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).

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.

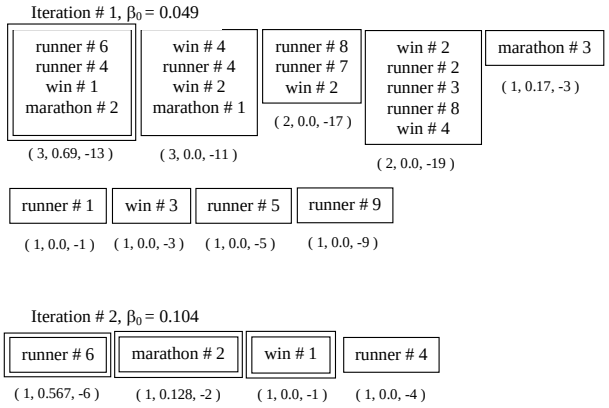

Fig. 2. Disambiguation of words in "*The runner won the marathon*".

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 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.