Aspects

Reactions of teachers and students. The current complete version of IToldU (http://opus.grenet.fr/itoldu/ITOLDU) was used for the first time in 2004-05 by all the students of EFPG, an engineering school that is attached to INPG, with a clear positive pedagogical impact. A total of 250 students were involved in the beta test, spread out over the three years of engineering school and one year of professional BA (licence) work. As far as English teaching was concerned, there were 17 groups, 6 teachers, and 1 coordinating teacher (the third author).

IToldU already addresses quite well the need felt by the coordinating teacher for a computer tool improving management of training, teachers' work, and students' learning of specialized English technical vocabulary.

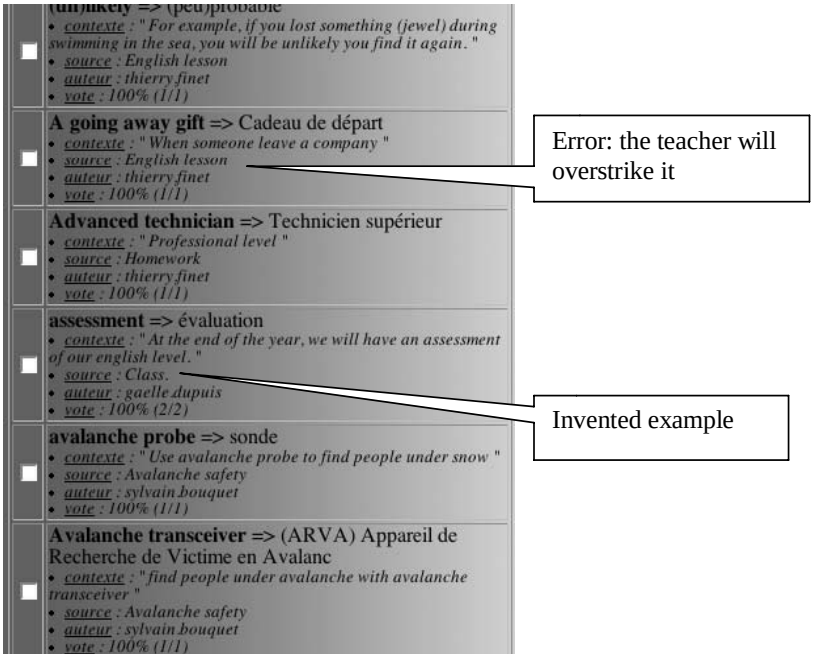


Fig. 5. Fragment of the (sub)dictionary of a class.

perks	Avantages	gaelle.dupuis	DUTpromo13
jobless	Au chômage	gaelle.dupuis	DUTpromo13
Employment agency	Agence de placements	thierry.finet	DUTpromo13
nine-to-five job		Ajouter	
Hire and fire		Ajouter	
Corporate culture		Ajouter	
Long-hours culture		Ajouter	
Casual Friday		Ajouter	
going rate		Ajouter	
cash in hand		Ajouter	
job with scope		Ajouter	

Fig. 6. Word hunt prepared by a teacher.

The use of IToldU has changed the behavior of most students for the better: they are more interested in taking notes. Further, using IToldU outside of classes is seen as a supplementary learning process in the acquisition of technical English vocabulary, and not only as a receptacle into which students are forced to put translations and examples, and which they later ignore.

IToldU not only motivates students by computing part of their grade as a function of their (correct) use of the site. It also allows teachers to establish a spirit of cooperation and emulation among students. On the one hand, as we have seen, students cooperate by "voting" for those whose entries they adopt. On the other hand, the system shows the students who have contributed most on a "scoreboard". Finally, word hunts give rise to a healthy and playful emulation.

Students now consider the long term, because they know they will be allowed to take ITOLDU with them in their professional life as an active copy of their personal dictionary (which can be installed and maintained on a Web site). If they wish, they can take along the entire dictionary built by their classmates.

However, it must be noted that not all teachers were as involved in the adoption and use of IToldU as the third author due to the difficulty of working conditions and lack of time; hence the inequality of the contributions of different classes.

Contributive aspects. The problem of motivating students to contribute and of automatically regulating the global contribution process is a particular case of a more general problem widely recognized as very difficult: that of motivating voluntary and free contributions to the population of knowledge bases. That problem is difficult because there are very few specialists in any field who are willing to give their hard-won the knowledge without return or reward.

Beyond such of rare contributions (which, even if they are large for individuals, represent only a small fraction of the desired knowledge), it is necessary to rely on large numbers of non-specialists, each contributing small, and even fragmentary, knowledge elements. However, in reality, it has always been difficult to obtain numerous individual voluntary and free contributions from a "community of interest".

If contributors gain something by contributing, then the contribution is not "free" in the strictest sense. For example, translators using Oki Electric http://www.yakushite.net/ web site put words in dictionaries because they use freely available online tools for translators (bilingual editor, online dictionaries, proposals from translation memories and from the MT system Pensée), in which contributed words become almost instantaneously active.

If, on the other hand, contributions are truly free, contributors are motivated in some way – of course, as discreetly and pleasantly as possible. That is the case of IToldU, in which almost all users – both teachers and students – are "strongly invited" to use the tool.

B. Dictionary Evaluation (First Year)

1) Quantitative aspect

In the first semester, about 12,000 English-French entries were entered into IToldU by the students, along with about 8,000 usage contexts.

At the end of the academic year, IToldU contained 17,062 English-French entries, and about as many usage contexts (only 157 entries lacked contexts).

2) Qualitative aspect

The second author quickly revised all the contributions of the first year, and about 10% in detail, thereby correcting them. Apart from errors arising from problems in inputting diacritics on the Web, the French translations of English terms are almost all correct. By contrast, 15% to 20% of usage contexts are not examples of use. Following are some details on these two types of contribution.

Translations. 95% of the translations seem correct to us. An interesting point is that only about 30% of the English terms chosen by the students concern a purely technical lexical field, one linked with students' studies (of manufacturing paper pulp, paper, cardboard, color processing, inks, rheology, etc.) while 70% concern "paratechnical" fields, such as business or job hunting, or general English.

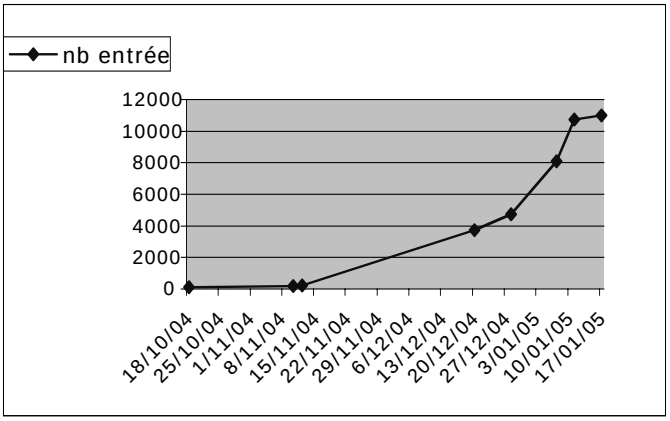


Fig. 7. Evolution of the number of entries in the first semester.

From usage contexts to examples of use. "Contexts" merit some comments. In the mind of the teachers, contexts should be citations of sentences in which the English terms had been encountered. But several unexpected things happened.

Certain students understood that they were being asked for the "domain" of the citation, selected from a list provided by IToldU. One finds for example:

5024	opportunity	possibilité, débouché	society
5025	to put up	ériger, construire	society
5026	to fulfill	accomplir, réaliser	society
5027	fulfilling	profondément, satisfaisant	society
15009	gas-fired	chauffé au gaz	used in paper mill

Others thought that they were being asked for definitions.

15049	a wind mill	une éolienne	an energy-producing facility
15065	a light bulb	une ampoule électrique	energy-related equipment
4632	TCF (totally chlorine free)	sans chlore	stade de blanchiment

The coordinator accordingly modified his description: he asked for "examples in use", and created some himself, putting "invented" in the source field. The students then understood that they, too, could invent examples, and did so. At the level of content, several cases arose:

- Some students created or adapted sentences containing the English terms in question, but in such a way that the word meaning could not be discriminated.

16070	collude	s'associer	they colluded last year
16990	telematics	télématique	it s telematics
16998	darts	fléchettes	he throws the darts
17003	potoling	spéléologie	the potoling is dangerous
17006	chiari-oscuro	clair-obscur	the is a chiarioscuro effect
17026	heir	héritier	you heir to your mother

- At the other extreme, other students used long sentences as examples.

12956	Falsification	Falsification	Some various documents to be protected from counterfeiting and falsification like service vouchers, security label and certificates of authenticity have special features.
12957	service vouchers	Tickets de prestation	
12958	security label	Etiquettes sécurisées	
12959	certificates of authenticity	Certificats d' authenticité	
12960	anti-counterfeiting features	Eléments anti-contrefaçon	
12961	anti-falsification feature	Eléments anti-falsification	

- Many proposals are intended as "honest examples", but are not in correct English.

6619	carriageway	chaussée	the carriageway is destroy by the cars
7073	union	syndicat	an union for help employees
7098	pythonesque	humour absurde	this joke are very pythonesque with his very absurd humor
9183	(to) insulate	isoler	insulating materials can be very useful in electronic

- A small percentage of students vented their frustration by putting "garbage" (silly examples or obscenities) in their examples.

In total, about 15% of the examples are incorrect with respect to content, again not counting input errors, and many

more are incorrect with respect to language, grammar, and spelling.

Hence, there is the origin of the idea to use IToldU not only for learning vocabulary, but also for language learning. Interestingly, students used some of these examples in class during oral performance.

IV. Perspectives

A. Encouraging More Contributions

Other possible ways to encourage more contributions:

- Generalize the "scoreboard" idea to show credits for each entry part.
- Introduce personalization facilities (i.e. automatic or semi-automatic user profiling), so that the system can suggest personalized lists of "things-to-do" or new contributions in the user's domain of interest.
- Allow users to self-organize in groups and groups of groups, each group having certain access rights and a profile.
- Give users access to tools that can extract potential translation pairs from related corpora (texts on the same domain in two or more languages, usually not parallel).
- Let users contribute directly through an "active reading" interface (translated words or idioms appear in annotations of read text).
- Make the importing environment accessible to users wishing to upload sets of translation pairs from any format (Excel, Word, FileMaker, XML, etc.).
- As the ultimate objective, integrate the lexical contribution function as an add-on (plug-in) in as many applications as possible, to be used by the general public.

B. Synchronize Papillon with IToldU

Since the Papillon platform (in particular, its CDM part) accepts any kind of dictionary, provided it is formatted in XML and can be mapped to the CDM DTD, the first problem in linking IToldU and Papillon is to define the mapping of information: are IToldU entries Papillon "lexies", or lemmas, or *vocables*? As seen in the examples above, they are in fact only *vocables* – citation forms without any disambiguating part-of- speech tags.

The second problem is maintenance: the periodic updating of information from IToldU in Papillon.

The fact that the information can be modified under Papillon as well as under IToldU should not be a major problem, as Papillon is designed to keep the contributions of each contributor in his or her private work space, and to allow the creation of groups of contributors. It should then suffice to create one IToldU contributor. Alternatively, if one wishes to keep track of the student contributors in Papillon, one could create a Papillon user for each IToldU student. Papillon groups would correspond to IToldU classes, with one main group for IToldU itself.

The basic idea for maintenance, found to be valid in other contexts, is to compute the differences between two successive states of the IToldU database, and then to compute an update program which can be executed by the Papillon API as if modifications had been made interactively using the Papillon web interface.

C. Extension to Other Language Pairs or Triples

Nothing in IToldU is specific to the English-French language pair, and the software is easy to localize: a language teacher with no programming skills can do it by editing text files.

However, one necessary change is that IToldU should be able to handle three languages in parallel (thereby integrating a second foreign language that a student may also be studying as a course requirement): the two languages used in the classroom and English if it is not one of these.

D. Other Information Types

In the current context of engineering schools, it does not seem possible to obtain sophisticated types of information beyond the lexicographical, such as DiCo semantic formula, definitions, regimes[1], lexico-semantic functions, and other types of collocations. Perhaps the parts-of-speech could be contributed by our students, but nothing more.

Hence, we are trying to find other learning contexts in which such advanced information types are more likely to be contributed by users, such as language schools and translation or interpretation schools.

V. Conclusion

The collaborative construction of free lexical resources is currently hampered by the difficulty of obtaining many small unpaid contributions. IToldU is a light web service which, in its first year of use for teaching technical English in French engineering schools, has led to the contribution of more than 17,000 English terms, in about 20 technical domains, with their French translations (95% correct) and almost as many examples of use (about 85% correct). The quality level has been raised in the second year. In 2 years, 22,000 entries have been created.

IToldU should now be extended to other language pairs, and to language triples. It is also a testbed for a user-friendly method to localize the interface to any language.

It remains to be seen whether IToldU can be synchronized with Papillon, a much more ambitious multilingual lexical database, and to what other contexts of use it could be extended to obtain other types of information, such as regimes, semantic formulas, lexico-semantic functions, or free collocations.

[1] Melchuk's term for the syntactic-semantic valencies, *aka* subcategorization frames.

ACKNOWLEDGMENTS

We would like to thank our reviewers for many useful comments, and Mark Seligman for a very detailed revision of the paper and improvement of its language. All remaining errors are of course ours!

REFERENCES

[1] V. Bellynck, "Bases lexicales multilingues et objets pédagogiques interactifs : Sensillon pour Papillon," in *Proceedings of Papillon 2002 Seminar*, NII, Tokyo, July 2002, 13 p.

[2] V. Bellynck, C. Boitet and J. Kenwright, "Resource pooling for technical English learning via lexical access," in *Proc. Papillon-04 seminar*, UJF, Grenoble, 30 Aug.-2 Sept. 2004, 5 p.

[3] V. Bellynck, C. Boitet, and J. Kenwright, "ITOLDU, a Web Service to Pool Technical Lexical Terms in a Learning Environment and Contribute to Multilingual Lexical Databases," in *Computational Linguistics and Intelligent Text Processing (Proc. CICLING-2005)*, A. Gelbukh (Ed.), Springer, LNCS 3406, pp. 319-327.

[4] M. Mangeot-Lerebours, "Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue," PhD in Computer Science, Université Joseph Fourier, Grenoble I, 280 p., Grenoble, France, 2001.

[5] M. Mangeot-Lerebours, G. Sérasset, and M. Lafourcade, "Construction collab-orative d'une base lexicale multilingue, le projet Papillon," *TAL*, 44/2, pp. 151-176.

[6] T. Murata, M. Kitamura, T. Fukui, and T. Sukehiro, "Implementation of Collaborative Translation Environment `Yakushite Net'," in *Proceedings of MT Summit VIII*, New Orleans, Sept. 2003.

[7] N. Tokuda and L. Chen, "An Online Tutoring System for Language Translation," *IEEE Transactions on Multimedia*, vol. 8, no. 3, pp. 46-55, July-September 2001.

Tecnología RFID
Aplicada al Control de Accesos

Juan Carlos Herrera Lozada, Patricia Pérez Romero y Magdalena Marciano Melchor

***Resumen*—En el presente trabajo se expone una introducción a la tecnología RFID (Identificación por Radio Frecuencia) que prometedoramente comienza a notarse como una alternativa viable para la captura de datos y el control de recursos varios en todos los sectores. En este mismo documento se incluye un análisis de las perspectivas propias y se culmina mostrando una aplicación práctica relacionada con el control de acceso.**

***Palabras clave*—RFID, Identificación por Radio Frecuencia, captura de datos, control de acceso.**

RFID Technology Applied to Access Control

***Abstract*—In this paper we present the perspectives of the technology RFID (Radio Frequency Identification), which is a notorious alternative for data capture and control of resources in many industrial sectors. After the discussion of its perspectives, we present a practical application of this technology related to access control.**

***Index Terms*—RFID, Radio Frequency Identification, data capture.**

I. Introducción

La tecnología RFID (Identificación por Radio Frecuencia, en inglés *Radio Frecuency IDentification*), nace como una alternativa de identificación automática de productos u objetos, similar a la lectura de códigos de barras que parece ser ya obsoleta e ineficiente. Comparando ambos casos, RFID no sólo tiene la ventaja de facilitar la creación de sistemas que almacenen mucho más información, sino que también permite identificar un producto u objeto como único, aunque sea de una misma clase, en contraparte, la lectura del código de barras considera un solo código de identificación por cada clase.

El sistema completo de RFID representa un método para almacenar y recuperar datos remotos a través de proximidad, éste se compone de tres partes o módulos básicos: Una tarjeta o etiqueta (*tag*), un dispositivo lector y un sistema de cómputo que contiene una base de datos [1, 2]; como puede observarse en la Fig. 1.

Manuscrito recibido el 17 de mayo del 2009. Manuscrito aceptado para su publicación el 20 de agosto del 2009.

J: C Herrera Lozada trabaja en Centro de Innovación y Desarrollo Tecnológico en Cómputo del Instituto Politécnico Nacional, México, D. F. (e-mail: jlozada@ipn.mx).

P. Pérez Romero trabaja en Centro de Innovación y Desarrollo Tecnológico en Cómputo del Instituto Politécnico Nacional, México, D. F. (e-mail: promerop@ipn.mx).

M. Marciano Melchor, Centro de Innovación y Desarrollo Tecnológico en Cómputo del Instituto Politécnico Nacional, México, D. F. (e-mail: mmarciano@ipn.mx).

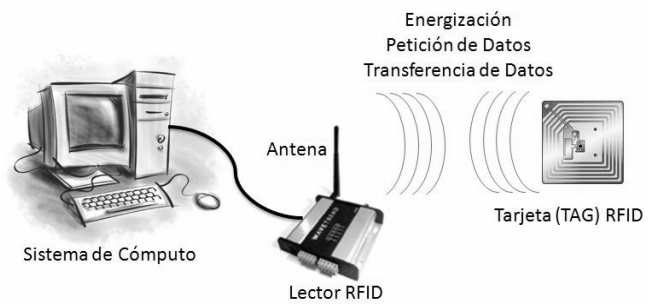


Fig. 1. Sistema básico de RFID.

El lector RFID genera un pequeño campo de radiofrecuencia que estimula e induce una antena en miniatura contenida en el encapsulado de la tarjeta, generándose en ésta una corriente eléctrica que permite que un microcircuito sea capaz de transmitir sus datos al lector. Así, cuando el lector hace una petición de datos, la tarjeta responde a dicha solicitud.

Los datos extraídos por el lector RFID pueden ser almacenados en una base de datos para realizar alguna consulta; en realidad, el sistema de cómputo de adecuará a las necesidades específicas de la aplicación.

La tarjeta se comporta como un Transponder (transmite y responde); el encapsulado de este dispositivo puede ser tan delgado como una hoja de papel y de un tamaño minúsculo. En este contexto, se dispone de tarjetas pasivas (sin alimentación interna, menor tamaño, menor coste) o tarjetas activas (alimentación interna, mayor almacenamiento). En las de tipo pasivo, la alimentación se obtiene de la misma frecuencia de trabajo y el sistema funciona mediante la técnica de modulación digital por frecuencia (FSK), con la que se facilita la adquisición pero está limitada en la distancia entre el lector y la tarjeta (de 2 a 10 centímetros) y en el número de lecturas que se pueden realizar. En las tarjetas activas de RFID, se utiliza comúnmente la alimentación por batería, propiciando alcances mayores en la proximidad (de 50 centímetros hasta 25 metros) [3, 4,5].

Los datos dentro de cada tarjeta se guardan en una memoria. Cada objeto a identificar tiene un código único y puede extraerse a distancia y sin tocarlo mediante el lector. Esta información puede ir desde un Bit hasta KBytes, dependiendo principalmente del sistema de almacenamiento que posea el transponder

El lector RFID consiste en una antena, un transceptor y un decodificador; éste envía señales periódicas para averiguar información de cualquier tarjeta/etiqueta en la vecindad.

El subsistema de procesamiento de datos (sistema de cómputo) provee los medios para procesar y almacenar los datos.

El funcionamiento de los dispositivos de RFID se realiza entre los 50 KHz y 2.5 GHz. Las unidades que funcionan a bajas frecuencias (50 KHz-14 MHz) son de bajo coste, corto alcance, y resistentes al "ruido" entre otras características. No se requiere de licencia para operar en este rango de frecuencia. Las unidades que operan a frecuencias más altas (14 MHz-2.5 GHz), son sistemas de mayor coste y tecnología más compleja.

De manera formal, para caracterizar un sistema RFID sería necesario profundizar en los temas de codificación y modulación de datos, control de errores, y colisiones ocasionadas por varias etiquetas cercanas que son estimuladas a la vez por un mismo lector [6, 7, 8].

II. Actualidad y Perspectivas de la Identificación por Radio Frecuencia

Dada la naturaleza de esta tecnología, la captura y recuperación confiable y eficaz de los datos presupone una mejor organización de procesos logísticos en almacenes y centros de distribución, aunado a las aplicaciones que conlleven a la identificación de códigos para validar alguna acción. En la actualidad, los sistemas de información implementados con tecnología RFID se utilizan ampliamente para catalogar y controlar recursos; por ejemplo, la clasificación de productos de un supermercado, la autentificación de documentos, la identificación de animales en granjas, acceso y control de vehículos, seguridad para medicamentos controlados y en el sector del consumo y del transporte, como sucede con las tarjetas recargables del Metro y del Metrobús de la Ciudad de México.

III. Análisis del Sistema de Control de Acceso Propuesto

Si consideramos que es posible implantar un sistema RFID para controlar el acceso a un recinto, se predispone el uso de una tarjeta que contenga el código correcto. Se parte de la idea de una empresa con n número de empleados, donde cada uno de estos tiene una tarjeta RFID con un código de identificación único. Solamente unos cuantos códigos autorizados tendrán acceso a cierta área restringida.

Para este ejercicio utilizamos un kit de desarrollo comercial que incluye un dispositivo lector RFID cuyo módulo principal es un circuito integrado (microcontrolador firmware), y unos cuantas etiquetas RFID en forma de tarjeta bancaria con códigos diferentes entre sí. El lector adquiere el código de la respectiva tarjeta y lo envía como un dato binario en forma serial.

En la particularidad del diseño presentado, para recibir y decodificar el dato que entrega el lector del kit de desarrollo, se utiliza un microcontrolador como subsistema de procesamiento de datos que recibe el dato serialmente y otorga validez al código, permitiendo el acceso al área restringida. La interfaz para visualizar la operación es una pantalla de cristal líquido (LCD) que indica si es un acceso positivo o no, aunque el microcontrolador se programó con la posibilidad de poder enviar datos a cualquier sistema de cómputo fijo (PC, laptop) o móvil (PDA, SmartPhone), conectándose a través de un puerto serie estándar.

IV. Características del Kit comercial

El sistema TIRIS *Micro-reader Module* (serie 2000) de Texas Instruments soporta datos de comunicación serial de la PC al micro lector [11]. Su interfaz de comunicación serial soporta comunicaciones TTL que permiten una comunicación estándar (RS232 y RS485). El módulo puede observarse en la Fig. 2.

Fig. 2. Módulo comercial TIRIS Micro-reader.

El micro lector puede trabajar remotamente enviando comandos a su interfaz de comunicación serial, que pueden ser manejados con o sin sincronización. La sincronización puede ser alambrada o inalámbrica, permitiendo una transferencia confiable en un ambiente que tenga uno o más dispositivos lectores. Dos salidas muestran el estado del micro lector e informan al usuario acerca del éxito del envío de los comandos. La antena del micro lector opera a 47 µH con una Q (factor de calidad) de entre 10 y 20 que genera una frecuencia de excitación de 134.2 Khz.

V. Desarrollo de la Aplicación

Una vez que una tarjeta es leída por el módulo comercial, el dato se envía hacia un microcontrolador PIC16F628 (con prestaciones superiores a otros y disponible en el mercado nacional a un bajo costo), que evaluará dicha información para desplegar un mensaje en respuesta por medio de una pantalla de cristal líquido; otro pin del mismo microcontrolador envía una señal que activa la bobina de una cerradura en caso de ser válido.

Es importante mencionar que el Micro Reader tiene comunicación hacia la PC por medio de su interfaz serial con un conector DB9 estándar; sin embargo, en la particularidad de este trabajo se utilizó la comunicación con el microcontrolador PIC16F628.

El lector RFID envía una frecuencia de 134.2 KHz por medio de la antena portadora durante un lapso de 50 ms (induciendo el circuito integrado de la tarjeta RFID para que ésta comience el envío del dato), en este período de tiempo la tarjeta procesa la información que transmite hacia el lector.

El lector tiene una apertura de tiempo de 20 ms, para recibir los datos. La antena para el lector se fabricó de acuerdo a las especificaciones del fabricante del módulo. En la Fig. 3 se muestra la antena y las tarjetas RFID utilizadas.

Fig. 3. Antena del módulo lector y tarjetas RFID (transponder).

El módulo lector no puede recibir durante el tiempo de carga o inducción de la tarjeta; con una señal la tarjeta indica que ha finalizado la carga y comienza a enviar datos usando el cambio de frecuencia FSK (Frecuency Shift Keying) como lo infiere la Fig. 4.

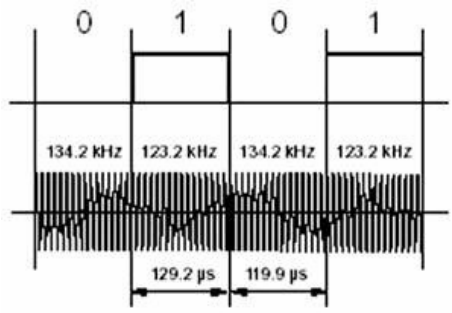


Fig. 4. Codificación FSK empleada.

La información que transmite cada tarjeta está modulada en frecuencia. Aunque la longitud de la trama es constante en bits, es variable en tiempo. La trama más larga durará unos 18ms.

Las secuencias de carga y lectura se controlan en los módulos de identificación mediante la señal de control TX (trasmisión) clásica en la comunicación serial. La duración de la fase de carga dependerá del tipo de tarjeta RFID, la distancia de paso, forma y tamaño de la antena del lector.

Para la lectura se utiliza un formato definido por el fabricante de la siguiente manera:

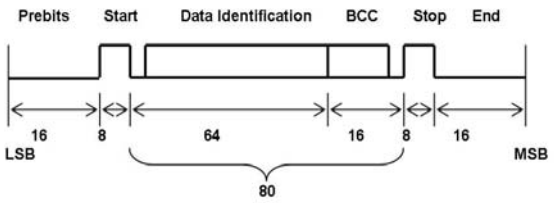


Fig. 5. Formato de lectura.

- *Prebits*, estos indican que es un RFID de sólo lectura (0000Hex).
- *Start byte* indica el comienzo del mensaje (SOS, start of header, 01 Hex).
- *Status byte* provee retroalimentación de la lectura precedente u operación de programa.
- *Length* Indica la longitud en bytes del siguiente campo de datos.
- *Data Identification* de acuerdo a ciertos bits relevantes este campo de datos se enviará al RFID o no, estos datos se programan de fábrica.
- *End Bits* son validados por el módulo de control RF.

El campo de datos está integrado por 80 bits que se encuentran entre el *Start* y el *Stop Byte*, que se programan de fábrica, es decir un código único de 64 bits, que excluyen los 16 bits del BCC (bits de protección de datos).

Después del *Stop Byte* se transmiten 16 bits; los primeros 15 bits, comenzando por el menos significativo, se chequean en el módulo de control. Durante el 16 bit el transponder termina el formato de datos.

VI. Integración del Hardware

En este apartado se muestra la integración de los elementos, el montaje del modulo Micro-reader TIRIS con el microcontrolador PIC16F628 y éste a su vez con la pantalla de cristal líquido (LCD). Se recomienda revisar la hoja de especificaciones del módulo TIRIS.

La programación del microcontrolador PIC16F628 se realizó con ayuda del lenguaje de alto nivel Pic Basic Pro, que facilita en gran medida el diseño, dado que se tienen instrucciones interconstruidas especiales para la comunicación serie.

A continuación se muestra un fragmento del código escrito en el lenguaje anteriormente referido para la programación del microcontrolador, que se encarga de hacer la lectura de los datos seriales enviados por el módulo lector Micro-reader, enviando resultados de la validación hacia una pantalla convencional de LCD de dos líneas con 16 caracteres en cada una de ellas.

```
TRISB = 2      'Pb.1, como entrada 'serial., Los
               demás como 'salida.

TRISA = 0  'PA.X como salidas LCD.

'Inicialización de LCD.
Pause 500      'Inicialización    de    LCD,    '0.5
               segundos.

lcdout  $fe, 1 ' Limpia pantalla LCD.
pause 250

'Inicio de programa principal
inicio:
lcdouT  $fe, 1  ' Limpia pantalla LCD.
pause 250
Lcdout "  LECTOR RFID "
Lcdout $fe, $C0       'Salta a segunda 'línea
```

```
PAUSE 250
INI:
Lcdout $fe, $C0
Lcdout " NO HAY TARJETA"

Tipo:
Serin PORTB.1,T9600, EPC
IF EPC = $09 then GOTO V
        goto Tipo

V:
Serin PORTB.1,T9600
If EPC = $AA then GOTO B
Lcdout $FE,1
Lcdout " NO AUTORIZADO"
pause 3000
goto inicio

B:
Lcdout $fe, 1
Lcdout " AUTORIZADO "
Lcdout $fe, $C0
For i=0 to 15
lookup i,["Bienvenido....."],aux
Lcdout aux
pause 200
next i
Lcdout $fe, $C0
portb.2 = 1
Lcdout " PUERTA ABIERTA "
Pause 3000
portb.2 = 0
goto inicio
end
```

Cuando el módulo Micro-reader detecta un código RFID (proveniente de una tarjeta) a través de su antena, lee el código y lo envía al microcontrolador PIC que continuamente está leyendo el pin de recepción de datos seriales (portb.1); lo que hace realmente éste último es evaluar los dos primeros datos (09hex) que luego por medio de una sentencia de decisión procede a evaluar los otros dos datos (AAhex) del código único del RFID, que de ser aceptado envía a la pantalla LCD un mensaje de autorización y bienvenida. También por el portb.2 del PIC envía una señal que puede activa un cerrojo electrónico que concede el acceso.

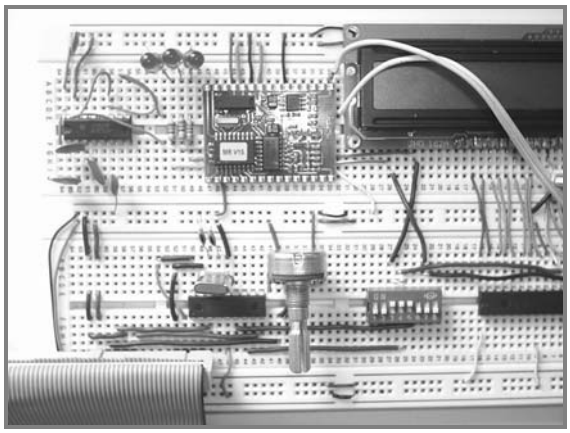


Fig. 6. Montaje completo de la aplicación.

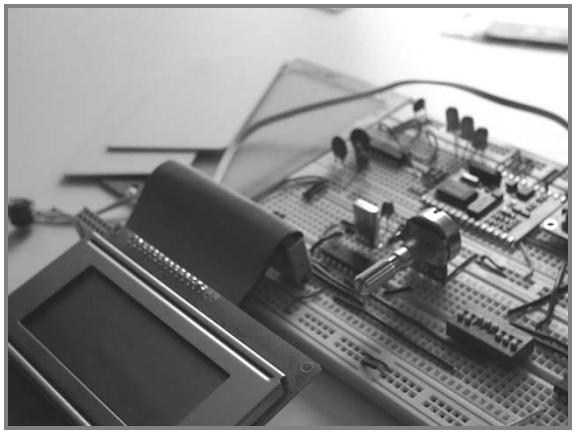


Fig. 7. El mismo montaje con una LCD más especializada.

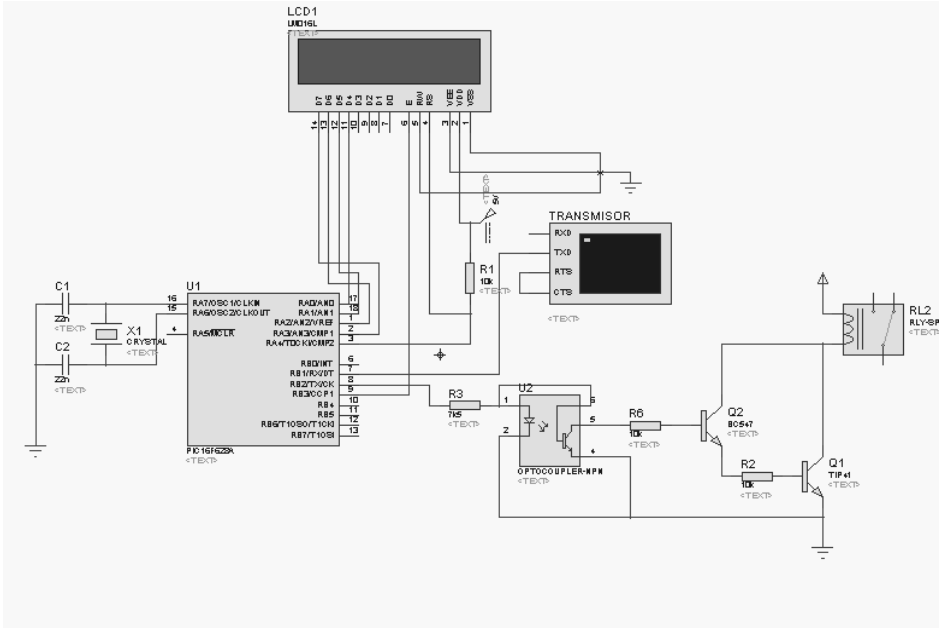


Fig. 8. Diagrama del control de acceso por RFID.

VII. Prototipo de un Sistema de Cómputo Móvil

Como se comentó con anterioridad, el microcontrolador se programó con la posibilidad de sustituir la pantalla de LCD por un dispositivo de cómputo fijo o móvil, a continuación se presenta una adecuación para monitorear los datos en la pantalla de un PDA, lo que infiere una mayor robustez en el prototipo. La idea general versa de tres partes: el kit de desarrollo RFID, el microcontrolador y el propio PDA. El esquema de conexiones se muestra en la Fig. 9.

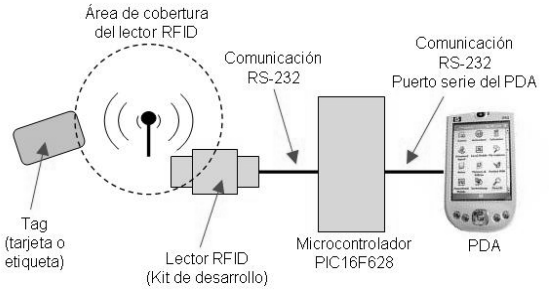


Fig. 9. Prototipo para monitorear datos en un PDA.

Se observa en la Fig. 8, que el microcontrolador sugerido se debe programar con una funcionalidad serial con el protocolo RS-232 alambrado, tanto para recibir los datos del kit lector RFID, como para enviar los resultados al PDA utilizando el puerto serie de éste último.

El PDA debe monitorear y supervisar la funcionalidad del microcontrolador a través de un programa residente. El microcontrolador se encarga de recibir secuencialmente el dato proveniente del PDA en formato estándar binario (también podría enviarse en formato ASCII) con una velocidad predeterminada de 9600 baudios, sin paridad y con un bit de paro.

A continuación se lista un fragmento del código que se programó en el PIC16F628 para establecer comunicación entre el PDA y el lector RFID del kit de desarrollo.

```
'Inicio de programa principal
inicio:
serout (establece comunicación con el PDA)
pause 250
serout "  LECTOR RFID "--imprime el PDA

Tipo:    lee dato del lector
Serin PORTB.1,T9600, EPC
IF EPC = $09 then GOTO V
        goto Tipo

V:
Serin PORTB.1,T9600
If EPC = $AA then GOTO B
Serout  " NO AUTORIZADO" –imprime PDA
pause 3000
goto inicio

B:
Serout  " AUTORIZADO " –imprime PDA
Pause 1000
Serout "Bienvenido" –imprime PDA
goto inicio
```

En los proyectos realizados en el CIDETEC se han utilizado frecuentemente PDAs de la familia iPAQ Pocket PC, fabricadas por HP, con sistema operativo Windows Mobile, por lo que el ambiente de desarrollo óptimo para programar estos dispositivos es Visual Studio. NET.

De manera alambrada, para un puerto serial de comunicación, Visual Basic contiene el control MS COMM con la opción a disparo, es decir, al depositar un byte en el buffer del puerto automáticamente se dispara el evento correspondiente. También es posible realizar un poleo al buffer del puerto, cada determinado tiempo, buscando el byte recibido. Este control incorpora todas las funciones para configurar el puerto, para mayor información referirse a [12]. En las Fig. 10 y 11, se aprecia la aplicación programada en el PDA.

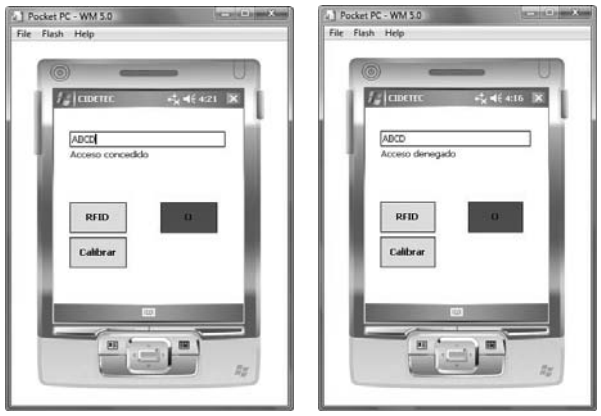


Fig. 10. Pantallas en tiempo de ejecución (simulación), para el acceso concedido y el acceso denegado, respectivamente.

Fig. 11. Aplicación ejecutándose en PDA.

VIII. Pruebas y Resultados

Se consideró de inicio el prototipo con la LCD, posteriormente se realizaron las mismas pruebas en el prototipo con el PDA; una vez realizada la integración y comprobando los datos de las tarjetas disponibles, el microcontrolador decide cuál de las tarjetas contiene la información correcta y establece comunicación con el despliegue respectivo (LCD o PDA).

La proximidad de la tarjeta hacia el módulo lector soportó distancias hasta de 6 centímetros. De acuerdo al fabricante, una antena bien construida podría permitir distancias de hasta 25 centímetros. Cabe mencionar que no se provocaron

colisiones acercando al lector dos tarjetas al mismo tiempo, lo anterior debido a que no está dentro de los alcances de esta propuesta darle tratamiento a este problema.

La bobina de la cerradura electrónica se activó de manera correcta sólo en el caso válido, por lo que el control de acceso funcionó correctamente.

IX. Conclusiones

Se presentó un panorama general de la tecnología de Identificación por Radio Frecuencia (RFID), el objetivo principal consistió en proponer el desarrollo de aplicaciones que utilicen la identificación de códigos y el procesamiento de datos bajo este esquema.

Este trabajo incluyó una aproximación real que puede hacerse extensiva a otras aplicaciones sin cambios drásticos. El kit de desarrollo utilizado puede ser sustituido por otro de características similares. En consecuencia al diseño mostrado, es posible resumir que el microcontrolador que recibe el dato serial proveniente del lector, es un *core* o núcleo reutilizable.

Si bien, el control de acceso diseñado no es un sistema completamente robusto, sirve para determinar claramente la intención de su aplicación.

El modo que se eligió para trabajar con el módulo lector Micro-reader TIRIS fue el de sólo lectura de tarjeta (*RO-read only*), pues cuenta con otros dos modos de operación: lectura-escritura (R/W) que no sólo lee la tarjeta RFID, si no que también puede modificar sus datos, y el modo multipágina (*MPT multi-page*) que tiene mucha más capacidad en cuanto a almacenamiento de datos. Estas características son las que permiten que un RFID sea un gran candidato para sustituir a los muy limitados códigos de barras.

Referencias

[1] B. Glover, “RFID Essentials Theory in Practice,” O’Reilly press, 2005.

[2] M. Bhuptani, “RFID Field Guide: Deploying Radio Frequency Identification Systems,” Prentice Hall, 2005.

[3] S. Garfinkel, “RFID: Applications, Security, and Privacy”, Addison-Wesley Professional, 2005.

[4] H. Vogt, “Efficient object identification with passive RFID tags,” in *Proc. of Int. Conf. on Pervasive Computing*, LNCS, Zurich, 2002, pp. 98–113.

[5] V. Stanford, “Pervasive computing goes to the last hundred feet with RFID system,” *IEEE Pervasive Computing*, 2 (2), 2003, 9–14.

[6] P. Hernandez, J. D. Sandoval, F. Puente, and F. Perez, “Mathematical model for a multiread anticollision protocol,” in *IEEE Pacific Rim Conf. on Communication, Computer and Signal Processing*, Vol. 2, Victoria, Canada, 2001, 26–28.

[7] Li Lu, Jinsong Han, Lei Hu, Yunhao Liu, and Lionel M Ni, "Dynamic Key-Updating: Privacy-Preserving Authentication for RFID Systems," in *IEEE PerCom 2007*, White Plains, NY, USA, March 2007.

[8] D. Engels and S. Sarma, “The reader collision problem,” in *Proc. IEEE Int. Conf. on Systems, Man and Cybernetics*, Vol. 3, Hammamet, Tunis, 2002, 6 pp.

[9] K. Finkenzeller, “RFID handbook: Radio-frequency identification fundamentals and applications,” 2nd ed. (New York: John Wiley & Sons, 2003).

[10] S. E. Sarma, S. A. Weis, and D.W. Engels, “Low cost RFID and the electronic product code,” in *Workshop on Cryptographic Hardware and Embedded Systems*, LNCS, Berlin, Germany: Springer-Verlag, 2002.

[11] www.ti.com/rfid/docs/manuals/refmanuals/micro_8.pdf.

[12] J. C. Herrera, I. Rivera, M. Olguín, “Computadoras de Bolsillo como una Alternativa al Control de Servomotores en Robótica,” *Polibits*, 38, 2008, pp. 75-79.

An Extended Payment Model for M-Commerce with Fair Non-Repudiation Protocols

Tran Khanh DANG and Thi Thanh Huyen PHAN

***Abstract*—Non-repudiation in e-commerce has recently gained a lot of interest but its successor brother, non-repudiation in m-commerce, is still at the start. In this paper, we propose an extension of existing mobile payment models to introduce an extended mobile payment service (EMPS) model, which is based on assumptions about the cooperation between mobile network operators and financial institutions to deal with different payment amounts ranging from micro to macro payment. The novel model focuses on enhancement of non-repudiation problem. Fair non-repudiation protocols are developed for not only payment phase but also other phases in a typical m-commerce transaction, including price negotiation and content delivery. Joint signatures method is used in protocols to overcome the limitations in mobile handheld device capability and to reduce the trust dependence totally on the payment service. As with the proposed non-repudiation protocols, EMPS plays the role of a semi-trusted third party and is an indispensable factor for creating the fairness property. Non-repudiation analyses of these protocols are also conducted besides some guidelines for ensuring non-repudiation in m-commerce.**

***Index Terms*—Communication system security, M-commerce security, non-repudiation, semi-trusted 3rd party, payment model.**

I. Introduction

In recent years, *m-commerce* with many advantages such as the ubiquity, reachability, localization has emerged as a new potential application and research area. However, its inherently secure weaknesses, resulted from the limited capacity and the mobility of mobile handheld devices, insecure wireless channel, etc are the main obstacles on the path of success. Basically, security in m-commerce also deals with the fundamental issues as authentication and authorization, confidentiality, integrity, availability, and non-repudiation. Among these issues, non-repudiation, one of the services used to cope with internal attack risks, almost has not been studied thoroughly.

Manuscript received June 29, 2009. Manuscript accepted for publication November 17, 2009.

This work was supported in part by Advances in Security & Information Systems (ASIS) Lab, Faculty of Computer Science & Engineering, HCMUT, Vietnam.

T. K. Dang is with the Faculty of Computer Science & Engineering, HCMC University of Technology, VNUHCM, Ho Chi Minh City, Vietnam (phone:+84-8-38647256, ext. 5841, e-mail: khanh@cse.hcmut.edu.vn).

T. T. H. Phan is with the Faculty of Computer Science & Engineering, HCMC University of Technology, VNUHCM, Ho Chi Minh City, Vietnam (phone:+84-8-38647256, ext. 5842, e-mail: huyenttp@cse.hcmut.edu.vn).

Repudiation is the false denial of having been involved in a communication. The goal of the non-repudiation service is to generate, collect, maintain, make available and verify evidence concerning a claimed event or action in order to resolve disputes about the occurrence or non-occurrence of the event or action [12]. Currently, most non-repudiation protocols use digital signature in generating non-repudiation evidences. Among the properties of a non-repudiation protocol, fairness may be the most desirable. This feature helps the protocol execute fairly, i.e. at the end of the protocol, either both entities get the expected evidences, or none of them get any valuable information. Using a trusted third party (TTP) is a common approach to resolve this problem.

The power of non-repudiation services creates its importance to the commercial transactions in e-/m-commerce environments where the parties participating in may not trust each other. Non-repudiation in e-commerce has generated a lot of interests recently and built a relatively sound foundation while this issue in m-commerce is still at a start. Although m-commerce can be considered as mobile e-commerce, we can not apply the same non-repudiation protocols in e-commerce to the new environment because of the inherently insecure nature of wireless network and limited capability of mobile devices. Therefore, we need lightweight but sufficiently secure non-repudiation protocols to protect transactions conducted in wireless environment. Non-repudiation protocols in m-commerce should be based on existing non-repudiation protocols in e-commerce and adjusted to suit the resource constraints of mobile devices as well as specific requirements of different transaction types. Nearly all currently existing research mentioning the non-repudiation in m-commerce just pays attention to this problem in mobile payment, one of the most important commercial transactions in m-commerce. Mobile payment or billing is defined as the process of two parties exchanging financial value using a mobile device in return for goods or services [13]. A general mobile payment system, along with a typical transaction, is described in figure 1. It uses a third party which could act at the same time as a payment service provider and a TTP to support the financial transaction between mobile customer and service provider. As with m-commerce, none of the proposed solutions gives a complete analysis of non-repudiation properties such as non-repudiation evidences, the fairness property, the timeliness properties or a formal verification and so on. Moreover, some

forget the limited capability of mobile handheld devices while other solutions are suitable for only some specific cases.

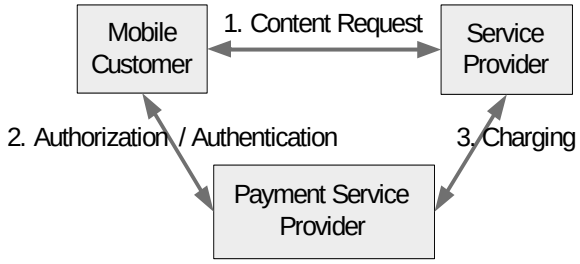


Fig. 1. A general mobile payment system.

In this paper, we propose a new mobile payment system founded on the extension of existing ones to support not only the non-repudiation protocol in the payment stage but also the other phases of a general commercial transaction such as price negotiation, content delivery, placing particular emphasis on the payment phase. In the proposed system, in addition to the traditional role, payment service provider also takes the role of a semi-trusted third party in non-repudiation protocols and supports mobile devices in generating non-repudiation evidences to help them overcome their limitations in computational power. The model and protocols proposed also address a variety of payment methods like credit card/account based payment methods, phone bill charging method and payment amounts like macro/micro payment.

The rest of this paper is organized as follows. Section 2 briefly discusses the related work. Section 3 presents common requirements and properties of non-repudiation protocols in m-commerce. In section 4, we introduce an extension of existing mobile payment systems, the overall architecture and innovations of our approach, and present three non-repudiation protocols for mobile transactions. Next, we carry out theoretical analyses of the proposed protocols in section 5. Finally, section 6 gives concluding remarks and presents future work.

II. Related Work

There are several mobile payment models, ranging from the concept to universal model [2], and most of them do not refer to non-repudiation problems. Moreover, the other systems mentioning non-repudiation in their solutions either have too simple and inflexible models or give incomplete analysis and unsuitable non-repudiation protocols for m-commerce environment. Some typical previous work is discussed as follows.

- The limited models such as PayBox, Vodafone m-PayBill, iPIN [10] are restricted in payment methods, customer, secure mechanism and none of them provides non-repudiation protocols.
- SEMOPS [9] is a typical example of universal models. The model looks prettily perfect because it is capable of supporting all transactions values, operating in any channels and supporting any transaction type with a domestic and/or international geographic coverage. However, SEMOPS does not give any formal protocols for its transactions and non-repudiation is not handled too. Another limitation of SEMOPS is that the customer and service provider have to trust the payment processor absolutely. Furthermore, using data center increases the number of steps in a transaction and reasonable solutions to traditional problems around data center such as bottleneck, attack risks are not presented.
- The payment model presented in [2] is derived from SEMOPS model with some enhancements for tackling the signature validating and privacy issues. A protocol which is a formal representation of the payment process and some initial non-repudiation analyses are discussed in [2]. This is a very first effort for non-repudiation in m-commerce, but the proposed protocol does not take into account the limited capacity of mobile devices when using traditional signatures to generate non-repudiation evidences and the non-repudiation analysis is just the case of non-repudiation of origin. Moreover, the given protocol skips the differences in nature of different payment methods and payment value.
- Other solutions to non-repudiation in mobile payment concerning evidence generation cost are given in [1, 3]. Both of them use the joint signature instead of traditional digital signature to reduce cost but one is for home network and the other is for foreign network. Although they are better than the aforementioned ones due to low cost, deeper non-repudiation analysis, they are only suitable for small payment which charges mobile customers through their phone bills.

III. Non-Repudiation Considerations in M-Commerce

An m-commerce transaction taking place between mobile customer (MC) and service provider (SP) usually involves three phases: price negotiation, payment and content delivery. The non-repudiation requirements for this transaction include:

- *Non-repudiation in price negotiation phase:* MC and SP can not falsely deny having involved in the communication and agreed on the given price.
- *Non-repudiation in payment phase:* MC can not falsely deny having agreed to pay her bill and SP can not falsely deny having received the payment for the invoice of MC.
- *Non-repudiation in content delivery phase*: MC can not falsely deny having received goods and SP can not falsely deny having not delivered the goods.

By examining the existing non-repudiation protocols in e-commerce and specific properties of m-commerce like the limited computational capability, inherently insecure wireless network, we identify some requirements for building non-repudiation protocols in m-commerce:

- They should be built on the non-repudiation foundation in e-commerce.
- Number of messages originated from mobile customer should be minimized.

- Cost for non-repudiation evidence generation and verification should be low but the used methods must reach an acceptable level of data security.
- A third party supporting the data delivery and evidence generation should be employed and the candidate for this role will vary according to transaction type and be chosen from the main players in mobile commerce environment such as mobile network operator (MNO), bank. Moreover, we should reduce the trust dependence on these third parties.
- Specific properties of each transaction type should be examined.

A. Non-Repudiation Evidence

Most of the existing solutions for non-repudiation in mobile commerce are reducing cost in non-repudiation evidence generation resulted from the limited computational capability of mobile devices. They can be divided into two groups: one based on the symmetric key technique and the other founded on the digital signature technique.

- *Symmetric key technique:* The idea is to use the symmetric cryptographic technique to create evidence at a low cost. However, if we use just a secret key k shared between two parties, generated evidence can not be irrefutable. The solution here is different from the rule of secure envelops mechanism in non-repudiation in e-commerce. It combines using 2 secret key k_1, shared between MC and SP, k_2, shared between MC and TTP, with other techniques such as hash, keyed hash, MAC. Therefore, the evidence containing both k_1 and k_2 must be generated by MC. A number of proposals like [8] can be counted in this group.
- *Digital signature technique:* Although digital signature can ensure the non-repudiation of origin of evidence, the cost of generating it is too high for limited computational devices to execute. Some schemes have been proposed to address this problem by designing more efficient mathematical algorithm. Other proposals use a third party to sign message on the original signer's behalf such as joint signature, proxy signature or server-supported signature [1].

B. Trusted Third Party

The particular properties of m-commerce environment influence the choice of candidate for TTP role in a non-repudiation protocol. Besides the traditional TTP which is indispensable for a non-repudiation protocol such as time-stamping authority (TSA), certification authority (CA), TTP assisting in fair exchange of the message and/or non-repudiation evidence can be one or combination of the following players:

- *Mobile Network Operator (MNO):* MNO owns the channels and almost all communications in mobile environment must pass through it. Besides its large customer bases, MNO has a lot of experience in the fields of billing and roaming.
- *Financial Institution/Bank (FI/B):* Its strengths lie in the trust of customers and long-standing customer relationships. Stemming from its expertise to handle transaction and risk, the necessary licenses, large customer and merchant bases, etc, FI/B is a valuable candidate for the role of a TTP, especially in the case of payment services.
- *Independent agent (IA)*: Although IA does not have advantages like MNO or FI/B such as the trust of customers and large customer bases; it can be more flexible and faster to explore new technologies than MNO or FI/B. Moreover, IA can collaborate with different mobile network operators and financial institution to offer its services to a variety of customers.

IV. EMPS System Model with Non-Repudiation Protocols

This section presents our main contributions, solutions to non-repudiation in m-commerce, by building an extended mobile payment service (EMPS) to support non-repudiation protocols in not only the payment phase but also other phases in a general m-commerce transaction including price negotiation and content delivery.

A. EMPS System Model

EMPS system model is based on the models introduced in [2, 9] because of their extensibility and universality. Some improvements are suggested to meet our requirements.

- A data center is not used in our model because it increases the complexity of the non-repudiation protocol with many steps, third parties and the trust level to third parties. Moreover, the other problems in a data center such as bottleneck, attack risk, message integrity can arise. In our model, customers of different EMPSs can do business together if their EMPSs have made a deal.
- Our model is also a payment service, so the value of transaction greatly affects the proposed protocols. Payment amounts are usually categorized in micro and macro payment. Micro payment refers to small purchases, usually less than 10 Euro and macro payment is about large purchases over 10 Euro. EMPS assumes that MNO and FI/B will collaborate on payment phase. Micro payment has low requirements for security but cost efficiency, hence it is reasonable to ask MNO to charge customers through their mobile phone bills. Marco payment requires higher security level, thus it should be paid by customer's bank account or card. FI/B with a lot of experience in payment services and risk management will responsible for macro payment.
- An innovation of EMPS is that it not only features mobile payment service but also supports MC and SP in price negotiation and good delivery in order to obtain a fair non-repudiation transaction. This is the reason we name this model *Extended Mobile Payment Service*.
- To reduce computational load on mobile user without affecting the system security, we use the idea of joint signature [1] in generating non-repudiation evidences. MC will have 2 secret keys: $k_{mc,emps\text{-}mc}$ shared between MC and EMPS of MC, $k_{mc,sp}$, shared between MC and SP. This means MC is the originator of messages

containing both $k_{mc,emps-mc}$ and $k_{mc,sp}$. EMPS of MC, which has large computational capability and also involves in the transaction between MC and SP, will sign on these messages to create the irrefutable evidences of non-repudiation protocols.

– A strong point of this model is the ability to reduce the trust dependence of MC to EMPS. In the system, EMPS can be regarded as a semi-trusted third party. This implies that EMPS just helps MC sign evidences and transfer them to other parties but it cannot modify or forge these evidences because of the presence of the secret key $k_{mc,sp}$, which is known only by MC and SP, in these evidences.

The system model of EMPS is shown in figure 2. There are four main parties participating in this model: MC, SP, EMPS-MC, and EMPS-SP. To gain the generality, we assume that MC and SP register to different EMPSs and these EMPSs trust each other. EMPS-MC and EMPS-SP are the payment service providers of MC and SP respectively. In our model, MNO collaborates with FI/B to build EMPS. While FI/B deals with macro payment, MNO is responsible for micro payment and supports MC in generating joint signature besides the traditional role of wireless access provider. In addition to these main parties, TSA and CA which are the essential TTP in most non-repudiation protocols also appear in our model. TSA is used to add trusted time information to evidence; and in the non-repudiation protocols, the step in which evidence is time-stamped is usually omitted for simplicity. CA is another TTP that issues public key certificates to guarantee the authenticity of public verification keys used for non-repudiation purpose.

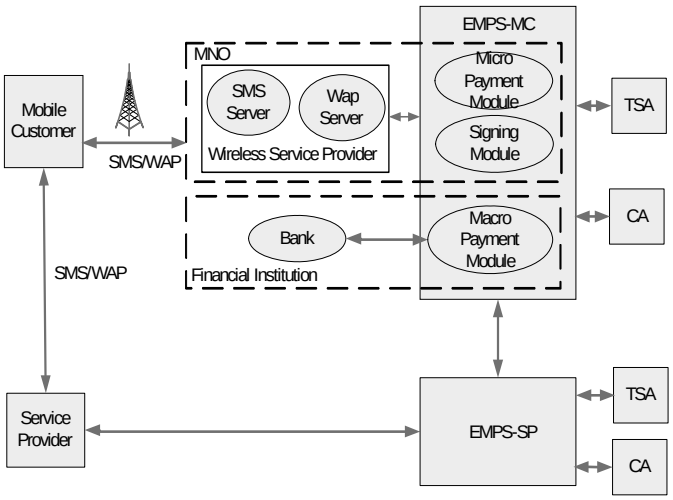


Fig. 2. EMPS System Model

Due to the space limitation, we just summarize main features of crucial modules in EMPS. Micro Payment Module handles micro payment and Macro Payment Module involves in macro payment. Signing Module helps MC generate joint signature on evidences and messages. Price Negotiation Module manages the related message in price negotiation phase and Content Delivery Module deals with managing the related messages in content delivery phase. Another module is User Management Module which manages customers of EMPS. Customers of EMPS can be categorized into MC, SP and other EMPS. Therefore, we need some communication modules serving the interaction between MC and EMPS, SP and EMPS as well as between EMPSs. In order to facilitate customers' easy access to the services, EMPS also supplies the front-end modules to MC and SP, especially MC front-end module which can assist MC to carry out an m-commerce transaction with necessary functions such as price negotiation, payment, personalization, and security.

B. Fair Non-Repudiation Protocols of M-Commerce Transactions in EMPS

TABLE I
NOTATIONS.

h(m)	Collision resistant hash function
m_1, m_2	Concatenation of data item m_1 and m_2
$k_{A,B}$	Session key shared between A and B
ID_A	Identity of entity A
$pk_A=(e_A, n_A)$	Public key of entity A
$sk_A=(d_A, n_A)$	Private key of entity A
$s_A=(h(m))^{d(A)} \bmod n_A$	Entity A signature over message m
Ek	A symmetric key encryption function under key k
Dk	A symmetric key decryption function under key k
c=Ek(m)	Cipher of message m under the key k
E_A	A public key encryption function under A's public key
D_A	A public key decryption function under A's private key
$cert_A$	Digital Certificate of entity A
L	A label uniquely identifies a protocol run
F	A flag indicating the purpose of a message
y=HOAC	Hash Origin Authentication Code
x=HMAC	Hash Message Authentication Code
ts_A	Time stamp of entity A
d_{A-P}	A deadline for response which is imposed by A in protocol P
$dl=[t_s, t_e]$	A time interval
OI	Order Information
P_A	Price suggested by entity A
PID	Identity of Product which MC intends to buy
N	Quantity of a Product which MC intends to buy
Adr_A	Delivery address of A
$Account_A$	Information about account of entity A
SC	Shipping cost

In this section, we present three fair non-repudiation protocols built for three phases of an m-commerce transaction: price negotiation, payment and content delivery. Assume that the communication channels among EMPS, between EMPS and both SP/MC are resilient. The communication between MC and SP may be unreliable.

Firstly, there is an initiation process occurring before these three phases for establishing a session key shared between MC and SP. As MC and EMPS-MC share a private key $k_{mc, emps-mc}$ which is issued to MC when she registers for services of EMPS-MC, we can apply the NAETEA protocol [4] in this case. At the end of the initiation process, MC and SP share a session key $k_{mc,sp}$ and SP also receives a hash value of the secret key shared between MC and EMPS-MC: $h(k_{mc,emps-mc})$. Secondly, joint signature method [1] is used in our protocols. To help readers grasp the general idea of the joint signature scheme we briefly explain the used notations in Table I.

Price Negotiation Phase

In this phase, MC negotiates with SP for certain goods. First, MC sends the order containing information about product identity (PID), amount (N), bidding price (P_{mc}): OI_{mc} = PID1, N1, $P1_{mc}$, PID2, N2, $P2_{mc}$.... along with HOAC, HMAC to EMPS-MC. y_{pn} is the HOAC and includes the secret $k_{mc,sp}$ which is not known to EMPS-MC, thus EMPS-MC can not forge a valid y_{pn} to SP. In addition, EMPS-MC can not get $k_{mc, sp}$ from y_{pn} since it is hashed. HOAC also embeds the hash secret $h(k_{mc,emps-mc})$ to protect the SP against false accusation by MC, or impersonation attacks by the SP or other entities against the MC. x_{pn} is the HMAC and can be used for source authentication because it contains $k_{mc, emps-mc}$ which is shared between MC and EMPS-MC only. Moreover, using the received HMAC, EMPS-MC can verify the integrity of $EK_{mc, sp}$ (OI) and y_{pn}. In short, the HOAC y_{pn} indicates to SP that the original of OI_{mc} is from MC and the HMAC x_{pn} indicates to EMPS-MC that the HOAC y_{pn} is from MC. dl is also introduced to check the freshness of the timestamp ts_{mc} and prevent EMPS-MC from deliberately replaying the signature generation so as to gain advantage. d_{mc-pn} is the deadline in which MC wants to receive the response of SP for its bidding prices. If the user is successfully authenticated, then in step 2, EMPS-MC will construct the joint signature from these messages and send to SP. In step 3, SP verifies the authenticity and integration of request. After successful request verification and validation, SP considers the bidding price of MC and replies with an OIResponse before the deadline d_{mc-pn}. SP also sets a deadline d_{sp-pn} for MC's feedback and generates a label l used for future communications. If the prices in OIResponse are the same as those in OI_{mc}, SP and MC reach an agreement. On the contrary, this phase will be repeated until one agrees with the prices given by the other or decides to give up. l and OIResponse will be used in the other phases of the transaction.

1. MC → EMPS-MC: f_{pn}, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, d_{mc-pn}, dl, ts_{mc}, $Ek_{mc, sp}$ (OI_{mc}), y_{pn}, x_{pn}
y_{pn} = $h(OI_{mc}$, dl, $k_{mc,sp}$, d_{mc-pn}, $h(k_{mc,emps-mc}))$ and x_{pn} = $h(Ek_{mc,sp}(OI_{mc})$, f_{pn}, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, ts_{mc}, $k_{mc,emps-mc}$, $y_{pn})$.

2. EMPS-MC → SP: f_{pn}, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, dl_{pn}, d_{mc-pn}, $ts_{emps-mc}$, $cert_{emps-mc}$, $Ek_{mc, sp}$ (OI_{mc}), y_{pn}, x_{pn}, $signpn_{emps-mc}$
The joint signature $signpn_{emps-mc}$ = $s_{emps-mc}(f_{pn}$, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, dl_{pn}, d_{mc-pn}, $ts_{emps-mc}$, $Ek_{mc,sp}(OI_{mc})$, y_{pn}, $x_{pn})$.

3. SP → MC: f_{pn}, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, $ID_{emps-sp}$, d_{sp-pn}, $Ek_{mc,sp}$(OI$_{mc}$, OIResponse, pk_{sp}, l), s_{sp}(OI$_{mc}$, OIResponse, pk_{sp}, l)
OIResponse = OI_{sp}, SC, d_{sp-pn} and l = $h(ID_{mc}$, ID_{sp}, $ID_{emps-mc}$, $ID_{emps-sp}$, h(OIResponse), $h(k_{mc, emps-mc}))$, OI_{sp} = PID1, N1, $P1_{sp}$, PID2, N2, $P2_{sp}$....

Payment Phase

If MC and SP reach an agreement at the end of the negotiation phase, MC will conduct the payment phase. Depending on the payment amount, MC will choose the micro payment protocol or macro payment protocol.

1. MC → EMPS-MC: f_{mip}, l, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, $ID_{emps-sp}$, dl_{mip}, ts_{mc}, d_{mc-mip}, $Ek_{mc, emps-mc}$ (OIResponse), $y_{mip, x_{mip}}$
y_{mip} = h(OIResponse, l, dl_{mip}, $k_{mc, sp}$, d_{mc-mip}, $h(k_{mc, emps-mc})$) is a HOAC showing SP that the original of the request for payment is from MC.
x_{mip} = h(OIResponse, f_{mip}, ts_{mc}, $k_{mc, emps-mc}$, y_{mip}) is a HMAC showing EMPS-MC that the HOAC y_{mip} is from MC.

2. EMPS-MC → SP: f_{mip}, l, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, dl_{mip}, $ts_{emps-mc}$, d_{mc-mip}, E_{sp}(OIResponse), $y_{mip, x_{mip}}$, $signmip_{emps-mc}$
The joint signature $signmip_{emps-mc}$ = $s_{emps-mc}(f_{mip}$, l, dl_{mip}, $ts_{emps-mc}$, d_{mc-mip}, OIResponse, $y_{mip, x_{mip}})$

3. SP → EMPS-SP: f_{mip}, l, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, $ID_{emps-sp}$, d_{mc-mip}, $E_{emps-sp}$ (Bill), s_{sp} (f_{mip}, l, Bill, d_{mc-mip})
Bill = OIResponse, Approval

4. EMPS-SP → EMPS-MC: f_{mip}, l, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, $ID_{emps-sp}$, $E_{emps-mc}$ (Bill, $Account_{emps-sp}$), $s_{emps-sp}$ (f_{mip}, l, Bill, $Account_{emps-sp}$)

5. EMPS-MC → MC: f_{mip}, l, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, $ID_{emps-sp}$, $Ek_{mc, emps-mc}$ (Bill), $s_{emps-mc}$ (f_{mip}, l, Bill)

Micro Payment Protocol: MC sends the request for payment to SP through EMPS-MC. EMPS-MC creates the joint signature and encrypts the OIResponse by SP's public key after checking the authentication and integrity of the message as well as the state of customer's account in case of prepaid account. These messages are sent to SP in step 2. In step 3, if SP accepts payment request of MC, he will create a Bill and asks his EMPS EMPS-SP to contact with EMPS-MC. Next, EMPS-SP transfers this Bill along with information about its account to EMPS-MC. EMPS-MC will pay for EMPS-SP through this account. This transaction will depend

on the deal between 2 EMPSs. EMPS-MC charges MC through her phone bill and notifies MC of payment completion in step 5.

Macro Payment Protocol: In contrast to micro payment which is charged through phone bill, macro payment is paid by bank account or card. Therefore, this protocol will require some information involving in customer account or card. Some assumptions are made for this case. First, MC shares information about account or his card (AI) with the FI/B which issues the card or account of MC. The second assumption is that MC is also given a PIN shared between MC and FI/B only. When MC registers with EMPS-MC for macro payment service, she must supply information about her FI/B. The macro payment protocol is very similar to the micro payment protocol and the difference between them is slim. The first difference lies in the information sent to EMPS-MC in step 1. Besides the information like in step 1 in micro payment protocol, MC also sends z_{map}, a HOAC used to show FI/B that the request for payment is from MC. The other differences are found in the internal processes of EMPS-MC at step 2' and 5'. Prior to transferring the request for payment of MC to SP, EMPS-MC will examine the financial situation of MC by sending the OIResponse and z_{map} to FI/B (step 2'.1). FI/B of EMPS-MC will contact with the FI/B of customer to get information. This process happens under the banking private network. The result will be returned to EMPS-MC in step 2'.2. If the result is positive, EMPS-MC will proceed to the remaining steps like in micro payment. The last difference is in step 5'. EMPS-MC requires its FI/B to link to FI/B of MC to conduct the transaction. The result is returned to EMPS-MC in step 5'.2. The final step of this protocol is the same as step 5 in micro payment protocol.

1. MC → EMPS-MC: f_{map}, l, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, $ID_{emps-sp}$, dl_{map}, ts_{mc}, d_{mc-map}, $EK_{mc,sp}$(OIResponse), y_{map}, x_{map}, z_{map}
y_{map} = h(OIResponse, l, dl_{mip}, $k_{mc,\ sp}$, d_{mc-map}, h($k_{mc,\ emps-mc}$)) is a HOAC indicating to SP that the original of the request for payment is from MC.
z_{map} = h(OIResponse, AI, dl_{map}, PIN) is a HOAC indicating to FI/B that the original of the request for payment is from MC.
x_{map}= h(OIResponse, ts_{mc}, $k_{mc,emps-mc}$, y_{map}, z_{map}) is a HMAC showing EMPS-MC that the HOAC y_{map} and z_{map} are from MC.

2'. Inside EMPS-MC
2'.1. EMPS-MC → FI/B: f_{map}, ID_{mc}, OIResponse, z_{map}, dl_{map}, d_{mc-map}, $s_{emps-mc}$(OIResponse, z_{map}, dl_{map}, d_{mc-map})
2'.2. FI/B → EMPS-MC: f_{map}, ID_{mc}, OIResponse, Result, $s_{fi/b}$(OIResponse, z_{map}, Result)
Result = Yes/No

2. EMPS-MC → SP: f_{map}, l, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, dl_{map}, $ts_{emps-mc}$, d_{mc-map}, E_{sp}(OIResponse), y_{map}, x_{map}, $signmap_{emps-mc}$
$signmap_{emps-mc}$ = $s_{emps-mc}$(f_{map}, l, dl_{map}, $ts_{emps-mc}$, d_{mc-map}, OIResponse, y_{map}, x_{map})

3. SP → EMPS-SP: f_{map}, l, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, $ID_{emps-sp}$, d_{mc-map}, $E_{emps-sp}$ (Bill), s_{sp}(f_{map}, l, Bill, d_{mc-map})
Bill = OIResponse, Approval

4. EMPS-SP → EMPS-MC: f_{map}, l, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, $ID_{emps-sp}$, $E_{emps-mc}$ (Bill, $Account_{emps-sp}$), $s_{emps-sp}$(f_{map}, l, Bill, $Account_{emps-sp}$)

5'. Inside EMPS-MC
5'.1. EMPS-MC → FI/B: f_{map}, ID_{mc}, OIResponse, dl_{map}, $E_{fi/b}$ (Bill, $Account_{emps-sp}$), $s_{emps-mc}$ (f_{map}, Bill, $Account_{emps-sp}$)
5'.2. FI/B → EMPS-MC: f_{map}, ID_{mc}, OIResponse, dl_{map}, Result, $s_{fi/b}$(f_{map}, Bill, Result)
Result = Yes/No

5. EMPS-MC → MC: f_{map}, l, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, $ID_{emps-sp}$, $Ek_{mc,\ emps-mc}$(Bill, Result), $s_{emps-mc}$(f_{map}, l, Bill, Result)

Content Delivery Protocol

This protocol is based on the assumption that the content delivered is the electronic goods, for example software, music, films, financial report, and can be considered as a message m in general. If the content is physical goods, we can use the traditional delivery method such as transportation companies and there is no concern of the repudiation problem. The protocol is divided into three sub-protocols, a main, a recovery and an abort protocol. In case of problems, the abort or recovery protocol can be involved.

1. SP → MC: f_{cd}, l, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, $ID_{emps-sp}$, c, $E_{emps-sp}$ ($Ek_{mc,\ sp}$ (k)), EOO_c
EOO_c = s_{sp} (f_{cd}, l, c, $E_{emps-sp}$ ($Ek_{mc,\ sp}$ (k)))

2. MC → EMPS-MC: f_{cd}, l, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, $ID_{emps-sp}$, dl_{cd}, ts_{mc}, d_{mc-cd}, h(c), h($E_{emps-sp}$ ($Ek_{mc,\ sp}$ (k))), y_{cd}, x_{cd}
y_{cd} = h(dl_{cd}, h(c), h(EK), dl_{cd}, d_{mc-cd}, $k_{mc,sp}$, h($k_{mc,emps-mc}$)) is a HOAC indicating to SP that the response for cipher c is from MC.
x_{cd} = h(f_{cd}, l, ts_{mc}, h(c), h(EK), $k_{mc,emps-mc}$, y_{cd}) is a HMAC indicating to EMPS-MC that the HOAC y_{cd} is from MC.

3. EMPS-MC → SP: f_{cd}, l, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, $ID_{emps-sp}$, dl_{dc}, $ts_{emps-mc}$, d_{mc-cd}, h(c), h($E_{emps-sp}$ ($Ek_{mc,\ sp}$ (k))), y_{cd}, x_{cd}, $signcd_{emps-mc}$
$signcd_{emps-mc}$ = $s_{emps-mc}$(f_{cd}, l, dl_{dc}, $ts_{emps-mc}$, d_{mc-cd}, h(c), h($E_{emps-sp}$ ($Ek_{mc,\ sp}$ (k))), y_{cd}, x_{cd})
If SP times out then abort

4. SP → MC: f_{cd}, l, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, $ID_{emps-sp}$, $Ek_{mc,\ sp}$ (k), s_{sp} (f_{cd}, l, k)
If MC times out then recovery

Main Protocol: SP sends the cipher of message and the evidence of origin for cipher EOO_c to MC in step 1 and waits for the non-repudiation of receipt evidence NRR. If MC carries out the step 2, the step 3 must be executed because EMPS-MC is a trust party of MC. So we can consider step2

and step 3 as two small steps in a unique step 2-3. After receiving NRR in step 3, SP will give the decryption key k to MC.

Abort Protocol: If SP doesn't receive the third message of the main protocol, SP initiates the abort protocol by sending to EMPS-SP an Abort request.

1. SP → EMPS-SP: f_{abort}, l, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, $ID_{emps-sp}$, Abort
Aborted = true
2. SP → MC: f_{abort}, l, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, $ID_{emps-sp}$, Abort

Recovery Protocol: MC executes the recovery protocol if she does not receive the message in step 4 of the main protocol. MC asks EMPS-MC to transfer its recovery request to EMPS-SP. EMPS-SP recovers the decryption k and sends it along with evidence back to MC through EMPS-MC.

1. MC → EMPS-MC: f_{rec}, l, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, $ID_{emps-sp}$, dl_{cd}, ts_{mc}, d_{mc-cd}, $E_{emps-sp}$ ($Ek_{mc,\ sp}$ (k)), h(c), y_{cd}, x_{cd}
If aborted or recovered then stop, Else recovered = true

2. EMPS-MC → EMPS-SP: f_{rec}, l, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, $ID_{emps-sp}$, $E_{emps-sp}$($Ek_{mc,\ sp}$ (k)), $s_{emps-mc}$ (f_{rec}, l, $E_{emps-sp}$ ($Ek_{mc,\ sp}$ (k)))
If aborted or recovered then stop, Else recovered = true

3. EMPS-SP → EMPS-MC: f_{rec}, l, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, $ID_{emps-sp}$, $E_{emps-mc}$($Ek_{mc,\ sp}$ (k)), $s_{emps-sp}$ (f_{rec}, l, $Ek_{mc,\ sp}$ (k))

4. EMPS-SP → SP: f_{rec}, l, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, $ID_{emps-sp}$, $s_{emps-sp}$ (f_{rec}, l, $Ek_{mc,\ sp}$ (k))

5. EMPS-MC → MC: f_{rec}, l, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, $ID_{emps-sp}$, $Ek_{mc,emps-mc}$($Ek_{mc,\ sp}$ (k)), $s_{emps-mc}$ (f_{rec}, l, $Ek_{mc,\ sp}$ (k))

C. Non-Repudiation Analysis

Non-Repudiation Analysis of Price Negotiation Protocol

Non-repudiability: The non-repudiation of origin and receipt evidences for OI_{mc} are NRO_{pn} = $signpn_{emps-mc}$ = $s_{emps-mc}$ (f_{pn}, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, dl_{pn}, d_{mc-pn}, $ts_{emps-mc}$, $Ek_{mc,sp}(OI_{mc})$, y_{pn}, x_{pn}) and NRR_{pn} = s_{sp}(OIResponse, pk_{sp}, l). If MC denies having sent OI_{mc}, SP has to present to the judge f_{pn}, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, dl_{pn}, d_{mc-pn}, $ts_{emps-mc}$, OI_{mc}, y_{pn}, x_{pn}, $k_{mc,sp}$, NRO_{pn}. The judge verifies that $Ek_{mc,sp}(OI_{mc})$ is the cipher of OI_{mc} under the session key $k_{mc,sp}$, NRO_{pn} is the signature of EMPS-SC on (f_{pn}, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, dl_{pn}, d_{mc-pn}, $ts_{emps-mc}$, $Ek_{mc,sp}(OI_{mc})$, y_{pn}, x_{pn}). As HOAC y_{pn} contains $k_{mc,sp}$, it can be created by only MC and SP. Similarly, HMAC x_{pn} must be generated by only MC and EMPS-MC because of $k_{mc,emps-mc}$. Therefore, it must be MC who produces both of HOAC y_{pn} and HMAC x_{pn}. If SP can present all of the items and all the check hold, the adjudicator concludes that MC is at the origin of OI_{mc}. If SP denies receipt of OI_{mc} and offered prices of products in OI_{mc}, MC gives the judge NRR_{pn}, $Ek_{mc,sp}$(OI_{mc}, OIResponse, pk_{sp}, l), OIResponse, l, OI_{mc}, pk_{sp}. The judge checks that $Ek_{mc,sp}$(OI_{mc}, OIResponse, pk_{sp}, l) is the cipher of (OI_{mc}, OIResponse, pk_{sp}, l) under the session key $k_{mc,sp}$ and NRR_{pn} is the signature of SP on (OI_{mc}, OIResponse, pk_{sp}, l). If all checks are valid, the adjudicator claims that SP received the OI_{mc} and replied with OIResponse.

Fairness: If SP does not send message in step 3, the protocol will not be strong fairness. However, if step 3 is not executed, SP will lose its customer and gain nothing from that. In other words, SP would harm himself. Consequently, he should carry out step 3 and that means the strong fairness feature of the protocol can be achieved.

Non-Repudiation Analysis of Payment Protocol

Non-repudiability: Non-repudiation evidences of micro payment protocol are NRO_{mip} = $signmip_{emps-mc}$ = $s_{emps-mc}$ (f_{mip}, l, dl_{mip}, $ts_{emps-mc}$, d_{mc-mip}, OIResponse, $y_{mip,}$ x_{mip}) and NRR_{mip} = $s_{emps-mc}$ (f_{mip}, l, Bill). In case of non-repudiation of origin, SP has to present to a judge NRO_{mip}, f_{mip}, l, dl_{mip}, $ts_{emps-mc}$, d_{mc-mip}, OIResponse, $y_{mip,}$ x_{mip}, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, $ID_{emps-sp}$, h(OIResponse), $h(k_{mc,\ emps-mc})$. The judge verifies that l = $h(ID_{mc}$, ID_{sp}, $ID_{emps-mc}$, $ID_{emps-sp}$, h(OIResponse), $h(k_{mc,\ emps-mc}))$, and NRO_{mip} is the signature of EMPS-MC on (f_{mip}, l, dl_{mip}, $ts_{emps-mc}$, d_{mc-mip}, OIResponse, $y_{mip,}$ x_{mip}). If all the checks hold, the adjudicator concludes that MC is the originator of payment request. On the other hand, MC can prove that SP has received her payment by presenting NRR_{mip}, f_{mip}, l, ID_{mc}, ID_{sp}, $ID_{emps-mc}$, $ID_{emps-sp}$, d_{mc-mip}, h(OIResponse), $h(k_{mc,\ emps-mc})$, Bill to the judge. EMPS-MC provides $s_{emps-sp}$ (f_{mip}, l, Bill, $Account_{emps-sp}$) and EMPS-SP provides s_{sp}(f_{mip}, l, Bill, d_{mc-mip}), $Account_{emps-sp}$. The arbitrator checks that l = $h(ID_{mc}$, ID_{sp}, $ID_{emps-mc}$, $ID_{emps-sp}$, h(OIResponse), $h(k_{mc,\ emps-mc}))$, s_{sp}(f_{mip}, l, Bill, d_{mc-mip}) is the signature of SP on (f_{mip}, l, Bill, d_{mc-mip}), $s_{emps-sp}$(f_{mip}, l, Bill, $Account_{emps-sp}$) is the signature of EMPS-SP on (f_{mip}, l, Bill, $Account_{emps-sp}$) and NRR_{mip} is the signature of EMPS-MC on (f_{mip}, l, Bill). If all verifications are valid, MC wins. We can have similar verifications in macro payment protocol with NRO_{map} = $signmap_{emps-mc}$ = $s_{emps-mc}$(f_{map}, l, dl_{map}, $ts_{emps-mc}$, d_{mc-map}, OIResponse, $y_{map,}$ x_{map}) and NRR_{map} = $s_{emps-mc}$(f_{map}, l, Bill, Result).

Fairness: These two payment protocols are strong fairness because the transaction are intervened by the trust parties of both MC and SP. EMPS-MC represents MC, EMPS-SP represents SP and the payment actually happens among these EMPSs which trust each other.

Non-repudiation Analysis of Content Delivery Protocol

Non-repudiability: If the recovery protocol is not invoked, NRO_{cd} = s_{sp} (f_{cd}, l, c, $E_{emps-sp}$ ($Ek_{mc,\ sp}$ (k))), s_{sp} (f_{cd}, l, k) and NRR_{cd} = $signcd_{emps-mc}$ = $s_{emps-mc}$(f_{cd}, l, dl_{dc}, $ts_{emps-mc}$, d_{mc-cd}, h(c), h($E_{emps-sp}$ ($Ek_{mc,\ sp}$ (k))), y_{cd} , x_{cd}). On the contrary, NRO_{cd} = s_{sp} (f_{cd}, l, c, $E_{emps-sp}$ ($Ek_{mc,\ sp}$ (k))), $s_{emps-mc}$ (f_{rec}, l, $Ek_{mc,\ sp}$ (k)) and NRR_{cd} = $s_{emps-mc}$(f_{cd}, l, dl_{dc}, $ts_{emps-mc}$, d_{mc-cd}, h(c), h($E_{emps-sp}$ ($Ek_{mc,\ sp}$ (k))), y_{cd} , x_{cd}), $s_{emps-sp}$ (f_{rec}, l, $Ek_{mc,\ sp}$ (k)). The checking process is similar to the above verifications.

Fairness: The protocol is strong fairness. If the step 2-3 of main protocol is not executed, SP can invoke the abort protocol and no party can obtain correct evidence anymore. If MC launches a recovery protocol, both MC and SP will receive all expected evidences, and hence the protocol remains fair.

Timeliness: Timeliness is provided by the fact that at each moment in the protocol, both MC and SP can stop the protocol while preserving fairness.

V. Conclusions and Future Work

In this paper we have identified the most common requirements and properties of concern for dealing with non-repudiation problem in m-commerce. Then, we have introduced an extension of mobile payment model named EMPS for solving the above problem. Based on the EMPS, three non-repudiation protocols for *all fundamental phases* (price negotiation, mobile payment, content delivery) in m-commerce transactions have been built. To the best of our knowledge, this holistic approach to the non-repudiation problem in m-commerce is among the vanguard solutions to address it. Analyses of the non-repudiability, fairness and timeliness of the proposed protocols are also carried out. They are the solid basis for our further improvements in the future.

In the future, we plan to standardize communications among parties in EMPS model, especially between MNO and FI/B and among EMPSs. Web service standard is one of our targets for this purpose. Additional formal analyses along with improvements to achieve the timeliness of the proposed non-repudiation protocols will be of our great interest. Moreover, generating long-term key from session key, combining the initiation process with price negotiation process, etc. are also of our concerns. We also intend to perform empirical evaluations on the proposed protocols' performance to establish their practical value.

Acknowledgment

We would like to thank all members of ASIS Lab at CSE/HCMUT for their enthusiastic supports during carrying out this research.

References

[1] L. He and N. Zhang, "A New Signature Scheme: Joint Signature," in *ACM Symposium on Applied Computing*, 2004, pp. 807 – 812.

[2] J. Liu, J. Liao, and X. Zhu, "A System Model and Protocol for Mobile Payment," in *Proc. of IEEE International Conference on e-Business Engineering*, 2005, pp. 638 – 641.

[3] R. K. Tiwari, "Fair Non Repudiation in Mobile Communication using Joint Signatures," in *Proc. of IEEE International Conference on Personal Wireless Communication*, 2005, pp. 438 – 440.

[4] L. He and N. Zhang, "An asymmetric authentication protocol for M-Commerce applications," in *Proc. of IEEE International Symposium on Computers and Communication*, Vol. 1, 2003, pp. 244 – 250.

[5] C. Chen, H. Lin, Y. Chen, and J. Jan, "A Fair Transaction Model in Mobile Commerce," in *Proc. of IEEE International Symposium on Signal Processing and Information Technology*, 2006.

[6] S. Kremer, O. Markowitch, and J. Zhou, "An Intensive Survey of Non-repudiation Protocols," *Computer Communications*, pp. 1606 – 1621, 2002.

[7] Jianying Zhou, "Non-repudiation in Electronic Commerce," Artech House Computer Security Series, 2001.

[8] S. Kungpisdan, B. Srinivasan, and P. D. Le, "A Secure Account-Based Mobile Payment Protocol," in *Proc. of International Conference on Information Technology: Coding and Computing*, 2004, pp. 35 – 39.

[9] A. Vilmos and S. Karnouskos, "SEMOPS: Design of a New Payment Service," in *Proc. of International Workshop on Database and Expert Systems Applications*, 2003, pp. 865 – 869.

[10] S. Nambiar and C.T. Lu, "M-Payment Solutions and M-Commerce Fraud Management," as Chapter IX of Book: Advances in Security and Payment Methods for Mobile Commerce, pp. 192 – 213, Idea Group Inc., 2005.

[11] C. Lee, W. Hu, and J. Yeh, "A System Model for Mobile Commerce", in *Proc. of International Conference on Distributed Computing Systems Workshops*, 2003, pp. 634 – 639.

[12] ISO/IEC 10181-4. Information Technology – Open Systems Interconnection – Security Frameworks in Open System – Part 4: Non-repudiation Framework, ISO/IEC, 1996.

[13] S. Nambiar, C.T. Lu, and L.R. Liang, "Analysis of Payment Transaction Security in Mobile Commerce," in *Proc. of IEEE International Conference on Information Reuse and Integration*, 2004, pp. 475 – 480.

Análisis Numérico de Pérdidas de Inserción de Conmutadores Diseñados con Diodos *p-i-n*

Alejandro Iturri Hinojosa, Cirilo Leon Vega, Gabriela Leija Hernández

***Resumen*—Se presenta un análisis numérico de la pérdida de inserción de conmutadores de microondas diseñados con diodos *p-i-n*. Se analizan las características de resistencia serie, R_s, y la capacitancia de unión, C_j, propias del modelo de circuito equivalente de los diodos *p-i-n*. Así mismo, se presenta a detalle la teoría de funcionamiento de los diodos *p-i-n* y de los conmutadores de microondas.**

***Palabras Clave*—Diodos *p-i-n*, conmutadores de microondas, pérdida de inserción.**

Numeric Analysis of the Insertion Loss in Switches Designed using the *p-i-n* Diodes

***Abstract*—We present numeric analysis of the insertion loss in the microwave switches designed using the *p-i-n* diodes. We analyze the characteristics of series resistance R_s, and junction capacitance C_j that are part of the equivalent circuit model of the *p-i-n* diodes. Also, we present necessary background for explanation of functioning of the *p-i-n* diodes and microwave switches.**

***Index Terms*—*p-i-n* diodes, microwave switches, insertion loss.**

I. Introducción

Un diodo *p-i-n* es un diodo semiconductor que consta de dos regiones, una tipo *P* y otra tipo *N* altamente dopadas y separadas por una región intrínseca de mayor resistividad, como se puede apreciar en la Fig. 1. Estos dispositivos son muy utilizados en desplazadores de fase y conmutadores de señales microondas. Los dispositivos diseñados con diodos *p-i-n* se destacan por sus bajas pérdidas de inserción y elevado desempeño con señales de altas frecuencias [1].

La principal ventaja del diodo p-i-n frente a un diodo convencional es la mejora en la respuesta de conmutación de señales microondas.

Cuando se manejan señales de baja frecuencia los efectos reactivos que se presentan en las uniones del diodo se consideran despreciables. Estos efectos son asociados con la difusión de portadores, electrones y huecos, a través de la unión.

Manuscrito recibido el 7 de mayo del 2009. Manuscrito aceptado para su publicación el 6 de octubre del 2009.

Los autores son de ICE – ESIME Zacatenco, Instituto Politécnico Nacional, México D. F., México (e-mail: aiturri@ipn.mx, cleonv@ipn.mx, gleijah@yahoo.com.mx).

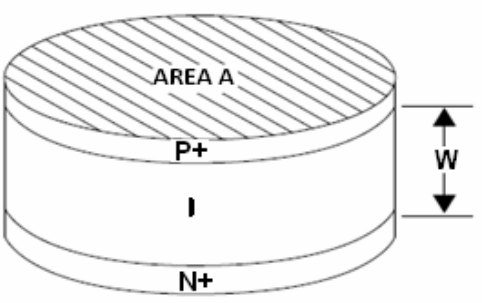


Fig. 1. Diagrama de un diodo *p-i-n* [6].

II. Características de Funcionamiento de los Diodos *P-I-N*

En estado de baja impedancia, situación de polarización directa, el diodo tiene una excelente linealidad y baja distorsión. En estado de alta impedancia, situación de polarización inversa, la región intrínseca produce valores muy altos de voltaje de ruptura y de impedancia. A medida que el ancho de la región de agotamiento, región I, aumenta, la capacitancia formada en las uniones del diodo disminuye. Así, el diodo se comporta como un circuito abierto.

A. Estructura y Funcionamiento del Diodo p-i-n

Un material altamente dopado es aquel que tiene un mayor número de impurezas, generalmente de tipo P o de tipo N. Por lo mismo, ofrecerá una menor resistencia al paso de la corriente. Para un material no dopado, es decir intrínseco, existirá una resistencia mucho mayor, dependiendo del material semiconductor que se esté utilizando [1].

En la práctica un diodo *p-i-n* tiene una alta resistividad en la parte media de la zona P o N. Mientras que existe una baja resistividad en los límites en las zonas P y N.

La nomenclatura P+ y N+ indica un alto dopaje de los materiales P y N, respectivamente. Un material tipo "π" indica que el material es de tipo P y además se dice que es un material "no dopado" (idealmente). Por otra parte, un material "N" no dopado es llamado material tipo "ν". El material usado en la región I puede ser tipo "π" o "ν". Utilizando cualquiera de estas dos estructuras no se presentan cambios en el desempeño de un dispositivo. En la práctica, generalmente se utiliza el silicio como el material semiconductor, el cual no es totalmente intrínseco.

En la Fig. 2 se presentan dos estructuras posibles del diodo *p-i-n* (P+, π, N+) y (P+, ν, N+).

En la Fig. 2b se muestra el perfil de impurezas de un diodo *p-i-n* con estructura (P+, ν, N+), en el cual la región intrínseca de alta resistividad concentra pocos átomos de impurezas tipo N que se ionizan, mientras que la región de agotamiento se extiende a lo largo de la región intrínseca incluyendo una pequeña cantidad de penetración en las regiones conductoras. La región de agotamiento no se extenderá mas allá de los límites de la región I debido al elevado dopaje de las regiones P+ y N+, siendo la zona de agotamiento esencialmente igual al ancho de la capa I, "W". La unión PN que se forma será en la zona P+ [1].

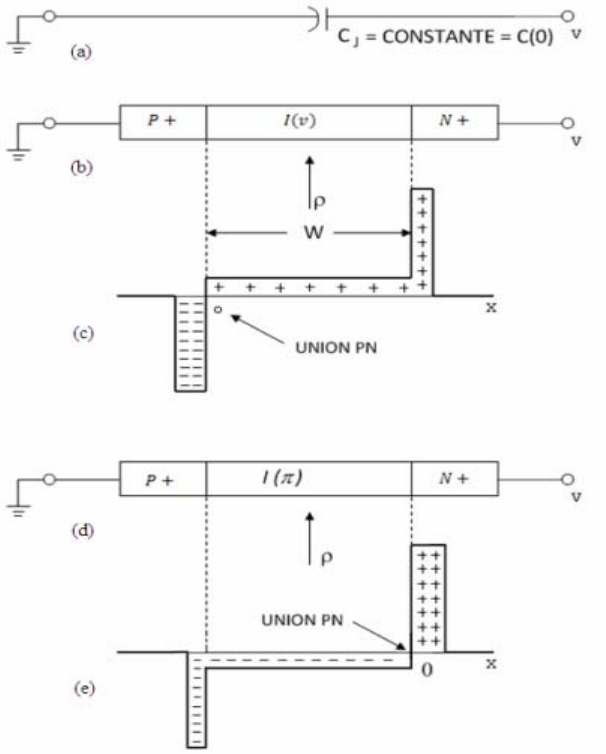


Fig. 2. Perfiles de los dos tipos de diodo *p-i-n*.(a) capacitancia aproximada de circuito equivalente, (b) cristal *p-ν-n*, (c) perfil de impurezas *p-ν-n*, (d) cristal *p-π-n*, (e) perfil de impurezas *p-π-n* [1].

Una característica importante del diodo *p-i-n* radica en la ampliación de la zona de agotamiento que se obtiene con la ionización de la misma.

En la Fig. 2d se muestra la estructura (P+, π, N+) la cual tiene una región intrínseca con concentración de impurezas de material tipo P+. La unión PN que se forma por la expansión de la zona de agotamiento se encuentra en la zona N+.

B. Voltaje de Perforación

Como la región intrínseca es altamente resistiva, la zona de agotamiento se extiende hasta las regiones de alta conducción, incluso aún cuando no se ha polarizado al diodo. Se dice que la capacitancia equivalente del diodo no se ve alterada por el voltaje suministrado, Fig. 3b.

A causa de altas concentraciones de impurezas y de la facilidad de ionización de los electrones y los huecos en los materiales P+ y N+ la zona de agotamiento tenderá a ensancharse más.

Un diodo libre de voltaje de polarización tiene la característica de tener una región I que se va extendiendo hacia las zonas P+ y N+ sin requerir la aplicación de algún voltaje.

Sin voltaje de polarización parte de las impurezas en la región I del diodo se ionizan y la zona de agotamiento cubre una parte de la capa intrínseca (Fig. 3b). Por otra parte en polarización inversa la capa de agotamiento se propaga por la zona intrínseca a medida que la capacitancia por unidad de área en la unión PN va decreciendo. Como se muestra en la Fig. 2c el ancho de la capa de agotamiento es casi igual al ancho de la capa intrínseca. Y el voltaje en este punto de operación es llamado voltaje de perforación (V_{PT}).

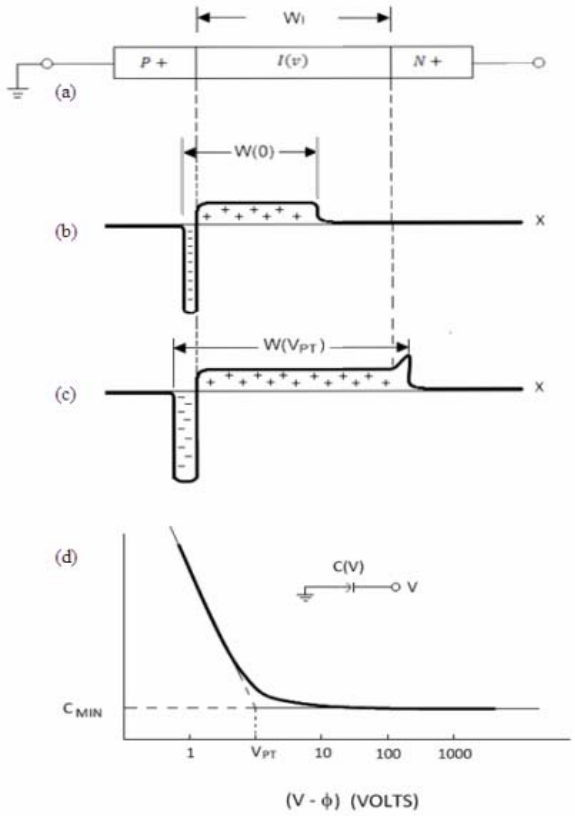


Fig. 3. Características del diodo *p-i-n* y su respuesta de voltaje de perforación. a) diodo *p-i-n* práctico, b) perfil de impurezas ionizadas en estado de polarización cero, c) perfil de impurezas ionizadas en estado de perforación, d) característica C(V) a 1 MHz [1].

En la práctica, para medir el voltaje de perforación se utiliza la característica de capacitancia vs. voltaje del diodo que se forma en la unión PN, ver Fig. 3d. La gráfica *C* vs. *V* muestra la relación cuando el diodo opera a alrededor de 1 MHz, considerada bajas frecuencias.

La medición del voltaje de perforación se obtiene en la intersección de las tangentes de los declives, como se observa en la Fig. 3d.

Sin embargo, el funcionamiento con microondas dependerá de la susceptibilidad del material semiconductor que se esté utilizando. El silicio presenta una susceptibilidad mayor que la conductancia de las impurezas en el material intrínseco.

C. Medición de Capacitancia y Relajación Dieléctrica

Debido a la elevada constante dieléctrica del silicio y también por ser un material semiconductor con conductancia variable, la capacitancia será mayor cuando el diodo opere con señales de baja frecuencia. Mientras que para señales de frecuencia alrededor de 1 GHz la capacitancia medida será mucho menor [1].

Por lo tanto, el circuito equivalente que representa al diodo *p-i-n* está conformado por el paralelo de una capacitancia y una conductancia. La división de corriente entre ellos varía con la frecuencia de la señal aplicada. Las corrientes de mayor frecuencia se conducirán mayormente por el trayecto capacitivo.

En la Fig. 4 se muestra la representación de un diodo *p-i-n* por debajo del voltaje de perforación. Las regiones de P+ e I que son reducidas, representan a la zona de agotamiento o región libre de portadores. La zona restante de la región I está intacta y puede modelarse por el circuito de la Fig. 4c, como el circuito en paralelo de una resistencia y un capacitancia, representados por los elementos C_{US} y R_{US}, respectivamente.

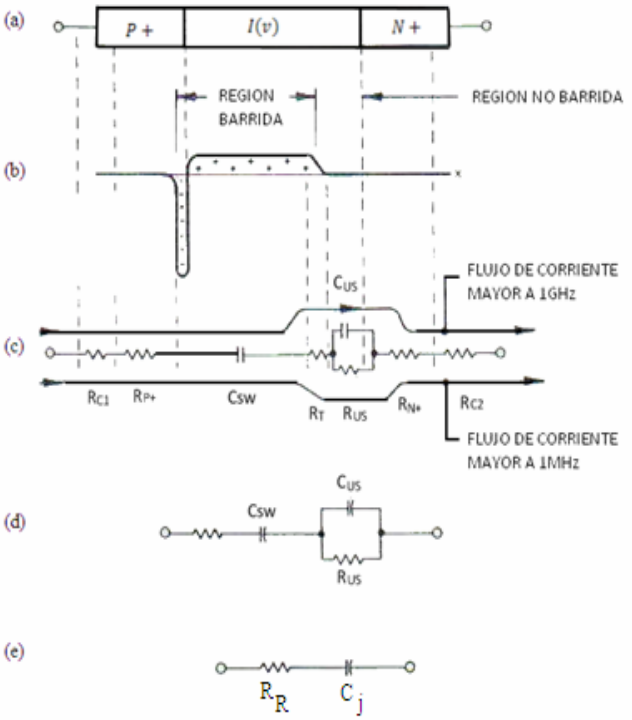


Fig. 4. Circuito Equivalente de un diodo *p-i-n* en polarización Inversa. (a) modelo del diodo *p-v-n*, (b) zona de agotamiento antes del voltaje de perforación, (c) circuito equivalente detallado, (d) circuito equivalente a bajas frecuencias, (e) circuito equivalente para microondas [1].

La división de corriente a través de C_{US} y R_{US} depende de la relación de la susceptancia de C_{US} y de la conductancia ($1/R_{US}$). Esta relación depende directamente de la constante dieléctrica que se crea en el material semiconductor. La constante de relajación dieléctrica, f_R, se define como la frecuencia a la cual la división de corriente entre los dos elementos es la misma, es decir, cuando los valores de susceptancia y conductancia son iguales.

Una característica importante en el funcionamiento de los diodos *p-i-n* operados a una frecuencia de microondas como se muestra en la Fig. 4e, es la capacitancia C_J que se forma cuando se maneja una frecuencia del triple o más alta que f_R, esto debido al efecto que produce tener varios capacitores conectados en serie.

Por lo tanto las mediciones de capacitancia pueden desarrollarse a una frecuencia baja pero con un voltaje suficiente para poder operar adecuadamente el diodo. Así los resultados que se obtienen a una frecuencia de 1 MHz son una buena aproximación a los obtenidos a microondas.

Esta frecuencia será una estimación de la frecuencia de relajación debido a que en esta la capacitancia mínima es aproximada al valor de C_j (capacitancia para microondas).

Los diodos *p-i-n* fabricados con material de silicio altamente puro, tienen una resistividad de aproximadamente 500 a 100 KΩ-cm. Valores típicos de resistividad de la región I son de 100 a 1000 Ω-cm. En la práctica, los diodos *p-i-n* utilizados para la conmutación de microondas tienen anchos de región I de alrededor de 25 a 250 μm.

La frecuencia de relajación del dieléctrico del perfil del diodo mostrado en la Fig. 4 se obtiene mediante.

$$f_{R,GHz} = \frac{153}{\rho_{\Omega-cm}} \qquad (1)$$

Por ejemplo, un diodo *p-i-n* con una resistividad de al menos 100 Ω-cm en la capa intrínseca, tendrá aproximadamente una $f_R = 1.53$GHz. Así para poder trabajar con frecuencias de 5GHz o más, se utiliza comúnmente el circuito que se muestra en la Fig. 4e.

En donde "ρ" es la resistividad que se forma por el ensanchamiento de la zona de agotamiento comúnmente llamado "*bulk resistivity*" en el área libre de carga.

$$\rho_{\Omega-cm} = \frac{2.4\times10^{8} W^{2}}{V_{PT}} \qquad (2)$$

Donde W, es el espesor de la región I y V_{PT} es el voltaje de perforación.

El espesor W correspondiente a un valor de tiempo de vida de portadores (τ) del diodo *p-i-n*, puede encontrarse por medio de la relación [4]:

$$W = \sqrt{\mu_{AP} V_T \tau}\,(W/L) \qquad (3)$$

El voltaje de perforación es calculado por la siguiente expresión [1]:

$$V_{PT} = \frac{eN_D W^2}{2\varepsilon_0 \varepsilon_R} \quad (4)$$

Donde e, es la carga de un electrón, ε_R es la constante dieléctrica del material utilizado en el diodo *p-i-n* (ε_R =11.8 para el silicio) y N_D es la concentración de impurezas en la región I.

D. Tiempo de Vida de Portadores

Para poder calcular la resistencia de la región I del diodo con señales de microondas, se debe tener en cuenta la carga almacenada en esta región. La carga, consiste en electrones y huecos que se inyectan en la región I cuando se aplica una polarización directa al diodo [3].

El tiempo de vida de portadores "τ" es el tiempo que existe a partir de la inyección de huecos y electrones en la región I cuando se polariza directamente el diodo, hasta el instante en que se recombinan totalmente las cargas [2].

De otra manera, el tiempo de vida es proporcional a la improbabilidad de que un electrón y un hueco se recombinen [2]. Por ejemplo, un cristal de silicio puro tiene un tiempo de vida de portadores calculado de 3.7 s.

Con impurezas dopantes de 10^{15} cm^{-3} se tiene un tiempo de vida de portadores de 0.1 ms [2]. El tiempo de vida en diodos actuales, oscila entre 0.1 y 10 microsegundos ordenes en magnitud por debajo de este valor teóricamente obtenido [2].

Para mantener una conducción de la señal de microondas y una densidad de carga, se requiere de poca corriente de polarización directa y de un tiempo de vida de portadores mayor [1].

Un tiempo de vida de portadores largo no necesariamente implica una velocidad de conmutación lenta [2].

En la Fig. 5a se muestra un circuito equivalente para la medición del tiempo de vida de portadores τ. El método consiste en inyectar una cantidad de carga conocida "Q_0" en la región I y medir el tiempo "τ_S" que se necesita para extraer esta carga usando una corriente de polarización inversa.

El funcionamiento de este circuito se basa en el valor de la resistencia R_R en comparación con R_F para cuando S está abierto, el diodo estará polarizado directamente. Para S cerrado, V_R es aplicado al diodo y este a su vez produce una corriente inversa de magnitud I_R-I_F. Después la carga almacenada se remueve hasta dejar completamente agotada esta zona. Si se cumple $\tau_S << \tau$ donde τ_S es el periodo de descarga, entonces ocurre una recombinación de electrones y huecos despreciable durante el apagado, y la carga total almacenada es recuperada. En este caso $Q_0=I_F*\tau$ y τ será

$$\tau \approx \tau_S \left(\frac{I_R}{I_F - 1} \right) \quad (5)$$

donde $\tau_s << \tau$.

E. Característica de Voltaje - Corriente del Diodo p-i-n

En la Fig. 6 se muestra la característica de V-I del diodo *p-i-n* cuando apera con grandes cantidades de corriente de microondas. Se observa que tiene la misma respuesta a voltajes altos.

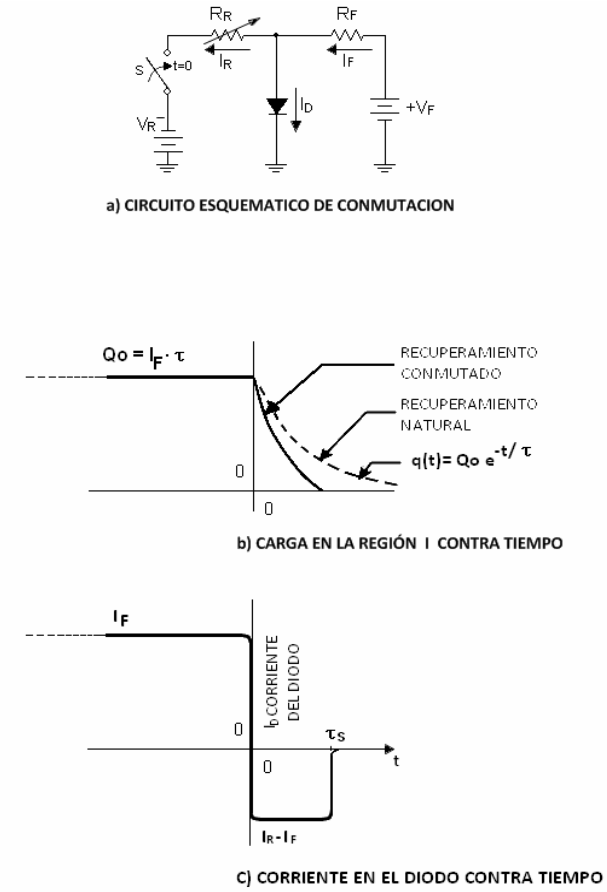


Fig. 5. Método de medición del tiempo de vida [2]

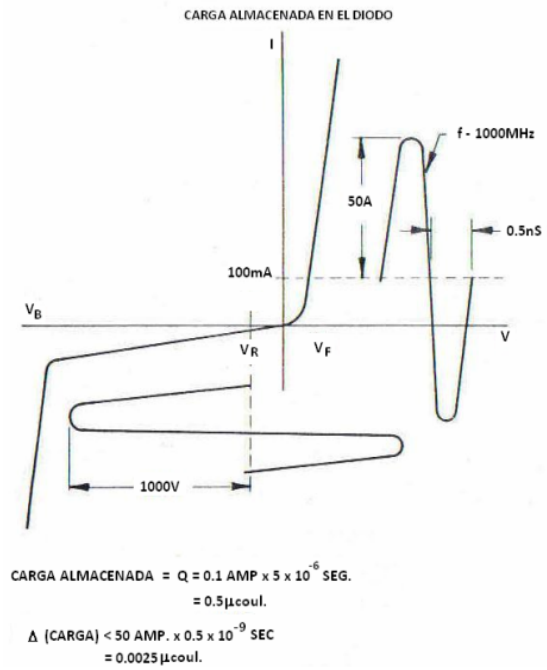


Fig. 6. Ejemplo de carga almacenada contra movimiento de carga debido a una señal de microonda [2].

La resistencia que se forma con una corriente de 100mA será menor que 1Ω. También se observa la linealidad que presenta este dispositivo incluso para altas corrientes [2].

$$Q = I_0 \tau \tag{6}$$

Por ejemplo, asumiendo un tiempo de vida de portadores de 5 μs típico de un voltaje elevado aplicado al diodo, unos 100 mA producen una carga de 0.5 μC.

F. Resistencia de Polarización Directa

La resistencia relacionada con la región intrínseca se puede calcular con la siguiente expresión, válida para altas frecuencias [4]:

$$R_I = \frac{W^2}{2\mu_{AP}\tau I_0} \tag{7}$$

donde $V_T = KT/q = 0.025852$, μ_{AP} es la movilidad ambipolar de las cargas [2]:

$$\mu_{AP} = \frac{2\mu_p\mu_n}{\mu_p + \mu_n} \tag{8}$$

Por ejemplo, para el silicio $\mu_{AP} = 610$ cm^2/V-s [2].

Y τ es el tiempo de vida de portadores.

Para frecuencias mayores a $1/\tau$ ($\approx 10/\tau$) se puede utilizar:

$$R_I = \left(\frac{W}{L}\right)^2 \frac{V_T}{2I_0} \tag{9}$$

Siendo

$$L_{AP} = \sqrt{D_{AP}\tau} \tag{10}$$

la longitud de difusión, que indica la variación de la densidad de portadores minoritarios de la región I.

A medida que la frecuencia de trabajo disminuye, los efectos de las uniones en el diodo pueden contribuir significativamente a la resistencia total R_T, de tal manera que R_I llega a ser solamente una pequeña parte de R_T, a bajas frecuencias [5].

En la región intrínseca, los efectos reactivos causados por la modulación de la conductividad también contribuyen significativamente a la impedancia R_T del diodo *p-i-n* [2]. De tal forma que la resistencia total del diodo en situación de polarización directa será [5]:

$$R_T = R_I + 2R_J(f) \tag{11}$$

donde:

$$R_J(f) = \frac{KT}{qI_0}\beta \tanh\left(\frac{W}{2L}\right)\cos\left(\Phi - \frac{\Theta}{2}\right) \tag{12}$$

Con parámetros dados por:

$$\beta = \frac{\sqrt{\left[\coth a\left(1+\cot^2 b\right)\right]^2 + \left[\cot^2 b\left(1-\coth^2 a\right)\right]}}{\left(1+4\pi^2 f^2\tau^2\right)^{\frac{1}{4}}\left(\coth^2 a + \coth^2 b\right)} \tag{13a}$$

$$a = \frac{W}{2L}\left(1+4\pi^2 f^2\tau^2\right)^{\frac{1}{4}}\cos\left(\frac{\Theta}{2}\right) \tag{13b}$$

$$b = a\tan\left(\frac{\Theta}{2}\right) \tag{13c}$$

$$\Phi = \tan^{-1}\left[\frac{\cot b\cdot\left(1-\coth^2 a\right)}{\coth a\cdot\left(1+\cot^2 b\right)}\right] \tag{13d}$$

$$\Theta = \tan^{-1}(\omega\tau) \tag{13e}$$

G. Parámetros R_R y C_J en Modelo de Circuito para Polarización Inversa

Bajo polarización inversa el diodo *p-i-n* permanece como una capacitancia constante a microondas debido a que la región I está agotada.

El parámetro R_R en su conexión en paralelo o en serie se refiere a la presencia de perdidas disipativas localizadas en las uniones de los contactos óhmicos y en la resistencia formada por las regiones P+ y N+.

A causa de la alta constante dieléctrica relativa para el silicio ε_R=11.8, la capacitancia formada en la región intrínseca es pequeña y se calcula con la expresión [2]:

$$C_j \approx \frac{\varepsilon_0\varepsilon_R\pi D^2}{4W} \tag{14}$$

donde ε_0 es la permeabilidad en espacio libre $8.85 \times \frac{10^{-14}\,F}{cm}$, D es el diámetro de la unión y W es la anchura de la región I.

III. Análisis Numérico de Algunos Parámetros de Funcionamiento de Dispositivos de Microondas

A. Análisis Numérico de los Parámetros de los Diodos p-i-n

Utilizando la expresión (11), logramos predecir el ancho de región I de algunos diodos comerciales muy utilizados en sistemas de microondas. Los diodos estudiados se muestran en la tabla siguiente (Tabla I), con sus respectivas características de resistencia serie R_S y capacitancia de unión C_j dados por el fabricante.

Para los diodos mostrados en la tabla encontramos los valores de ancho W aproximados. La Tabla II muestra los resultados obtenidos para frecuencias de microondas, de 10 GHz. Para estos cálculos se utilizaron las expresiones (3) y (6).

TABLA I
CARACTERÍSTICAS DE RESISTENCIA SERIE Y CAPACITANCIA DE UNIÓN DE LOS DIODOS *P-I-N* UTILIZADOS EN LOS CONMUTADORES.

Diodo	R_s	$C_j(p_F)$
HPND4005	4.7	0.017
HPND0002	3.5	0.2
HPND4028	2.3	0.025
HPND4038	1.5	0.045
5082-0012	1	0.12

TABLA II
RESPUESTAS DE ANCHO *W* PARA ALGUNOS DIODOS *P-I-N*.

Diodo HPND	Rs (Ω)	Cj (pF)	τ (ns)	I_0 (mA)	Q (nF)	W/L	W (μm)	L (μm)
00002	3.5	0.2	1500	50	75	7.0	371	53
4005	4.7	0.017	100	10	1	2.3	31.4	13.7
4028	2.3	0.025	36	10	360	1.5	11.9	82
4038	1.5	0.45	45	10	450	1.1	10.5	918

Se puede observar que diodos con característica de impedancia serie menor, llegan a percibir menor ancho W de región I. Aumentando la corriente de polarización directa del diodo, podemos observar que el ancho de la región I incrementa de manera proporcional.

La longitud de difusión es inversamente proporcional a la impedancia serie del diodo *p-i-n*.

Diodos *p-i-n* con valores bajos de ancho W de región I, poseen un tiempo de vida de portadores bajo e impedancia serie igualmente baja. Una relación W/L baja en un diodo *p-i-n* se caracterizará por una impedancia serie igualmente baja. Un valor de R_S bajo, resulta en un buen desempeño del diodo *p-i-n* en dispositivos de microondas. A continuación presentamos las características de impedancia serie (R_S) vs. tiempo de vida de portadores (τ) para diodos *p-i-n* operando a una frecuencia de 10 GHz, con ancho W relativamente bajo, fabricados de los materiales semiconductores Silicio y Arseniuro de Galio. Para estos cálculos reemplazamos las propiedades eléctricas de los materiales en las expresiones (8) y (11).

En la Fig. 7 se puede observar que los diodos *p-i-n* de Arseniuro de Galio tienen mejor respuesta de resistencia serie respecto a los diodos de Silicio para un valor común de tiempo de vida de portadores. Esto debido principalmente a que cuentan con una permitividad eléctrica mayor.

Los diodos *GaAs* tendrán mejor desempeño en dispositivos de microondas.

1) *Pérdidas de inserción y aislamiento*

Las perdidas por inserción (IL) y aislamiento son parámetros importantes utilizados para evaluar el desempeño de los conmutadores de microondas. El presente trabajo trata la pérdida de inserción de conmutadores de microondas.

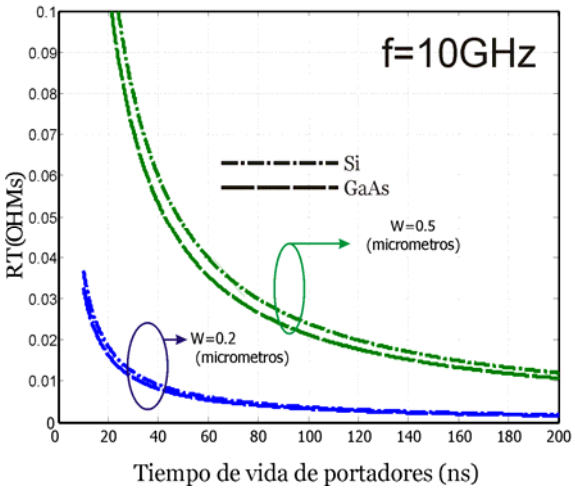


Fig. 7. Respuesta de R_S vs. τ en diodos *p-i-n* de Si y GaAs en altas frecuencias.

La pérdida de inserción se define como la relación, generalmente en decibelios, de la potencia entregada por un conmutador ideal en estado de conducción con la potencia real entregada por el conmutador en estado de conducción [15]. Es decir: $dL_{db} = 10 \log_{10}\left(\frac{P_{L,ON}^{ideal}}{P_{L,ON}^{real}}\right)$.

B. *Desempeño de los Diodos p-i-n en Conmutadores de Microondas*

Para el control de las señales de RF, mayormente se utilizan los diodos *p-i-n*. El uso de diodos *p-i-n* en diversas aplicaciones se basa en la polarización del diodo, en estado de alta o baja impedancia, dependiendo del nivel de carga que se encuentre almacenada en la región I.

La disposición de los diodos *p-i-n* en los conmutadores de un polo y un tiro, SPST, tipos serie y derivación es como se muestran en la Fig. 8.

Los conmutadores multi-tiro son mayormente utilizados. Un conmutador multi-tiro se puede diseñar usando un diodo *p-i-n* en cada brazo adyacente al punto común (GND). El desempeño que se obtiene puede ser mejorado con el uso de "conmutadores compuestos", los cuales consisten en la combinación en cada brazo de los conmutadores conectados en serie y los conmutadores conectados en derivación.

Para aplicaciones de banda estrecha, se usan líneas de transmisión de un cuarto de longitud de onda, separado por múltiples diodos, permitiendo ser usados para varios diseños de conmutadores y así lograr obtener una mejor operación de estos.

1) *Conmutadores conectados en serie*

Los conmutadores con diodos *p-i-n* conectados en serie, un polo – un tiro (SPST) y un polo – dos tiros (SPDT), mostrados en la Fig. 8a y 9 respectivamente, son comúnmente utilizados en aplicaciones de banda ancha. En ambos casos el diodo se encuentra en un estado de "paso de potencia". Lo que indica este estado es que conforme se incrementa la polarización

directa aplicada se presenta un pequeño aumento en la resistencia serie del diodo "Rs", ubicada entre la "RF" del generador y la carga. En cambio, bajo la condición de "interrupción de potencia" el diodo se encuentra en polarización inversa, por lo que presenta un estado de alta impedancia entre la fuente y la carga [6].

En los conmutadores conectados en serie, el máximo aislamiento que se puede obtener depende fundamentalmente de la capacitancia del diodo *p-i-n*, mientras que la pérdida de inserción y la disipación de potencia se encuentran en función de la resistencia serie del diodo [6].

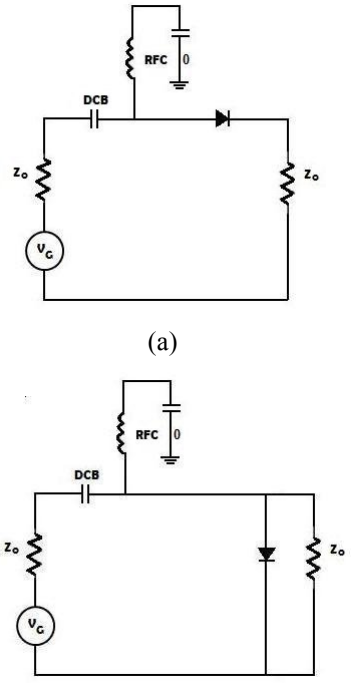


Fig. 8. Conmutadores SPST en base a diodo *p-i-n* dispuesto en (a) serie y (b) paralelo

Los principales parámetros de operación de los conmutadores conectados en serie es la pérdida de inserción, situación de baja impedancia del diodo por polarización directa, y el aislamiento, situación de alta impedancia del diodo por polarización inversa.

Para encontrar la pérdida de inserción en el conmutador conectado serie:

$$IL_{dB} = 20 \log_{10}\left[1 + \frac{R_S}{2Z_0}\right] \qquad (15)$$

Esta ecuación aplica para un conmutador SPST. Para los conmutadores multi–tiro las pérdidas por inserción son un poco mayores, debido a algún desacoplo con los conectores del diodo provocado por la capacitancia de los diodos *p-i-n*.

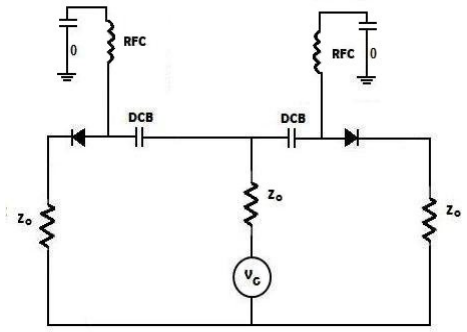


Fig. 9. Conmutador SPDT [6]

La ecuación para el cálculo del aislamiento para conmutadores SPST en base a diodos *p-i-n* es la siguiente:

$$I_{dB} = 10 \log_{10}\left[1 + \left(4\pi f C Z_0\right)^{-2}\right] \qquad (16)$$

Como se puede apreciar, un valor bajo de capacitancia permitirá obtener un aislamiento mayor. Esta es una característica requerida por las antenas de arreglos de fase en sistemas de comunicaciones.

Por cada conmutador SPNT se agregan 6 dB con relación al 50% del voltaje de reducción a través del diodo en estado de alta impedancia, debido a las limitaciones del generador en cuanto a sus características de impedancia [6].

2) Conmutadores conectados en derivación o paralelo

En la Fig. 8b y 10 se muestran los dos conmutadores típicos con diodos *p-i-n* conectados en derivación. Estos conmutadores ofrecen elevados aislamientos para muchas aplicaciones, debido a que el diodo permite disminuir el calor en un electrodo y así es capaz de manipular mayor potencia RF que un conmutador con diodo *p-i-n* tipo serie.

En los diseños del conmutador con derivación tanto el aislamiento como la disipación de potencia, se encuentran en función de los incrementos en la resistencia de los diodos, donde las pérdidas por inserción dependen principalmente de la capacitancia de los diodos *p-i-n*.

Los parámetros de operación de los conmutadores con derivación son descritos principalmente por las siguientes ecuaciones.

La expresión para el cálculo de pérdida de inserción para ambos conmutadores con derivación SPST y SPNT es la siguiente:

$$IL_{dB} = 10 \log_{10}\left[1 + \left(\pi f C_T Z_0\right)^2\right] \qquad (17)$$

La ecuación para el cálculo del aislamiento para un interruptor con derivación SPST es:

$$I_{dB} = 20 \log_{10}\left[1 + \frac{Z_0}{2R_S}\right] \qquad (18)$$

Para obtener un correcto aislamiento para un interruptor muti-tiro se agregan 6 dB a los valores obtenidos con (13) [6].

3) Conmutadores compuestos y sintonizados

Estos conmutadores presentan grandes ventajas en cuestión de un mejor desempeño en el efecto de aislamiento en comparación con utilizar un solo diodo *p-i-n* como un conmutador. Los conmutadores compuestos son combinaciones de arreglos en serie y en paralelo de los diodos *p-i-n*. Los conmutadores sintonizados son conmutadores con una estructura resonante. Con estas configuraciones se puede tener más de 40 dB de aislamiento.

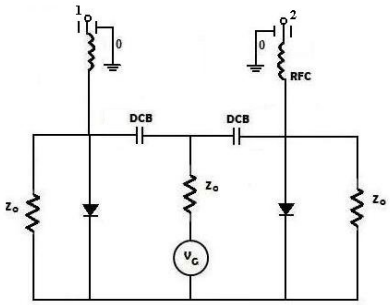


Fig. 10. Conmutadores SPDT con diodos en derivación [6].

En la siguiente figura se muestran dos conmutadores compuestos SPST.

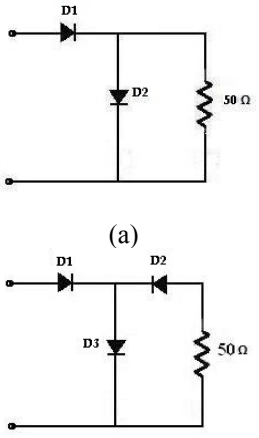


Fig. 11. Conmutadores SPST compuestos (a) serie-paralelo-ELL y (b) *Tee* [6].

En el conmutador compuesto, cuando los diodos *p-i-n* en serie se encuentran bajo polarización directa y los diodos en paralelo se encuentran en polarización inversa o estado cero, se trata del estado de pérdida de inserción. El caso inverso es el estado de aislamiento.

En estos casos, los circuitos de control de los diodos son más complejos en comparación con los conmutadores simples.

En la Fig. 12 se puede observar el esquema de un conmutador serie-paralelo.

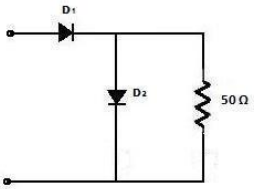


Fig. 12. Conmutador serie-paralelo [6].

Si se conectan dos conmutadores en serie o dos en paralelo separados una distancia de una longitud de onda, podemos obtener un conmutador sintonizado. Así se duplica el valor obtenido de aislamiento con un solo diodo. Sin embargo, la pérdida por inserción es mayor al obtenido por un conmutador de un solo diodo.

En el caso del conmutador sintonizado con configuración en paralelo las perdidas en comparación con un conmutador simple en paralelo serian menores debido a la resonancia que existe en este circuito.

Las configuraciones de estos conmutadores sintonizados se muestran en la Fig. 13.

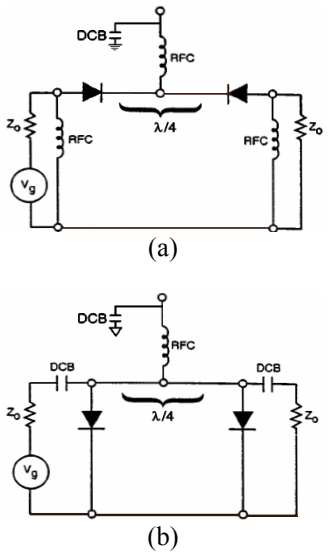


Fig. 13. Conmutador sintonizado (a) serie y (b) paralelo [6].

C. *Análisis de Pérdidas en Conmutadores de Microondas*

En la Tabla III se muestran las formulas para calcular las pérdidas de inserción en dB de los conmutadores simples y compuestos usando diodos *p-i-n*.

TABLA III
EXPRESIONES PARA EL CÁLCULO DE PÉRDIDA DE INSERCIÓN DE CONMUTADORES SPST [6].

Tipo	Pérdida de Inserción, dB
Serie	$20 \log_{10}\left[1+\frac{R_S}{2Z_0}\right]$
Derivación	$10 \log_{10}\left[1+\left(\frac{Z_0}{2X_C}\right)^2\right]$

Tipo	Pérdida de Inserción, dB
Serie – Derivación	$10 \log_{10}\left[\left(1+\frac{R_S}{2Z_0}\right)^2+\left(\frac{Z_0+R_S}{2X_C}\right)^2\right]$
Tee	$20 \log_{10}\left[1+\frac{R_S}{Z_0}\right]+10 \log_{10}\left[1+\left(\frac{Z_0+R_S}{2X_C}\right)^2\right]$

1) Pérdida de inserción de acuerdo a la frecuencia de operación

Para el análisis de pérdida de inserción respecto a la frecuencia de operación, consideramos la resistencia serie de 1.5 Ohms, en situación de polarización directa, y capacitancia de unión de 0.0045 pF, en estado de polarización inversa, correspondientes al diodo *p-i-n* HPND4038.

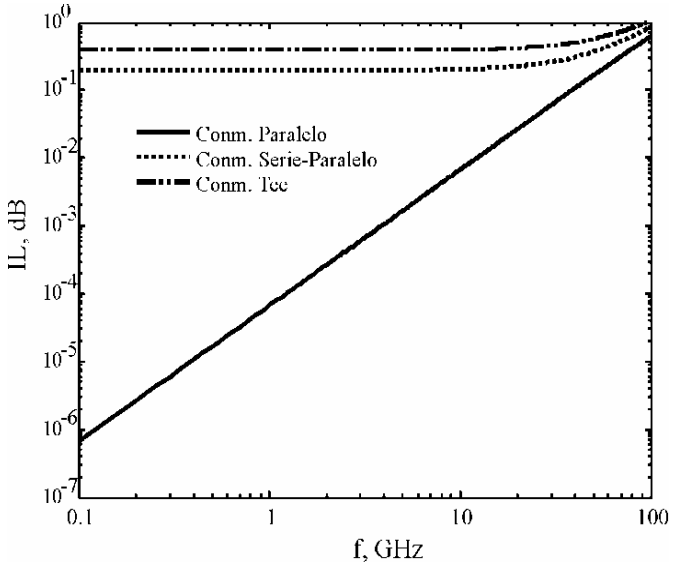


Fig. 14. Pérdida de inserción respecto a la frecuencia de operación.

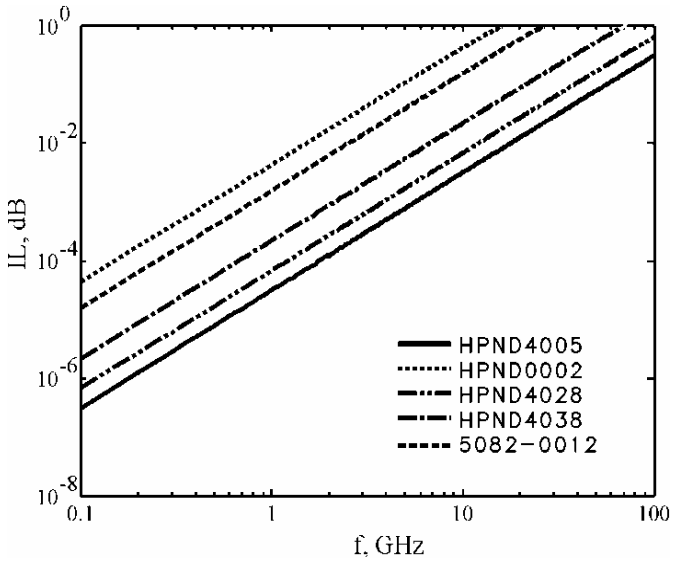


Fig. 15. Pérdidas de inserción de conmutadores en paralelo.

En la Fig. 14 se muestra la respuesta de pérdida de inserción de conmutadores serie-paralelo, paralelo y *Tee*, utilizando el diodo HPND4038.

Como se puede apreciar en la Fig. 14 los conmutadores tipos paralelo, serie-paralelo y *Tee*, tienen una pérdida de inserción, hasta los 10 GHz, por debajo de 0.0067 dB, 0.2045 dB y 0.398 dB, respectivamente.

La menor pérdida de inserción para frecuencias hasta los 10 GHz se obtiene utilizando el Conmutador Paralelo. Conforme aumenta la frecuencia de operación por arriba de 10 GHz, la pérdida de inserción en los tres conmutadores incrementa exponencialmente como se aprecia en la Fig. 14.

Se continuó con el análisis de las respuestas de pérdida de inserción de los conmutadores serie, paralelo, serie-paralelo y *Tee*, para dispositivos en base a los diodos presentados en la Tabla I.

La Tabla IV muestra la pérdida de inserción obtenida de los conmutadores en serie.

TABLA IV
VALORES DE PÉRDIDA DE INSERCIÓN EN LOS CONMUTADORES EN SERIE PARA CADA UNO DE LOS DIODOS UTILIZADOS.

Diodo	IL(dB)
HPND4005	0.40
HPND0002	0.30
HPND4028	0.19
HPND4038	0.13
5082-0012	0.08

Las Fig. 15, 16 y 17 muestran la pérdida de inserción en dB para los conmutadores en paralelo, serie-paralelo y *Tee* respectivamente, en base a los diodos *p-i-n* de la Tabla I.

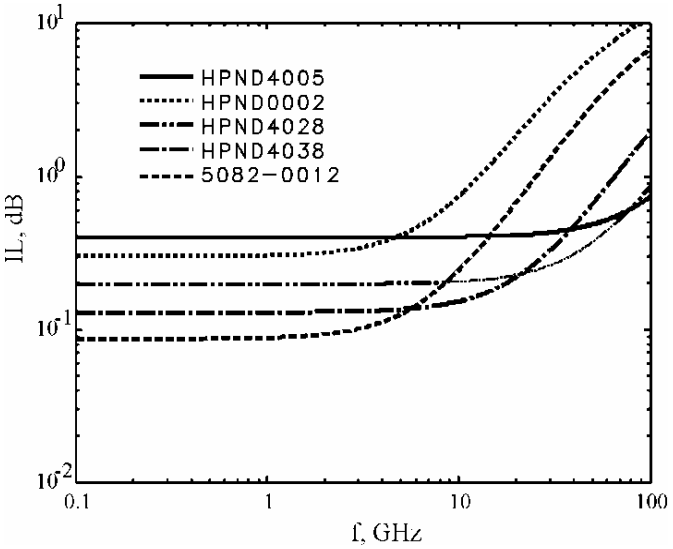


Fig. 16. Pérdida de inserción de conmutadores serie-paralelo.

Se puede observar que la pérdida de inserción en los conmutadores en paralelo es proporcional al valor de

capacitancia de unión del diodo *p-i-n* utilizado, y tiene una respuesta exponencial creciente a partir de, aproximadamente 2 GHz, como se puede ver en la Fig. 15.

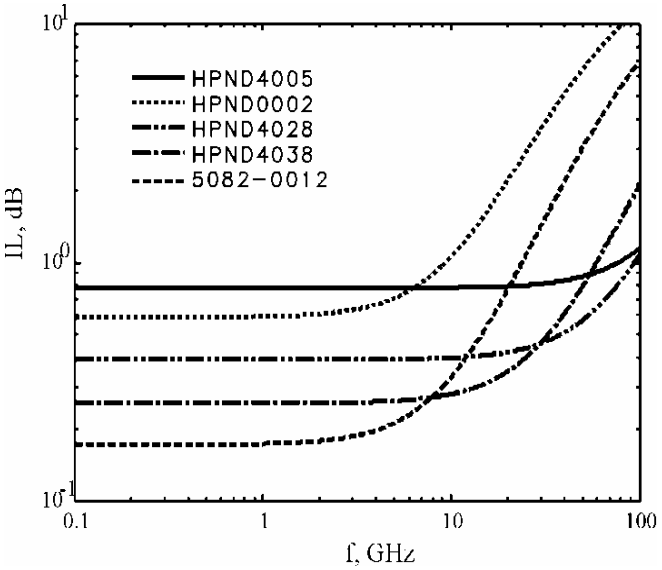


Fig. 17. Pérdida de inserción de conmutadores *Tee*.

La pérdida de inserción en los conmutadores *Tee* y serie-paralelo es directamente proporcional al valor de resistencia serie del diodo *p-i-n* a utilizarse, en bajas frecuencias, es decir menor a 1 GHz. A partir de esta frecuencia, el valor de pérdida de inserción aumenta exponencialmente, presentándose una dependencia más directa con la capacitancia de unión de los diodos utilizados.

IV. Conclusiones

Se describen a detalle los parámetros de desempeño y las características de funcionamiento de los diodos *p-i-n*.

Así mismo, se expone la teoría necesaria y suficiente de dispositivos conmutadores ampliamente utilizados en sistemas de comunicaciones por microondas.

Se presenta un análisis numérico de los parámetros de los diodos *p-i-n*, y las pérdidas de inserción de conmutadores serie, paralelo, serie-paralelo y *Tee* construidos con diodos *p-i-n*. El análisis se realiza en la frecuencia de operación de la mayoría de los dispositivos de microondas.

Agradecimientos

La investigación se llevó a cabo con apoyo parcial del proyecto de investigación SIP20090676 del Instituto Politécnico Nacional.

Referencias

[1] K. Chang, "Handbook of Microwave and Optical Components," Vol. 2 Microwave Solid-State Components, John Wiley & Sons, 1990.

[2] J. F. White, "Microwave Semiconductor Engineering," Van Nostrand Reinhold Co.1982.

[3] G. Mike, "The RF and Microwave Handbook," Mike Golio (Ed.), Boca Raton: CRC Press LLC, 2001, pp. 664- 671.

[4] R. H. Caverly, and G. Hiller, "The Frequency-Dependent Impedance of *P-I-N* Diodes", *IEEE Transactions on Microwave Theory and Techniques*, Vol. 37, No. 4, April 1989.

[5] R. H. Caverly and G. Hiller, "Microwave Resistance of Gallium Arsenide and Silicon P-I-N Diodes", *IEEE MTT-S Digest*, 1987.

[6] G. Hiller, "Design with PIN diodes," Alpha Industries.

Restricción del Uso de Teléfonos Celulares en Ambientes Controlados

Maria Aurora Molina Vilchis, Ramón Silva Ortigoza,
Yasania Joselín Escalona Bautista y Héctor Oscar Ramos García

***Resumen*—Es común que se provoquen interrupciones o interferencias por el uso indiscriminado de teléfonos celulares en eventos académicos, culturales o sociales, de ahí que surja la necesidad de evitar o disminuir la recepción o transmisión de llamadas. Otras restricciones pudieran estar relacionadas con el uso de las cámaras fotográficas que incorporan estos dispositivos, la transmisión de mensajes o grabaciones de videos sin autorización. En este artículo se presenta una aplicación basada en Bluetooth para el control del uso de estos dispositivos en ambientes con restricciones.**

***Palabras clave*—Bluetooth, restricción del uso de teléfonos celulares, protocolo de comunicación.**

Restriction of the Usage of Mobile Phones in Controlled Environments

***Abstract*—It often happens that interruptions or interferences occur due to indiscriminate usage of the mobile phones during academic, cultural o social events. Thus, there is a necessity for avoiding or diminishing transmissions of phone calls. Another important restriction is related with the unauthorized usage of cameras integrated in these devices, transmission of messages or video capture without permission. In this paper, we present a Bluetooth based application for the restriction of usage of the mobile phones in specially controlled environments.**

***Index Terms*—Bluetooth, usage restrictions for mobile phones, communication protocol.**

I. Introduction

La importancia que han cobrado las Redes de Área Personal, se basa en la popularidad que han alcanzado los dispositivos móviles tales como los teléfonos celulares, PDA y computadoras portátiles, ya que permiten la comunicación eficiente en cualquier momento y lugar en un entorno personal, de tal manera que se perciben estos dispositivos como partes de un sistema de comunicación integral a lo que se le ha dado en llamar computación pervasiva. Este concepto novedoso se le atribuye a Mark Weiser quien lo anticipó en sus escritos en 1988, cuando trabajaba para Xerox en el laboratorio de Palo Alto (PARC) en EUA.

La evolución que han tenido los teléfonos celulares ha logrado que este concepto sea ahora una realidad, al incorporar funcionalidades que no hace mucho parecían futuristas como la transmisión y recepción de archivos multimedia, ejecución de juegos, reproducción de archivos MP3, correo electrónico, Web, envío de mensajes de texto y fotografías, recepción de radio y televisión digitales, entre otros. Lo anterior los ha hecho ser los dispositivos móviles más populares por excelencia, sólo en México existen casi 65 millones de teléfonos celulares.

No obstante, el abuso indiscriminado de estos dispositivos ha creado serios problemas, sobre todo la recepción o transmisión de llamadas telefónicas en los ambientes con restricciones de comunicación, pues generan interrupciones no deseadas sobre todo en eventos culturales, ruedas de prensa, ámbitos académicos, etc. donde existen prohibiciones de uso.

En este artículo se presenta el desarrollo de una aplicación de una red Bluetooth para el control de comunicaciones telefónicas celulares en ambientes controlados. La sección II trata de los aspectos técnicos de la red Bluetooth. En la sección III se presenta el diseño de la aplicación. En la sección IV se presenta su desarrollo y los resultados de las pruebas de operación.

II. Trabajo Previo

La tecnología Bluetooth se ha incorporado de manera natural a estos dispositivos facilitando su interacción y comunicación, y se erige como el estándar de facto de las redes de área personal. Sin embargo, el origen de esta tecnología está relacionado estrechamente con las investigaciones en el campo de las comunicaciones inalámbricas [1]. Así, desde su origen en la década de los 70 hasta la actualidad el interés por estas redes se ha visto progresivamente incrementado convirtiéndola en la tecnología más popular de los últimos años. Los recientes avances en la materia se han centrado en las redes *ad hoc*. Este término hace referencia a redes flexibles inalámbricas sin infraestructura, donde los dispositivos se conectan de forma autónoma.

Estas redes deben poder adaptarse dinámicamente ante cambios continuos, como la posición de los dispositivos, la potencia de la señal, el tráfico de la red y la distribución de la carga. Su principal reto estriba en los continuos e impredecibles cambios de topología.

Manuscrito recibido el 18 de febrero del 2008. Manuscrito aceptado para su publicación el 20 de agosto del 2009.

M. A. Molina Vilchis es del Área de Telemática del Departamento de Posgrado, CIDETEC, Instituto Politécnico Nacional, México (mamolinav@ipn.mx).

R. Silva Ortigoza es del Area de Mecatrónica del Departamento de Posgrado, CIDETEC, Instituto Politécnico Nacional, México (rsilva@ipn.mx).

Y. J. Escalona Bautista y H. O. Ramos García son del Departamento de Comunicaciones y Electrónica, ESIME-Z, Instituto Politécnico Nacional, México (jossy_259_1@hotmail.com, kzper16@hotmail.com).

Recientemente la tecnología Bluetooth se ha mostrado como una plataforma de soporte prometedora para estas redes, su principal ventaja es su habilidad para localizar de forma transparente dispositivos móviles cercanos, así como los servicios que ofrecen.

Bluetooth ha es el estándar impulsado por Ericsson, IBM, Intel, Nokia y Toshiba. Las que formaron el Grupo de Interés Especial (*Special Interest Group*, SIG) Bluetooth en Mayo de 1998 [2]. Definida dentro del estándar IEEE 802.15 que se refiere a las redes inalámbricas tipo personal. La conexión inalámbrica Bluetooth opera en el rango de radiofrecuencia de los 2,4 GHz (2,400 a 2,485 GHz) [3] que no requiere licencia de uso en ningún lugar del mundo. Con una banda de guarda de 2 MHz a 3,5 MHz para cumplir con las regulaciones internacionales. Transmite en espectro disperso, con salto de frecuencia, en dúplex y hasta 1,600 saltos/s. La señal salta entre 79 frecuencias en intervalos de 1 MHz para tener un alto grado de tolerancia a las interferencias y obtener comunicaciones robustas. Además se dispone de comunicaciones punto a punto y multipunto, donde un dispositivo puede establecer de forma simultánea hasta siete canales de comunicación a la vez con un solo radio de cobertura.

Bluetooth transmite a una tasa de 1 Mbps en su funcionamiento básico, y de 2/3 Mbps en modo mejorado (Bluetooth 2.0). Utiliza modulación Gausiana por Desplazamiento de Frecuencia (*Gaussian Frecuency Shift Keying*, GFSK) para el modo básico, y en el modo mejorado la Modulación Diferencial por Desplazamiento de Fase en Cuadratura (*Differential Quadrature Phase-Shift Keying*, DQPSK) y la Modulación Diferencial por Desplazamiento de Fase (*Differential Phase-Shift Keying*, 8-DPSK). Para lograr que la comunicación sea en dúplex se divide el tiempo de transmisión en ranuras por medio de la técnica Dúplex por División de Tiempo (*Time-Division Duplex*, TDD). La cobertura es de 100 metros para los dispositivos de clase 1 (100 mW), de 20 metros para los de clase 2 (10 mW) y de 10 metros para los de clase 3 (1 mW). Siendo los dispositivos de clase 3 los mayormente usados.

Bluetooth permite conectarse casi con cualquier dispositivo compatible que se halle en las proximidades, cuando los dispositivos se conectan se forma una piconet, en esta los dispositivos conectados comparten el mismo canal y adquieren dos roles distintos: maestro o esclavo. En cada piconet solamente puede existir un maestro y un máximo de siete esclavos, estos últimos no pueden establecer enlaces entre sí, por lo que todo el tráfico es enviado al maestro. Para que el dispositivo maestro logre establecer la comunicación en un principio, deberá ejecutar el protocolo de descubrimiento, este envía una señal en multidifusión para solicitar, básicamente a los dispositivos cercanos, su dirección MAC y el PIN. De esta manera es posible registrar a los dispositivos que se encuentran en estado de activo, sin importar el servicio que estén realizando.

Diversas aplicaciones, basada en Bluetooth, se han propuesto como soluciones a diferentes problemas del quehacer humano. Así en [4] se presenta el desarrollo de un sistema de información contextual para terminales móviles que sirve como mecanismo de localización en ambientes tales como museos, monumentos, etc. Como apoyo a la red de transporte público en Italia, para llevar a cabo tareas de diagnóstico inalámbrico y mantenimiento preventivo en los autobuses [5]. Otra aplicación interesante es una alarma, como botón de pánico, que envía un mensaje de texto desde un teléfono móvil a un número de emergencia, en caso de que el portador del móvil sufra algún percance en la vía pública [6].

Por otro lado, se han desarrollado aplicaciones en el campo de la medicina, en el Centro Noruego para la Telemedicina, se ha desarrollado una aplicación que hace posible vigilar el nivel de glucosa en la sangre de manera remota. Esta aplicación está dirigida al auto cuidado de la diabetes en pacientes menores de edad [7]. Otras aplicaciones se centran en el mantenimiento y operación de estaciones de bombeo de agua en áreas urbanas para el control del agua [8].

No obstante, se han desarrollado muy pocas aplicaciones para el monitoreo, identificación y rastreo de dispositivos móviles y sus servicios, ese es el caso de *BlueSweep* de AirMagnet Inc que facilita a los usuarios de dispositivos móviles localizar sus dispositivos e identificar los servicios en los que se encuentran trabajando en tiempo real, con lo que se puede identificar las posibles amenazas de seguridad y llevar un registro de incidentes [20].

III. Diseño de la Aplicación

A. Antecedentes del Problema

Entre las soluciones propuestas para resolver el problema del control de teléfonos celulares en ambientes con restricciones de uso, podemos mencionar a los sistemas de bloqueo de llamadas que obstruyen la recepción en un radio de 30 metros aproximadamente, y que son utilizados principalmente para los sistemas penitenciarios. Entre los equipos que se ofertan podemos citar el *Radio Capsule* SRC-300 [9] y el *Portable Palm Phone Jammer* [10] que cuentan con un alcance aproximado de 10 a 50 metros, los cuales pueden trabajar en frecuencias de 800 a 900 MHz y de 1930 a 1990 MHz, con un consumo de 5 voltios y son de dimensiones muy pequeñas.

Si partimos del hecho de que muchos de los teléfonos celulares incorporan la tecnología Bluetooth, entonces es posible plantear una solución para el problema que nos ocupa, por medio de una aplicación que permita detectar o identificar distintos marcas de teléfonos celulares activos y sus servicios, de una forma rápida y automática en un radio de hasta 100 metros utilizando su propio protocolo de descubrimiento de dispositivos.

Respetando el derecho que todos los seres humanos tenemos para comunicarnos, esta aplicación no deberá contraponerse a este derecho, por lo que se establece como requisito en su diseño las siguientes características:

Que los teléfonos celulares cuenten con la tecnología Bluetooth para que puedan ser detectados mediante su protocolo de descubrimiento.

Una vez detectados se va a transmitir una aplicación que al ser aceptada por el usuario desplegará un menú de texto en donde se podrán ver las opciones a seguir.

La consola de control inicia la comunicación enviando mensajes de texto a los teléfonos celulares que estén dentro de su alcance, para obtener su identificación, es decir su dirección física (MAC).

B. Contexto de Operación

El sistema propuesto utilizará una consola de control para ser instalada en la cabina del auditorio, para propósitos de prueba. Esta consola enviará las señales de descubrimiento de los teléfonos celulares con un alcance de hasta 40 metros [11], si el auditorio sobrepasa la distancia estimada, se usarán equipos repetidores para establecer varias conexiones Bluetooth simultáneas con un alcance de hasta 200 metros.

C. Interfaz de Usuario

Los requisitos para el diseño de la interfaz de la aplicación del lado de la consola son: que cuente con la tecnología Bluetooth, con una tarjeta inalámbrica y que el sistema operativo sea compatible con el software utilizado.

La interfaz de los teléfonos celulares deberá ser diseñada de tal manera que sea compatible con las características de despliegue y operación de la mayoría de las marcas de teléfonos, es decir en modo de texto y con selección por menú, utilizando los botones de desplazamiento convencionales. Ver Fig. 1.

Fig. 1. Interfaz del Usuario.

D. Diseño de los objetos del sistema

El diseño de aplicaciones para teléfonos celulares es en base a la programación orientada a objetos, de esta manera la aplicación se basará en dos clases: *Listener* y *Auditorio*. La primera es esencial para el funcionamiento de la tecnología Bluetooth y la segunda es la aplicación en sí.

La función de *Listener*, propia de la tecnología Bluetooth, es detectar de forma automática los teléfonos celulares activos, confirmar su presencia y mostrarlos en la pantalla, además de reservar los servicios de operación del dispositivo maestro, en este caso de la consola de control. Utiliza el atributo Vector *Disp_Encontrados* para poder realizar la comunicación automáticamente. También se emplearán funciones relacionadas con los servicios como son *deviceDiscovered ()* que realiza la función de detectar los teléfonos celulares con tecnología Bluetooth, *serviceSearchCompleted()* que recibe una afirmación a la búsqueda de servicios y *servicesDiscovered()* para confirmar la localización de servicios y mostrarla en pantalla [12].

La clase llamada *Auditorio* es usada para ejecutar la aplicación, en esta se incluyen los atributos y servicios que tienen que ver con la operación y el despliegue de mensajes presentados en la interfaz del usuario del teléfono celular.

E. Funcionamiento esperado

En la Fig. 2., se muestra el protocolo para la instalación de la aplicación. La computadora de control inicia la comunicación enviando un mensaje en multidifusión (*Req_Inquiry*) a los teléfonos celulares para su descubrimiento y con ello identifica su nombre y dirección física.

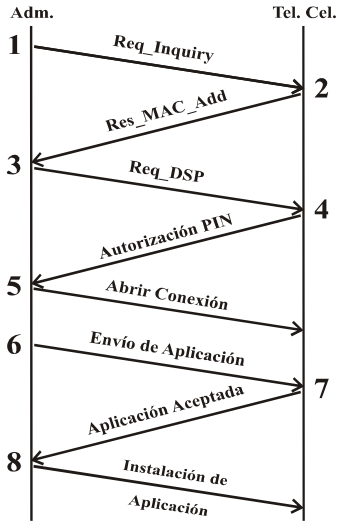


Fig. 2. Protocolo para la instalación de la aplicación.

En consecuencia, los teléfonos celulares envían el mensaje *Res_MAC_Add* como respuesta al requerimiento, enviando su dirección física y nombre. A continuación la consola de control enviará el mensaje Req_DSP para decidir con que dispositivo establecerá la conexión y el servicio que usará. El teléfono celular enviará la Autorización del PIN, para que el dispositivo responda con los diferentes servicios que puede ofrecer. Una vez recibido el PIN se abre la conexión con el mensaje *Abrir Conexión*. Esto permitirá enviar la aplicación al dispositivo seleccionado. Finalmente sólo es necesario que el usuario instale la aplicación y haga uso de ella.

IV. DESARROLLO DE LA APLICACIÓN

A. Plataforma de desarrollo

El desarrollo de aplicaciones para teléfonos celulares se basa en la plataforma Java [13], con los programas Java para móviles: *Java 2 Micro Edition* (J2ME) que utiliza un subconjunto reducido de componentes Java SE, máquinas virtuales y la interfaz de programación de aplicaciones (*Application Programming Interface*, API) [14], Sun

Microsystems, Java JDK (*Java Development Kit*) para aplicaciones de teléfonos celulares [15] y *NetBeans* [16].

B. Funciones

Para el desarrollo de la aplicación se definieron tres funciones: 1) Constructor de *Auditorio*, 2) Inicialización de la aplicación, y 3) Ejecución de la acción asignada a *Mobile Device*; mostradas en la Fig. 3.

```
 * The Auditorio constructor.
 */
public Auditorio() { //Constructor
}
//<editor-fold defaultstate="collapsed"desc="Generated Method: initialize">
/**
 * Initilizes the application.
*/
private void initialize() { //Comenzar a Ejecutar el Programa
    // write pre-initialize user code here
}
//</editor-fold>

//<editor-fold defaultstate="collapsed"desc="Generated Method: startMIDlet">
/**
 * Performs an action assigned to the Mobile Device - MIDlet Started point.
 */
public void startMIDlet() { //Ejecución de Aplicación
    // write pre-action user code here
    switchDisplayable(getAlertaBienvenida(), getMenu());
          // Muestra la alerta Bienvenido y después el Menú en Pantalla
}
//</editor-fold>
```

Fig. 3. Funciones de Inicialización.

Adicionalmente se definieron otras funciones: *switchDisplayable* para el despliegue de mensajes en la pantalla del celular, ver Fig. 4.; *commandAction* que es la que especifica la aplicación que se enviará al dispositivo, ver Fig. 6.; *exitMIDlet* mostrada en la Fig. 5, que realiza la salida definitiva de la aplicación.

```
public void switchDisplayable(Alert alert, Displayable nextDisplayable) {
    // write pre-switch user code here
    Display display = getDisplay();//Se asigna la variable de tipo Display
    if (alert == null) {//Se compara si Alert es de tipo nulo
        display.setCurrent(nextDisplayable); // Cambia de pantalla
    }else {// Si no se cumple la condición vuelve a cambiar de pantalla
        display.setCurrent(alert, nextDisplayable);
    }
    // write post-switch user code here
}
//</editor-fold>
```

Fig. 4. Función *switchDisplayable().*

```
/**
 * Exits MIDlet.
 */
 public void exitMIDlet() { //Sale del MIDlet
 switchDisplayable (null, null); //Coloca en nulo el switch
 destroyApp(true); //Recibe verdadero
 notifyDestroyed(); //Notifica la destrucción de la aplicación
}
/**
```

Fig. 5. Función *exitMIDlet().*

```
public void commandAction(Command command,Displayable displayable) {
   if (displayable == Alerta) {
        if (command == ComandoRegresaAlerta) {
            switchDisplayable(null, getMenu());
        }else if (command == ComandoSalirAlerta) {
            exitMIDlet();
        }
    }else if (displayable == Menu) {
        if (command == ComandoContinuar) {
            int Seleccion;
            Seleccion=GrupoElementos.getSelectedIndex();

            //La variable Seleccion contiene el índice seleccionado del grupo de
            elementos switch(Seleccion)
            {
                //Selecciona la opción alerta.
                case 0:
                {
                    switchDisplayable(null, getAlerta());
                    // Función de NetBeans que permite vibrar al seleccionar la opción
                    javax.microedition.lcdui.Display.getDisplay(this).vibrate(1000);
                    break;
                }
                //Selecciona la opción Recomendación.
                case 1:
                {
                    switchDisplayable(null, getRecomendacion());
                    // Función de NetBeans que permite vibrar al seleccionar la opción
                    javax.microedition.lcdui.Display.getDisplay(this).vibrate(1000);
                    break;
                }
                //Selecciona la opción Precaución.
                case 2:
                {
                    switchDisplayable(null, getPrecaución());
                    // Función de NetBeans que permite vibrar al seleccionar la opción
                    javax.microedition.lcdui.Display.getDisplay(this).vibrate(1000);
                    break;
                }
                //Selecciona la opción Salir.
                case 3:
                {
                    exitMIDlet();
                }
            }
        }else if (command == ComandoSalir) {
            exitMIDlet();
        }
    }else if (displayable == Precaución) {
        if (command == ComandoRegresaPrecaucion) {
            switchDisplayable(null, getMenu());
        }else if (command == ComandoSalePrecaucion) {
            exitMIDlet();
        }
    }else if (displayable == Recomendacion) {
        if (command == ComandoRegresaRecomendacion) {
            switchDisplayable(null, getMenu());
        }else if (command == ComandoSalirRecomendacion) {
            exitMIDlet();
        }
    }
}
//</editor-fold>
```

Fig. 6. Función *commandAction().*

En la Fig. 7. se especifican las Funciones para *Pausar* y *Salir* del MIDlet, con tres opciones: 1) para el reinicio de la aplicación, después de que ha sido pausada cuando alguna acción del teléfono tiene una mayor prioridad, como por ejemplo la entrada de una llamada telefónica o un mensaje de texto. 2) Pausa del MIDlet, da un valor verdadero para que siempre que suceda alguna de las acciones antes mencionadas automáticamente realice le pausa y no exista conflicto con la acción prioritaria. 3) Destrucción de la aplicación. Es decir, borra la aplicación en el dispositivo.

```
public void startApp() { // Reinicia la aplicación pausada
   if (midletPaused) { resumeMIDlet ();}
     else {
           initialize ();
           startMIDlet ();
          }
       midletPaused = false;
 }
 public void pauseApp() { //Pausa la aplicación cuando es necesario
       midletPaused = true;
 }
 /**
  * Hace el llamado para poner fin al MIDlet.
  */
 public void destroyApp(boolean unconditional) {
 }
}
```

Fig. 7. Funciones para Pausar y Salir del MIDlet.

C. *Pruebas*

Para efectos de prueba, se utilizó un auditorio común, durante la realización de un evento cultural, de 213 m^2 de área total. Se propusieron y se realizaron tres pruebas: 1) descubrimiento, 2) transmisión y 3) compatibilidad.

1) *Prueba de Descubrimiento*

Las primeras pruebas de alcance y descubrimiento tienen como finalidad determinar los teléfonos celulares reconocidos por la consola de control. Para iniciar la búsqueda, la consola de control envía una señal en multidifusión para localizar los teléfonos por su dirección física y su nombre. Ver Fig. 8.

2) *Pruebas de Transmisión*

A continuación se hace la transferencia de la aplicación en modo punto a punto. Sin embargo, hay que considerar que la transferencia puede ser interrumpida por diferentes causas: el rechazo en la transferencia por parte del usuario, pérdida en la conexión debido a ese rechazo, distancia mayor al alcance de la señal o interferencias con otras señales de radio. Para esos casos, la aplicación mostró en pantalla un mensaje de error.

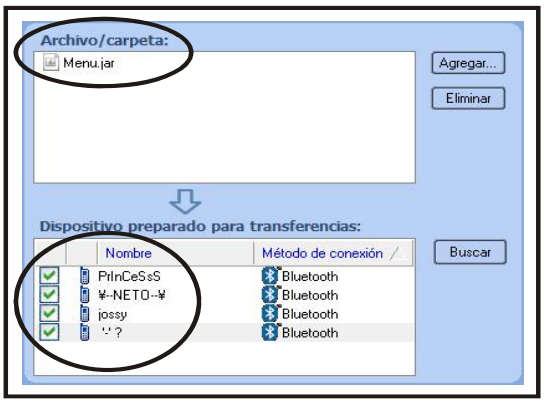


Fig. 8. Teléfonos Celulares Seleccionados para la Transmisión de la Aplicación.

3) *Pruebas de Compatibilidad*

Aquí se verificó si la aplicación, es compatible con los diferentes fabricantes de teléfonos celulares. En este caso se probaron las marcas Sony Ericsson, Nokia, LG, Samsung y Motorola. Esta prueba se realizó porque, cada fabricante tiene instalados diferentes sistemas operativos, plataformas y versiones de Bluetooth.

V. RESULTADOS

Para la obtención de resultados de las pruebas se empleó el programa SysNucles USBTrace [17]. Los resultados obtenidos de la primera prueba se muestran en la Fig. 9., donde se puede observar del lado izquierdo de la imagen los dispositivos detectados. Mientras a la derecha se muestran los detalles del registro, entre ellos el más importante para nuestros propósitos es el estado de la conexión [18].

Los resultados de la segunda prueba se muestran en la Fig. 10., en esta se observa el éxito o interrupción de la transferencia. Los teléfonos celulares que no pudieron recibir la aplicación fueron los modelos SGH-J700 de Samsung, MG300 y MG800c de LG.

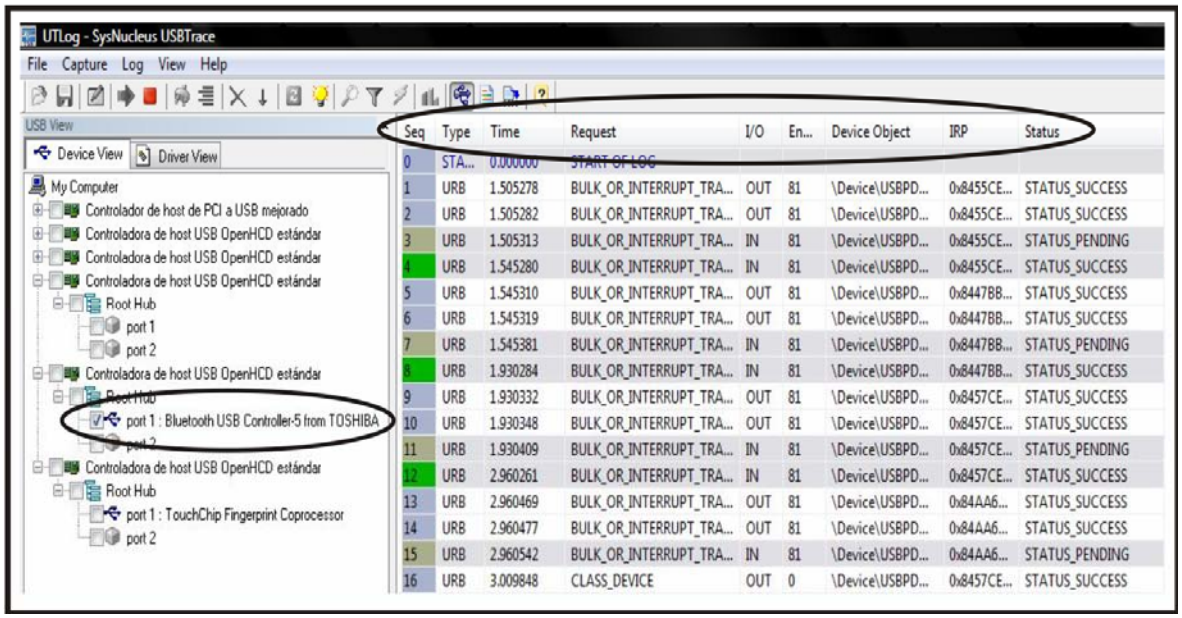
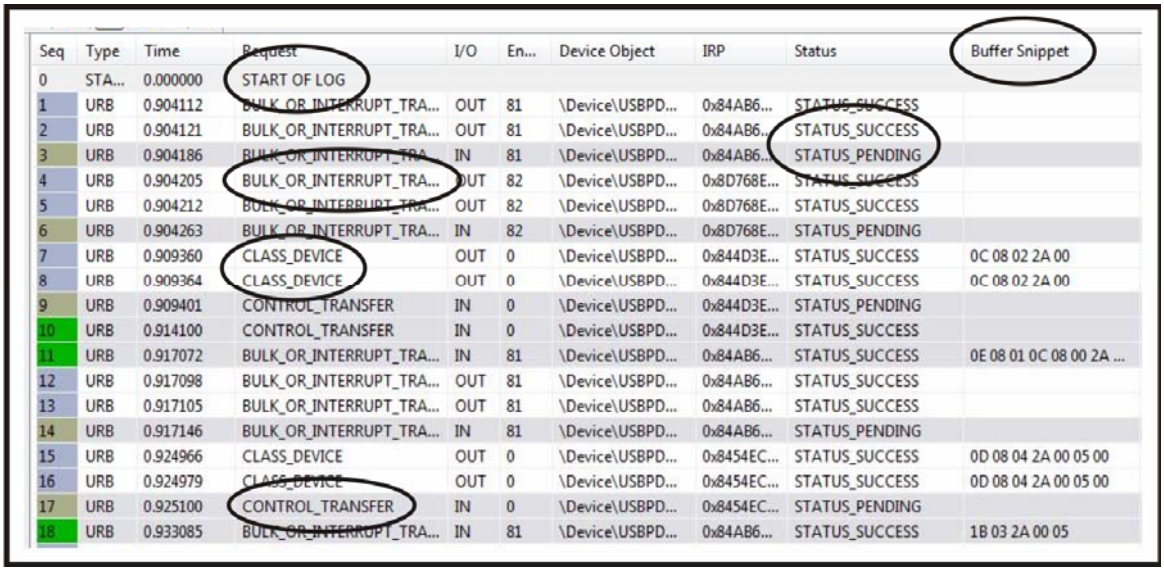


Fig. 9. Resultados de la prueba de detección.



Seq	Type	Time	Request	I/O	En...	Device Object	IRP	Status	Buffer Snippet
0	STA...	0.000000	START OF LOG						
1	URB	0.904112	BULK_OR_INTERRUPT_TRA...	OUT	81	\Device\USBPD...	0x84AB6...	STATUS_SUCCESS	
2	URB	0.904121	BULK_OR_INTERRUPT_TRA...	OUT	81	\Device\USBPD...	0x84AB6...	STATUS_SUCCESS	
3	URB	0.904186	BULK_OR_INTERRUPT_TRA...	IN	81	\Device\USBPD...	0x84AB6...	STATUS_PENDING	
4	URB	0.904205	BULK_OR_INTERRUPT_TRA...	OUT	82	\Device\USBPD...	0x8D768E...	STATUS_SUCCESS	
5	URB	0.904212	BULK_OR_INTERRUPT_TRA...	OUT	82	\Device\USBPD...	0x8D768E...	STATUS_SUCCESS	
6	URB	0.904263	BULK_OR_INTERRUPT_TRA...	IN	82	\Device\USBPD...	0x8D768E...	STATUS_PENDING	
7	URB	0.909360	CLASS_DEVICE	OUT	0	\Device\USBPD...	0x844D3E...	STATUS_SUCCESS	0C 08 02 2A 00
8	URB	0.909364	CLASS_DEVICE	OUT	0	\Device\USBPD...	0x844D3E...	STATUS_SUCCESS	0C 08 02 2A 00
9	URB	0.909401	CONTROL_TRANSFER	IN	0	\Device\USBPD...	0x844D3E...	STATUS_PENDING	
10	URB	0.914100	CONTROL_TRANSFER	IN	0	\Device\USBPD...	0x844D3E...	STATUS_SUCCESS	
11	URB	0.917072	BULK_OR_INTERRUPT_TRA...	IN	81	\Device\USBPD...	0x84AB6...	STATUS_SUCCESS	0E 08 01 0C 08 00 2A ...
12	URB	0.917098	BULK_OR_INTERRUPT_TRA...	OUT	81	\Device\USBPD...	0x84AB6...	STATUS_SUCCESS	
13	URB	0.917105	BULK_OR_INTERRUPT_TRA...	OUT	81	\Device\USBPD...	0x84AB6...	STATUS_SUCCESS	
14	URB	0.917146	BULK_OR_INTERRUPT_TRA...	IN	81	\Device\USBPD...	0x84AB6...	STATUS_PENDING	
15	URB	0.924966	CLASS_DEVICE	OUT	0	\Device\USBPD...	0x8454EC...	STATUS_SUCCESS	0D 08 04 2A 00 05 00
16	URB	0.924979	CLASS_DEVICE	OUT	0	\Device\USBPD...	0x8454EC...	STATUS_SUCCESS	0D 08 04 2A 00 05 00
17	URB	0.925100	CONTROL_TRANSFER	IN	0	\Device\USBPD...	0x8454EC...	STATUS_PENDING	
18	URB	0.933085	BULK_OR_INTERRUPT_TRA...	IN	81	\Device\USBPD...	0x84AB6...	STATUS_SUCCESS	1B 03 2A 00 05

Fig. 10. Resultados de la transferencia de la aplicación.

VI. Conclusiones

Los teléfonos celulares han facilitado en gran medida nuestra vida diaria, pero también se han convertido en un problema dado el abuso que se ha hecho de estos, sobre todo en aquellos ambientes donde está restringido su uso, las soluciones en hardware inhiben la recepción y transmisión de llamadas y atentan contra el derecho a la comunicación. Una alternativa es una aplicación de software basada en la tecnología Bluetooth, para la detección de teléfonos celulares activos y sus servicios, es decir que no sólo detecte la actividad en la recepción y transmisión de llamadas, sino en otras facilidades que requieren de autorización, y que además pueda adecuarse a las restricciones impuestas en distintos ambientes. Siendo la compatibilidad con distintas marcas de teléfonos celulares, el costo y la ubicuidad esenciales en su diseño.

Agradecimientos

M. A. Molina agradece el apoyo económico recibido del programa COFAA del IPN. R. Silva agradece el soporte económico recibido por la SIP y del programa EDI del IPN, así como del Sistema Nacional de Investigadores (SNI), México.

Referencias

[1] J. C. Cano, C. T. Calafate, P. Malumbres, M. Pietro Manzoni, “Redes inalámbricas ad hoc como tecnología de soporte para la computación Uuicua,” Departamento de Sistemas Informáticos y Computacionales, Universidad Politécnica de Valencia, disponible en: http://www.grc.upv.es/calafate/download/main_novatica.pdf, marzo 2008.

[2] J. Muller Nathan, “Tecnología Bluetooth,” Interamericana de España, McGraw Hill, 2002. Consulta: Febrero 2008.

[3] M. P. Mata Ramírez, “*Tecnología Bluetooth*”, GestioPolis, disponible en: http://www.gestiopolis.com/canales8/ger/tecnología-bluetooth.htm, marzo 2008.

[4] P. Pece-Juan, C. Fernández, C. Escudero, “Bluesic: Sistema de información contextual para terminales móviles basado en tecnología Bluetooth,” Departamento de Electrónica y Sistemas, Universidad de A. Coruña, marzo, 2008.

[5] Centro Riserche FIAT Scpa de Orbassano y Digigrup Sri de Torino, “ATM Spa-Reference Application”, disponible en: http://spanish. Bluetooth.com/NR/rdonlyres/235EC2BC-644F-4806-85ED-DD1666FEF43B/0/atm.pdf.

[6] O-S. Kell, “Head of business development, Bluetooth panic button potential to save lives,” Securecom Technologies Ltd., disponible en: http: Spanish.bluetooth.com/NR/rdonlyres/EC33B6D5-A831-4D3B-886B-48EDEF4E60D8/0/securecom.pdf, abril, 2008.

[7] The Norwegian Center for Telemedicine (NST), NST reference application, disponible en: http://spanish.bluetooth.com/NR/rdonlyres/BA25A6D5-15F0-437F-84C1- AABA58818BFD/0/Ddiabetes.pdf, consulta: diciembre 2008.

[8] Mr. O. Roland, ABBAS, Power Technologies Division, Norway, Oslo Municipaly Reference Application, disponible en: http:// spanish.bluetooth.com/NR/rdonlyres/4BCB7259-F643-4A27-855C-56875A78CDF8/0/oslo.pdf, consulta: diciembre 2008

[9] G. A. González Palacio, G.R. Peláez Gómez, “Bloqueo señal de celulares,” *Informática Jurídica*, Universidad Autónoma Latinoamericana, Facultad de Derecho, Medellín, Octubre 2005, disponible en: http://bloqueocelular.blogspot.com, consulta: marzo2008.

[10] A. Alegría, “Bloquea la señal de celular con Portable Palm Phone Jammer,” Gadgets, disponible en: http://tecnoday.net/2007/09/29/bloquea-la-senal-de-celular-con-portable-palm-phone-jammer/, consulta: marzo 2008.

[11] Toshiba. Satellite serie A210-A215 “Manual de Usuario,” pp. 124-140.

[12] . P.D. Borches Juzgado, C. Campo Vázquez, “Java 2 Micro Edition Soporte Bluetooth,” Versión 1.0, Universidad Carlos III de Madrid, 20 de Marzo del 2004, consulta Octubre 2008.

[13] Página Oficial de Java, “Java para Windows – Internet Explorer,” disponible en: http://java.com/es/download/windows_ie.jsp?locale=es&host=java.com:80&bhcp=1, consulta: octubre 2008.

[14] “Java ME Applications Learning Trail,” Página Oficial de NetBeans http://www.netbeans.org/kb/trails/mobility.html., consulta: Octubre 2008.

[15] Página Oficial de Java, “Conozca más sobre la tecnología Java,” disponible en: http://java.com/es/about/, consulta: octubre 2008.

[16] “Página Oficial de NetBeans IDE 6.1,” Disponible: http://www.netbeans.org/index.html, consulta: octubre 2008.

[17] “USBTrace Tour: User Interface,” Página Oficial de SysNucles USBTrace, disponible en: http://www.sysnucleus.com/usbtrace_tour.html, consulta: octubre 2008.

[18] “*USBTrace Tour: Viewing captured data*”, Página Oficial de SysNucles USBTrace, disponible en: http://www.sysnucleus.com/usbtrace_tour3.html, consulta: octubre 2008.

Evaluation of E-Learning Readiness: A Study of Informational Behavior of University Students

Michael Brückner and Orasa Tetiwat

***Abstract*—In this study we investigated the behavior of university students from different universities and faculties of Thailand with regard to search, evaluate, use and share information. Our goal was to prepare the introduction of personal information management into the e-learning curriculum. We compare our results with data reported by others. Method: For gathering the data we used a questionnaire in Thai language, which was actually translated from the English original and sent to various universities in Thailand. Follow-up interviews with an adapted set of questions were carried out to generate qualitative data and a deeper insight into the knowledge and practices of the students. Analysis: Both quantitative and qualitative analyses were carried out on the data coming from 1,317 university students. Quantitative analysis employed the statistical package SPSS. Results: We have got a picture of the present informational behavior of Thai students. The results showed some differences between Thai and foreign students, for example in the use of Internet search engines. The insights gained by this study will be applied in the generation of the part of the e-learning curriculum that deals with the students' personal information management and can be applied to informational behavior of students in other countries like Mexico, Brazil, etc.**

***Index terms*—Personal information management, Thailand, university students, e-learning curriculum.**

I. Introduction

THE university system in Thailand began in 1917 with the establishment of the Chulalongkorn University in Bangkok. Today there are many different public and private universities in the country. The established universities in both the government and private sectors offer excellent programs especially in the fields of Medicine, the Arts, Humanities, and Information Technology, although many students prefer to pursue studies of law and business in Western faculties abroad or in those which have created local facilities in Thailand. For an overview of Thailand's educational system, refer to [1]. During the first years of the 21st century, the number of institutions called universities increased dramatically. As university students all over the world Thai students have to deal with the growing amount of data, information and knowledge being created, evaluated, used and disseminated in many different forms, not at least in e-learning activities.

Thailand has set up a National ICT Education Master Plan recently, as stated in [2], aiming amongst others at integrating technological knowledge (i.e. ICT knowledge) and information management skills to develop the ability to analyze, think creatively, solve problems, and work in teams. This has led to a large number of distance learning activities and projects in the country.

E-learning in Thailand has been studied by Suanpang and Petocz [3] as a case study on a course in Business Statistics at Suan Dusit Rajabhat University with some 1,000 participants. They found that online courses can generate a more favorable learning outcome, for example significantly higher grades, than traditional courses. They explain this in part with the higher use of advanced technological support and diverse materials by online students for their learning than the traditional students had received. They also found that students in online courses assess themselves with a higher score than those in traditional courses.

Students participating in online courses have to know how to find, use, and share information for learning. Personal information management refers to both the practice and the study of human activities to gather, organize, store, retrieve and use information objects such as paper-based, digital and online documents, web pages and email messages for everyday use to complete job-related and other tasks. Since students have to deal with a huge amount of information, personal information management seems to be a necessary skill to achieve; this is especially useful for online courses.

There is not much literature available that deals with the behavior of Thai students using information and the Internet. Focusing on web page design Vitartas and Sangkamanee provided a study on Thai students' use of the Internet ([4]). They carried out a survey among 170 students of Assumption University in Bangkok with the help of a questionnaire. The authors concede that “the findings are limited to a sample size which may not provide adequate representation of all Thai University students”.

In this study we want to show how prepared Thai university students are with regard to online education. For this, we investigated the personal information behavior of

Manuscript received April 25, 2009. Manuscript accepted for publication August 14, 2009.

The authors are with the Department of Computer Science, Faculty of Science, Naresuan University, Phitsanulok 65000, Thailand (e-mail: michaelb@nu.ac.th).

Thai university students with the help of paper-based, online and interview questionnaires to gain quantitative as well as qualitative data. The results of this study can be useful for setting up online or distance learning courses that use the Internet as a technological basis.

This paper is organized as follows. After the introduction we outline the research method of the questionnaires that led to the quantitative results presented and briefly discussed given in table form. After that, we introduce some of the results gained through the interviews. Finally, we draw conclusions and give some hints on further work related to this research.

II. Research Method

The methods used in this research were questionnaires and interviews to get quantitative as well as qualitative data for answering the research questions. A questionnaire is a research instrument consisting of a series of questions and other prompts for the purpose of gathering information from respondents. Although they are often designed for statistical analysis of the responses, this is not always the case. As a type of survey, questionnaires also have many of the same problems relating to question construction and wording that exist in other types of opinion polls.

The construction of a questionnaire needs careful consideration. As Peterson states, “the quality of the information obtained from a questionnaire is directly proportional to the quality of the questionnaire, which in turn is directly proportional to the quality of the question construction process” ([5]).

In this research we used a self-administered questionnaire approach for the quantitative aspects, which helped us to reach more people in the data gathering process. Additionally, we carried out a series of interviews with students from different faculties and years of study to set a basis for the qualitative data and using an adapted set of questions. Table I gives an overview of the number of open and closed questions used in this research.

TABLE I
Overview of the questionnaires used in this research.

Questionnaire	No. of closed questions	No. of open questions	Sum
Self-administered questionnaire	24	2	26
Questionnaire for interviews	13	10	23

During a two-month period we piloted the questionnaire in Thai language and evaluated the results for comprehensibility and integrity. Possible ambiguities were removed by adapting the wording of the questions and translating them into English for the comparison group. The questionnaire was then sent out to different tertiary educational institutions in Thailand and in other countries to gather the data. From the returned answer sheets we discarded those that were filled out incorrectly, e.g. sheets with missing answers.

To measure the reliability of the questionnaire we used the split-half method. The split-half method is one of the methods used to assess reliability. With this method, the questionnaire is administered to a group of respondents and then the items are split in half, for example odds and evens, for purposes of scoring. The results of the two halves are then compared. The split-half method offers a clear advantage in terms of time and resources over the test-retest and the alternative form methods in that it does not require the test to be administered twice to the same group of respondents. In this case the results could be doubtful, because the respondents might remember their first answer and just repeat it or choose to answer the questions this time in the opposite way.

For a more detailed view on the questions for the follow-up interviews refer to section "Qualitative results from the interviews" and the Appendix below.

III. Quantitative Results from the Paper-based and Online Questionnaires

In this section we present the main body of the most significant data collected through the paper-based and online questionnaires. The qualitative data extracted from the follow-up interviews are dealt with in the following section. In Table II the proportion of the male and female students is shown for each of the research instruments, i.e. the questionnaires and interviews.

TABLE II
Sample sizes of paper-based/online and follow-up interviews.

Type	Male	Female	Totals
Paper-based questionnaire	250	452	702
Online questionnaire	336	256	592
Interviews	11	12	23
Total	597	720	1317
Percentage	45.3%	54.7%	100%

We used a paper-based and an online questionnaire to generate the data needed for the research.

In total there were 1,317 students that took part in this research compared to the total number of university students of 2,25 Mio ([6]), so the general level of accuracy of our data can be calculated as 2.7% for the confidence interval with a 95 % confidence level. The number of respondents to the questionnaires also show a Gender Parity Index (GPI, the female-to-male ratio) of 1.20, which is similar compared to the actual GPI of enrolled students in tertiary educational institutions in Thailand, which has risen between 1998 and 2002 from 1.15 to 1.17 ([6]).

The main body of data is about the students' habits related to information. The questions in the questionnaires and interviews covered the following topics for which the results are given in the tables below:

The main body of data is about the students' habits related to information. The questions in the questionnaires and interviews covered the following topics for which the results are given in the tables below:

- Recognizing information needs (Table III),
- Seeking for information (Table IV),
- Using search engines (Table V),
- Information sources being used (Table VI),
- Using and evaluating information (Table VII),
- Dissemination of information (Table VIII).

TABLE III
RECOGNIZING INFORMATION NEEDS.

Recognizing information needs	Percentage
I recognize information myself	70.3
My teacher tells me when I need	24.1
Others	5.6
When noticing information needs	**Percentage**
Before an assignment	16.9
Before a test	16.6
For preparing class lectures	13.8
For doing research	11.2
For writing reports and papers	23.1
For entertainment	16.8
Others	1.6

Table III is about recognizing information needs by the students. Because Thai universities are much more like a school, it is not surprising that students tend to recognize information by themselves through the help of assignments, papers (reports) and tests they have to return to their teachers. The many reports Thai students typically have to prepare during the terms help them to understand information needs rather naturally. On the other hand, quite a few students seem to notice information needs mostly for entertainment (16.8 %).

In Table IV the main data for Thai students' information seeking behavior are summarized. These data were derived from a part of the questionnaire, in which we used the Likert scale ([7]). Whereas it is not surprising that the use of Internet search engines has become second nature to the students, their willingness to plan for their search is remarkably high. 53.7 % of the students plans "always" or "mostly" for their information seeking task. The low rates for using library tools as valuable information sources have been already noticed by many librarians and have led to various initiatives to retain students in library activities. The Internet as a source for information outnumbers textbooks by far: 83.6 % use the Internet "always" or "mostly", in parallel 61.6 % use textbooks "always" or "mostly". We believe that this behavior together with the high planning rate mentioned above offers good chances for offering online courses successfully.

Since search engines are used extensively to get to informative web sites it is interesting to see which of them are actually used.

Table V shows the figures we have got for the use of major search engines. These figures differ surprisingly from those of the USA gained from Sullivan ([8]). Local Thai search portals, i.e. Sanook, Kapook and Hunsa, are in sum equally popular as Google, although Google is accessible via a convenient interface in Thai language. Not surprisingly, Sullivan does lack data for the locally oriented Thai search engines.

TABLE IV
SEEKING FOR INFORMATION (IN PERCENT).

Question	Always	Mostly	Sometimes	Seldom	Never
Do you plan for seeking information?	11.1	42.6	43.0	2.6	0.7
How often do you use the Internet for finding specific information?	44.3	39.3	16.0	0.3	0.1
Do you use Internet search engines for seeking information?	46.2	33.9	16.1	2.4	1.4
How often do you use the online catalog of your library?	5.5	16.0	42.7	21.4	14.4
How often do you go to the library for seeking information?	6.6	24.9	50.0	15.1	3.3
Do you use textbooks for seeking information?	15.5	46.1	32.8	4.6	1.0

TABLE V
USE OF SEARCH ENGINES.

Search engine	Preferred usage in %	
	Thailand	USA
Google	37.7	49.2
Sanook	16.0	-
Kapook	13.1	-
MSN	12.6	9.6
Yahoo	10.1	23.8
Hunsa	6.4	-
Altavista	1.1	-
Excite	1.1	-
Lycos	0.3	-
Others	1.4	8.5

In Table VI the main sources of information for Thai students are presented. As can be seen clearly, online journals offer the most relevant source of information for the students, followed by human sources. Librarians seem to play a marginal role as a source of information.

In Table VII the students' use and evaluation of information is briefly shown. As expected there is a major tendency to copy and paste information from the Internet. More than 80% of the students seem to do this on a regular basis. On the other hand, students seem to be aware of possible inaccuracies of information found in the Internet and in other information sources. More than 70% check information before they use them. The students rank the reliability of the information sources as follows: textbooks (74.3%), academic journals (67.5%), Internet (53.4%9, TV and radio (53.9%), and newspapers (47.1%).

TABLE VI
THAI STUDENTS' MAIN INFORMATION SOURCES.

Sources	Percentage
Online journals	34.0
Colleagues/peers	24.7
Paper journals	22.6
Supervisors	13.1
Librarians	3.7
Others	1.8

TABLE VII
INFORMATION USAGE AND EVALUATION.

Question	Always	Mostly	Sometimes	Seldom	Never
Do you copy and paste information when you find it in the Internet?	36.9	46.9	14.9	1.0	0.3
Do you check information before you use it?	24.3	46.6	25.3	3.4	0.3
Do you think the information in textbooks is reliable?	19.4	54.9	24.0	1.4	0.3
Do you think the information in academic journals is reliable?	14.8	52.7	32.0	0.3	0.3
Do you think the information in the Internet is generally reliable?	8.0	45.4	44.3	2.1	0.1
Do you think the information in newspapers is reliable?	7.0	40.1	48.6	4.0	0.3
Do you think the information in TV and radio is reliable?	9.6	44.3	42.8	3.0	0.3

TABLE VIII
INFORMATION DISSEMINATION.

Questions	>4 times a year	>2 times a year	> once in 2 years	once a year	never
How often do you publish papers in academic journals?	2.7	9.0	10.7	10.1	67.5
How often do you give talks about your work (on conferences or in the labs)?	11.4	17.2	18.7	21.0	31.6
How often do you publish web sites/blogs?	16.5	17.2	18.2	20.5	27.5
How often do you write articles for newspapers?	2.6	6.2	6.4	7.7	77.1
How often do you write reports?	43.1	27.1	16.1	10.0	3.6
How often do you make reportages for radio and TV?	5.1	8.7	10.0	11.7	64.5

In Table VIII the students' behavior related to the dissemination information is presented. Here we asked the students about the frequency by which they give talks, publish papers, turn in reports, upload to blogs or web sites.

The answers were standardized for each kind of dissemination, i.e. "never", "once a year", "more than once in 2 yrs", "more than twice a year" and "more than 4 times a year".

IV. Qualitative Results from the Interviews

In this section we present some of the qualitative results gathered by interviews with 23 students. The interviews lasted between 39 minutes and 77 minutes with an average duration of 61 minutes and were carried out in the university premises. During the interviews notes were taken according to the predefined questions handed over two or three days in advance to let the participants understand the questions and interview practice.

To obtain more detailed information we used structured interviews with 23 questions that were pre-determined and identical for every interviewee. The interviewer was a university lecturer with a background in computer science and information technology speaking Thai and English. We read out the questions and noted the answers without commenting them.

Because of the limited space we present in the following only some of the results that we find most interesting.

One of the questions was "Do you explore general information sources, such as encyclopedias, to become more familiar with a topic? Which sources do you use?" 20 out of 23 students actually use such sources, and they indicated that they benefit from online encyclopedias, e.g. Wikipedia. This is not surprising, since the Wikipedia project is well-known and has a Thai version available since the end of 2004. We later asked "Have you ever published valuable information in the Internet, such as writing articles for the Wikipedia (not social networks, Hi5)?" This question was answered positively by only one student. Compared with the questionnaire results presented above, which showed that more than 33% of the students publish web sites and blogs regularly, this might be seen as a contradiction. Here we asked for "valuable information" published by the student, so we would assume that a high number of students is active more on social networks, e.g. Hi5, which in fact is a popular element of Thai students' Internet life. This is clearly a point for further study.

The question "Do you know the difference between primary and secondary sources?" was not understood by all students. Since this question cannot be explained, otherwise the interviewer would have given the answer by himself, it works as an examination question about information management. Two students mixed primary and secondary with most important and less important, respectively. 16 students failed to answer at all, making up 5 students who knew the correct answer.

The question "How do you notice when you have enough information and when you need to get more? And if you need more, do you have processes and mechanisms for getting it (e.g., interlibrary loan; using resources at other locations; obtaining images, videos, text, or sound)?" is a complex one. It can be answered in different ways, e.g. in practical terms of a project and with respect to search strategies (precision and recall), and we were curious how the students would respond to it. The students' answers were quite pragmatic as the slight majority (13 students) stated that they stop searching when they feel time pressure to perform the following tasks, e.g. writing the main body of a report, but 8 of this pragmatic group of students return to the searching process if they feel a further need for information. 6 students related the problem to identify, whether or not they had enough information, to the production process during the project. They said if they had a source for every statement they made in their publication and at least another source for the main points they felt satisfied with the amount of information gathered for their work.

We asked the question "Are you comfortable citing different kinds of information?" and got the general impression that the students do not know enough about referencing methods and styles. None of the students answered positively, and from our experience of working with Thai students for some years we can confirm this result.

The questionnaire created for the interviews is presented in the appendix of this paper.

V. Conclusions

In this research we used questionnaires and interviews to gather data about the personal information habits of Thai students. We used quantitative and qualitative methods to analyze the results, which lead to new insights for setting up appropriate learning materials related to the students' personal information management.

From the data we see the need at least for

- Making the students more aware of the problems of copying and pasting materials in their academic texts without citing correctly, and
- Instructing common and appropriate citation methods for sources used in their materials.

All of this will be integrated into a new curriculum on "Methods of Research" introduced recently at the Faculty of Science at Naresuan University.

VI. Further Work

This study focused mainly on the personal information behavior of Thai students. The comparison with the behavior of students from other countries and educational cultures is difficult because of the different focus groups and different questionnaires used. It is desirable to carry out a cross-cultural study preferably with a common questionnaire and similar focus groups to compare results for the personal information behavior of students in tertiary education.

For a more thorough picture of the personal information management of students it would be desirable to include the use of appropriate tools into the research focus, for example electronic address books and instant message archiving.

Further work has also to be done to incorporate best practices and useful knowledge for personal information management into the curriculum. Moreover, it would be worthwhile to find out what kind of information Thai students contribute to the Internet, since there are a considerable number of them actually publishing.

APPENDIX: COMPLETE QUESTIONNAIRE USED FOR THE INTERVIEWS (IN ENGLISH)

1. *Do you talk with instructors and participate in class or electronic discussions to identify a research topic, or any other information need?*

2. Do you explore general information sources, such as encyclopedias, to become more familiar with a topic?

3. How do you work out a focus for a research question or a research paper that you can manage?

4. When you look at resources, how do you identify the purpose and audience of potential resources (e.g. popular vs. scholarly, current vs. historical)?

5. Do you know the difference between primary and secondary sources?

6. How do you notice that you have enough information or that you need more? And if you need more, do you have processes and mechanisms for getting it (e.g. asking peers, using resources at other locations, obtaining images, videos, text, or sound)?

7. When you use an information retrieval system, such as an online database, have you considered how it is organized and what is the best way to get information from it?

8. How do you search? Do you use key words, controlled vocabulary, Boolean operators, proximity searching, and truncation?

9. Do you use different formats of information? How do you get it?

10. Do you use different resources to get information (e.g. Internet, local and national library, professional associations, institutional research offices, experts and practitioners)?

11. When you cannot find information, what do you do? Do you take a different approach and try again, refine your strategy? Refocus? Figure out ways to fill in gaps? Do it all again?

12. Do you consider yourself a computer-literate person? Do you know how to use computer functions reasonably well? How about other technologies?

13. How do you organize the information you use? Folders for subjects/topics? Chronological order?

14. Are you comfortable citing different kinds of information?

15. How do you pick out the important parts of what you find and read in order to find the parts useful for you? How do you decide when to paraphrase things and when to quote things - do you have any rule of thumb?

16. Do you evaluate information that you want to use? Which criteria do you use? Do you look for bias, prejudice?

17. Do you ever question the accuracy of the information you have found? Ever compare information from different sources to see whether one makes more sense than the other?

18. Have you ever tried to contact an "expert" to validate information you are finding?

19. Do you provide outlines when you are writing something - organize the information you are using in some way?

20. In doing a paper or a project - have you ever kept a journal or log of activities about how you got or evaluated information and then written up what you found?

21. Have you ever considered that a written report is not the best way to convey information - that video or something else might provide a better way to convey information?

22. Have you given any thought to privacy and security issues in the electronic environment?

23. How about plagiarism - are you confident that you know enough to avoid it?

24. Have you ever published valuable information in the Internet, such as writing articles for the Wikipedia (but not social networks and Hi5)?

REFERENCES

[1] S. Pitiyanuwat, S. and S. Sujiva, "Civic Education in Thailand. Policies and practices in schools," Bangkok: Chulalongkorn University Press 2005.

[2] D. Ainley, P. Arthur, P. Macklin and B. Rigby, "Thai Learning Technologies 2010. Capacity building of Thai education reform (CABTER) Stage 1 – Learning technologies," 2003. Available: http://www.onec.go.th/publication/a_tech/a_tech.pdf (last accessed October 16, 2008).

[3] P. Suanpang and P. Petocz, "E-Learning in Thailand: An Analysis and Case Study". *International Journal on E-Learning*, Chesapeake, VA: AACE, vol. 5, no. 3, pp. 415-438. 2006.

[4] P. Vivartas and S. Sangkamanee, "Profiling Thai students' use of the Internet: implications for Web page design," in *ACSILITE Conference*, Southern Cross University, Coos Harbour, Australia, 9-14 Dec 2000. Available: http://www.ascilite.org.au/conferences/coffs00/papers/peter_vitartas.pdf (last accessed October 14, 2008).

[5] R. A. Peterson, "Constructing effective questionnaires," London: Sage 2000.

[6] "UNESCO, Global Education Digest 2005. Comparing Education Statistics across the World," Montreal: UNESCO Institute for Statistics 2005. Available: http://www.uis.unesco.org/template/pdf/ged/2005/ged2005_en.pdf (last accessed October 13, 2008).

[7] R. Likert, "A technique for the measurement of attitudes," *Archives of Psychology*, vol. 140, no. 1, pp. 1–55, 1932.

[8] D. Sullivan, "Nielsen net ratings search engine ratings," 2006. Available online at http://searchenginewatch.com/showPage.html?page=2156451 (last accessed Oct. 14, 2008).

Journal Information and Instructions for Authors

I. Journal Information

"*Polibits*" is a half-yearly research journal published since 1989 by the Center for Technological Design and Development in Computer Science (CIDETEC) of the National Polytechnic Institute (IPN) in Mexico City, Mexico. The journal solicits original research papers in all areas of computer science and computer engineering, with emphasis on applied research.

The journal has double-blind review procedure. It publishes papers in English and Spanish.

Publication has no cost for the authors.

A. Main topics of interest

The journal publishes research papers in all areas of computer science and computer engineering, with emphasis on applied research.

More specifically, the main topics of interest include, though are not limited to, the following:

- Artificial Intelligence
- Natural Language Processing
- Fuzzy Logic
- Computer Vision
- Multiagent Systems
- Bioinformatics
- Neural Networks
- Evolutionary algorithms
- Knowledge Representation
- Expert Systems
- Intelligent Interfaces: Multimedia, Virtual Reality
- Machine Learning
- Pattern Recognition
- Intelligent Tutoring Systems
- Semantic Web
- Database Systems
- Data Mining
- Software Engineering
- Web Design
- Compilers
- Formal Languages
- Operating Systems
- Distributed Systems
- Parallelism
- Real Time Systems
- Algorithm Theory
- Scientific Computing
- High-Performance Computing
- Geo-processing
- Networks and Connectivity
- Cryptography
- Informatics Security
- Digital Systems Design
- Digital Signal Processing
- Control Systems
- Robotics
- Virtual Instrumentation
- Computer Architecture
- other.

B. Indexing

LatIndex, Periodica, eRevistas.

II. Instructions for Authors

A. Submission

Papers ready to review are received through the Web submission system www.easychair.org/polibits. See also the updated information at the web page of the journal www.cidetec.ipn.mx/polibits.

The papers can be written in English or Spanish.

Since the review procedure is double-blind, the full text of the papers should be submitted without names and affiliations of the authors and without any other data that reveals the authors' identity.

For review, a file in one of the following formats is to be submitted: PDF (preferred), PS, Word. In case of acceptance, you will need to upload your source file in Word or TeX. We will send you further instructions on uploading your camera-ready source files upon acceptance notification.

Deadline for the nearest issue (January-June 2010): May 1, 2010. Papers received after this date will be considered for the next issue (July-December 2010).

B. Format

Please, use IEEE format[1], see section "Template for all Transactions (except IEEE Transactions on Magnetics)". The editors keep the right to modify the format and style of the final version of the paper if necessary.

We do not have any specific page limit: we welcome both short and long papers, provided the quality and novelty of the paper adequately justifies the length.

In case of being written in Spanish, the paper should also contain the title, abstract, and keywords in English.

[1] www.ieee.org/web/publications/authors/transjnl/index.html