Using the Multilingual Central Repository for Graph-Based Word Sense Disambiguation

Eneko Agirre, Aitor Soroa

IXA NLP Group
University of Basque Country
Donostia, Basque Contry
a.soroa@ehu.es

Abstract

This paper presents the results of a graph-based method for performing knowledge-based Word Sense Disambiguation (WSD). The technique exploits the structural properties of the graph underlying the chosen knowledge base. The method is general, in the sense that it is not tied to any particular knowledge base, but in this work we have applied it to the Multilingual Central Repository (MCR, (Atserias et al., 2004)). The evaluation has been performed on the Senseval-3 all-words task (Snyder and Palmer, 2004). The main contributions of the paper are twofold: (1) We have evaluated the separate and combined performance of each type of relation in the MCR, and thus indirectly validated the contents of the MCR and their potential for WSD. (2) We obtain state-of-the-art results, and in fact yield the best results that can be obtained using publicly available data.

1. Introduction

Word Sense Disambiguation (WSD) is a key enabling-technology, the aim of which is to automatically choose the correct sense of a word in a context. Supervised WSD systems are the best performing in public evaluations but they need large amounts of hand-tagged data, which is typically very expensive to build. On the other hand, knowledge-based WSD systems exploit the information present on a lexical knowledge base (LKB) to perform WSD, without using any further corpus evidence.

Traditional knowledge-based WSD system assign a sense to an ambiguous word by comparing each of its senses with those of the surrounding context. Typically, some semantic similarity metric is used for calculating the relatedness among senses(Lesk, 1986; McCarthy et al., 2004). One of the major drawbacks on these systems comes from the fact that the WSD process is done in a word-by-word basis, i.e., the words are disambiguated individually.

Recently, graph-based methods for knowledge-based WSD have gained much attention in the NLP community (Sinha and Mihalcea, 2007; Navigli and Lapata, 2007; Mihalcea, 2005). These methods use well-known graph-based techniques to find and exploit the structural properties of the graph underlying the LKB. Because the graph is analyzed as a whole, these techniques can help to find globally optimal solutions given the relations between entities, and disambiguate large portions of text in one go.

In this paper, we present a novel graph-based method for performing unsupervised WSD. The method is general, in the sense that it is not tied to any particular lexical resource, but in this work we have applied it to the Multilingual Central Repository (MCR, (Atserias et al., 2004)). The MCR is based on the EuroWordNet architecture to represent tightly connected wordnets for various languages. It also contains thousands of automatically acquired relations. The main goal of this work is to show that the relations in the MCR are valuable for WSD, and that our WSD algorithm obtains state-of-the-art results compared to other more complex graph-based algorithms.

The WSD method works as follows. Given an input context, the method first explores the whole LKB and finds a subgraph which is particularly relevant for the words of the context. Then, a random-walk algorithm is executed over the subgraph for obtaining a ranking of vertices. As a result, every word of the context is attached to the highest ranking concept among its possible senses.

The paper is organized as follows. We first describe our graph based method in Section 2. Section 3 presents the Multilingual Central Repository. Section 4 show experimental setting and results. Section 5 compares our system with related experiments on graph-based WSD. Finally, we draw conclusions and further work.

2. Graph-based techniques for WSD

In this section we will describe our general framework for graph-based WSD. The method is general and none of the processes of the method are particular to any particular knowledge base. However, in our work we have used the MCR as the source of lexical-semantic knowledge.

In a first step, we build a graph G_{MCR} which represents the MCR knowledge base: each synset (concept) of the MCR is represented by a node v_i on the graph, and each relation between concepts v_i and v_j is represented by an edge $e_{i,j}$.

Given an input text, we extract the list W_i $i = 1 \ldots m$ of content words (i.e. nouns, verbs, adjectives and adverbs), where each of them will have a set of synsets associated in the MCR. Let $Synsets_i = \{v_1, \ldots, v_{i_m}\}$ be the i_m associated synsets of word W_i in the MCR. Note that monosemous words will be related to only one synset in the MCR, whereas polysemous words may be attached to several synsets. As a result of the disambiguation process, for each word in the sentence we will select the sense in the MCR with the highest ranking synset in the graph. We will now explain how we create subgraph and the method to rank the synsets contained in it.

Name	Relations	#synsets	#relations
M16	WN1.6, REL2.0, XNET, sPref, sCooc	99,634	1,651,445
M16_wout_sPref	WN1.6, REL2.0, XNET, sCooc	99,634	1,519,833
M16_wout_sCooc	WN1.6, REL2.0, XNET, sPref	99,632	798,453
M16_wout_Xnet	WN1.6, REL2.0, sPref, sCooc	99,238	1,169,300
M16_wout_Semcor	WN1.6, REL2.0, XNET	99,632	637,290
M16_wout_WXnet	sPref, sCooc	27,336	1,024,698
M17	WN1.7, XNET	109,359	620,396

Table 1: Groupings of the relations in the MCR used in the paper.

Creating a disambiguation subgraph for a given context

The main idea of our method is that to extract the subgraph of G_{MCR} whose vertices and relations are particularly relevant for a given input context. We call such a subgraph a "disambiguation subgraph" G_D, and we build it in the following way. For each word W_i in the input context and each synset $v_i \in Synsets_i$, we perform a standard breath-first search (BFS) over G_{MCR} starting at node v_i. Each run of the BFS calculates the minimum distance paths between v_i and the rest of concepts of G_{MCR}. In particular, we are interested in the minimum distance paths between v_i and the concepts associated to the rest of the words in the context, $v_j \in \bigcup_{j \neq i} Synsets_j$. Let mdp_{v_i} be the set of these shortest paths.

We repeat this BFS computation for every synset of every word in the input context, storing mdp_{v_i} accordingly. At the end, we obtain a set of minimum length paths each of them having a different concept as a source. We then build the disambiguation graph G_D which is just the union of the vertices and edges of the shortest paths, $G_D = \bigcup_{i=1}^{m} \{mdp_{v_j} / v_j \in Synsets_i\}$.

The disambiguation graph G_D is thus a subgraph of the original G_{MCR} graph obtained by computing the shortest paths between the synsets of the words co-occurring in the context. Thus, we can assume that it captures the most relevant concepts and relations in the MCR for the particular input context.

Exploiting structural properties of the disambiguation graph

Once the G_D graph is built, we compute the PageRank (Brin and Page, 1998) algorithm over it. The intuition behind this step is that the vertices representing the correct synsets will be more relevant in G_D than the rest of the possible synsets of the context words, which should have less relations on average and be more isolated.

PageRank is an iterative algorithm that ranks all the vertices of a graph according to their relative importance within the graph, following a random-walk model. In this model, a link between vertices v_1 and v_2 means that v_1 *recommends* v_2. The more vertices recommend v_2, the higher the rank of v_2 will be. Furthermore, the rank of a vertex depends not only on how many vertices point to it, but on the rank of these vertices as well.

Specifically, let $G = (V, E)$ be a graph with the set of vertices V and set of edges E. For a given vertex v_i, let $\mathrm{In}(v_i)$ be the set of vertices pointing to it, and let d_j the degree of vertex v_j. The rank of v_i is defined as:

$$P(v_i) = (1 - \alpha) + \alpha \sum_{j \in \mathrm{In}(v_i)} \frac{1}{d_j} P(v_j)$$

where $0 \leq \alpha \leq 1$. α is called the *damping factor* and models the probability of a web surfer standing at a vertex to follow a link from this vertex (probability α) or to jump to a random vertex in the graph (probability $1 - \alpha$). The factor is usually set at 0.85.

The algorithm initializes the ranks of the vertex with a fixed value (usually $\frac{1}{N}$ for a graph with N vertices) and iterates until convergence below a given threshold is achieved, or, more typically, until a fixed number of iterations are executed. Note that the convergence of the algorithms doesn't depend of the initial value of the ranks.

After running the algorithm, the vertices of G_D are ordered in decreasing order according to its rank. Finally, the disambiguation step is performed by assigning to each word W_i the associated concept in $Synsets_i$ which has maximum rank (in case of ties, we assign all the concepts with maximum rank).

3. Multilingual Central Repository

The Multilingual Central Repository (Atserias et al., 2004) is a lexical knowledge base built within the MEANING project[1], and acts as a multilingual interface for integrating and distributing all the knowledge acquired in the project. The MCR constitutes a natural multilingual large scale linguistic resource for a number of semantic processes that need large amounts of multilingual knowledge to be effective tools.

The current version of the MCR contains more than $1,500,000$ semantic relations between synsets, most of them acquired by automatic means. This represents almost seven times more relations than the Princeton WordNet (Fellbaum, 1998) ($235,402$ unique semantic relations in WN 3.0).

The MCR follows the model proposed by the EWN project, whose architecture includes the Inter-Lingual-Index (ILI), a Domain ontology and a Top Concept ontology.

The current MCR integrates: (i) the ILI based in WN1.6, including EWN Base Concepts, the EWN Top Concept ontology, MultiWordNet Domains (MWND) and the Suggested Upper Merged Ontology (SUMO); (ii) Local WNs connected to the ILI, including English WN 1.5, 1.6, 1.7, 1.7.1, Basque, Catalan, Italian and Spanish wordnets; (iii)

---

[1] http://nipadio.lsi.upc.es/nlp/meaning

1389

Large collections of semantic preferences, acquired both from SemCor and from BNC; (iv) disambiguated glosses from the eXtended WordNet (Harabagiu and Moldovan, 2000); (v) Instances, including named entities.

The MCR is mainly based on WordNet 1.6. The reason is that most of the resources integrated on it where built linked to that version of WordNet. Unfortunately new versions of WordNet have been produced in later years, and although automatic mapping technology exists (Daude et al., 2000) and porting the relations among versions is possible, some errors are introduced, as we will see.

In this work we have used the following relations from the MCR:

- WN1.6: English WordNet 1.6 synsets and relations

- WN1.7: English WordNet 1.7 synsets and relations

- REL2.0: English WordNet 2.0 relations (to be added to WN1.6)

- XNET: eXtended WordNet (gold, silver and normal)

- sPref: Selectional preference relations

- sCooc: Coocurrence relations

The latter two types of relations (sPref and sCooc) are extracted from Semcor, a semantically hand-tagged corpus (Miller et al., 1993). They are thus essentially different from the other relations of the MCR; because they are extracted from a hand-tagged corpus, the system benefits from a "supervised" kind of information when exploiting sPref or sCooc relations in our algorithm.

One of our main objectives is to measure the impact of the different relations in the MCR on the WSD task. Therefore, we have grouped different MCR relation types together, and tried the same algorithm with each of them. Table 1 shows the relation groups and their sizes (number of synsets and relations). We basically have two main groups of graphs: those based on version 1.6 of WordNet (all of which start with M16) and those based on version 1.7 of WordNet (starting with M17). Version 1.7 is interesting because it is the one used in the evaluation, cf. Section 4..

4. Experiments

We have applied our graph-based technique to the Senseval-3 all words dataset (Snyder and Palmer, 2004), which is based on version 1.7 of Wordnet. For each sentence to be disambiguated we build a context of at least 20 words, taking the sentences immediately before and after it in the case that the original sentence was too short.

We have tried different sets of relations in several runs of the algorithm. Table 2 summarizes the performance of the algorithm for the different MCR relations. The table is divided in two sections, depending whether relations extracted from Semcor are used ("Semi supervised") or not ("Unsupervised").

We can see that adding supervised relations achieve the best overall results. In particular, the results indicate that coocurrence relations between synsets are a powerful source of information. If we discard coocurrence relations from the LKB the performance of the system drops

Relations	All	Noun	Verb	Adj.	Adv.
Semi supervised					
M16	57.30	62.30	49.00	**62.40**	92.90
M16_wout_sPref	**57.90**	**63.10**	**49.80**	61.80	92.90
M16_wout_sCooc	53.00	58.10	44.20	58.30	92.90
M16_wout_Xnet	57.60	**63.10**	49.60	61.00	92.90
M16_wout_WXnet	55.30	58.70	48.70	60.80	85.70
Unsupervised					
M16_wout_semcor	53.70	59.50	45.00	57.80	92.90
M17	**56.20**	**61.60**	**47.30**	61.80	92.90

Table 2: Results of Senseval-3 All Words as recall

more than 4 percentage points (as can be seen in row M16_wout_sCooc in Table 2). On the other hand, selectional preference relations don't seem to be as useful as coocurrence relations.

Somehow surprisingly, the system performs only slightly worse when using only supervised information (i.e. without any hierarchical relations of WordNet), as indicated by the row M16_wout_WXnet. Because the size of the LKB when considering only supervised relations is one third of the LKB with all relations (see Table 1), we would expect a heavy coverage penalty on the algorithm, and thus a performance drop in the overall recall. However, this drop is not produced, which indicates that large parts of WordNet itself are never considered when applying the WSD algorithm to this data set, and that the relations extracted from SemCor contain the most important synsets that need to be considered.

Regarding the unsupervised results (row M16_wout_semcor), the performance drops as expected, but the obtained results are of great merit, as no hand-tagged corpora are being used. The experiment we did with WordNet 1.7 (row M17), shows that it yields better results than WordNet 1.6. The reason, may lye on the fact that the Senseval 3 all words data set is tagged using WordNet 1.7.1 synsets, and thus the mapping step which is used with the LKBs based on WordNet 1.6 introduces considerable noise.

5. Comparison to related work

We have compared our results with similar work which also use graph techniques for performing unsupervised, knowledge based WSD, namely (Mihalcea, 2005; Sinha and Mihalcea, 2007; Navigli and Lapata, 2007). Table 3 shows a comparison of their results with ours. We only depict the results of out best unsupervised experiment M17, as the rest of the systems are also of unsupervised nature. A baseline which consists on selecting the most frequent sense in Semcor of every word is also included, as well as the result of the best system of Senseval-3 all word task (GAMBL). Note that GAMBL is a supervised system which learns from Semcor information.

The TexRank algorithm (Mihalcea, 2005) for WSD creates a complete weighted graph (e.g. a graph where every pair of distinct vertices is connected by an weighted edge) formed by the synsets of the words in the input context. The weight of the links joining two synsets is calculated by executing Lesk's algorithm between them, i.e., by calculating an over-

lapping factor of the words in the sense's definition glosses. Once the complete graph is built, PageRank algorithm is executed over it and words are assigned to the most relevant synset. Their results on Senseval-3 dataset are depicted in row Mih05 on Table 3.

(Sinha and Mihalcea, 2007) extends the previous work by using a collection of semantic similarity measures when assigning a weight to the links across synsets. They also compare different graph-based centrality algorithms to rank the vertices of the complete graph. Using different similarity metrics for different POS types and a voting scheme among the centrality algorimth ranks yields the results on row Sin07 on Table 3. Here, the Senseval-3 corpus was used as a development data set, and we can thus see the results as the upperbound of their method. Both (Mihalcea, 2005) and (Sinha and Mihalcea, 2007) use WordNet as a LKB, but neither of them specify which particular version did they use.

We can see in Table 3 that our method clearly outperforms both Mih05 and Sin07, which indicates that the initial extraction of a LKB's subgraph is an important step to perform. We can see that storing intermediate LKB nodes from the shortest paths between synsets adds useful information when performing WSD. The results of various inhouse made experiments with complete graphs a-la (Mihalcea, 2005) also confirm this observation. Note also that our method is simpler than the combination strategy used in (Sinha and Mihalcea, 2007), and that we did not perform any parameter tuning as they did.

The work presented here is very similar to (Navigli and Lapata, 2007). They also perform a two stage process for WSD, the first one consisting on the extraction of a subgraph from the LKB which better suits the concepts involved in a particular input context. Then, they study different graph-based centrality algorithms deciding the relevance of the nodes on the subgraph. The algorithm which performs best yields the results shown in row Nav07 on table 3.

The main difference between the method used in (Navigli and Lapata, 2007) and our method lies on the initial method for extracting the context subgraph. Whereas we rely on shortest paths between word synsets, they apply a depth-first search algorithm over the LKB graph, and restrict the deep of the subtree to a value of 3. As for the underlying LKB, they use WordNet 2.0 enriched with several new relations as described in (Navigli and Velardi, 2005). Unfortunately, those new relations are not publicly available, so we can't perform a direct comparison.

Navigli and Lapata don't report overall results and therefore, we can't directly compare our results with theirs. However, we can see that a on a POS-basis evaluation our results are rather similar for nouns and adjectives, and perform better for verbs.

6. Conclusion and further work

In this paper we have presented the results of a graph-based method for performing knowledge-based Word Sense Disambiguation (WSD). The technique exploits the structural properties of the graph underlying the chosen knowledge base. The method is general, in the sense that it is not tied to

System	All	Noun	Verb	Adj.	Adv.
Mih05	52.2	-	-	-	-
Sin07	52.4	60.45	40.57	54.14	100
Nav07	-	61.9	36.1	62.8	-
M17	56.20	61.60	47.30	61.80	92.90
MFS	60.9-62.4	-	-	-	-
GAMBL	65.1	-	-	-	-

Table 3: Comparison with related work on Senseval-3 All Words dataset. *Sin07*: Senseval-3 corpus used was development data set. *Nav07*: the authors report the results only by POS. *M17*: our best unsupervised experiment. *MFS*: baseline performance varies depending on treatment of multiwords and hyphenated words. *GAMBL*: best supervised system.

any particular knowledge base, but in this work we have applied it to the Multilingual Central Repository (MCR, (Atserias et al., 2004)).

The evaluation has been performed on the Senseval-3 all-words task (Snyder and Palmer, 2004). One of the contributions of this paper is that we have evaluated the separate and combined performance of each type of relation in the MCR, and thus indirectly validated the contents of the MCR and their potential for WSD. The analysis shows that all the relations in the MCR are valuable for performing WSD, and that the relations coming from hand-tagged corpora are the most valuable, as could be expected. The version of WordNet is highly relevant. In fact, using the same version as of the test corpus proved to be a key issue, as it allowed to obtain better results.

The other main contribution of this paper is to show that our graph-based WSD system is competitive with the current state-of-the-art, and in fact yields the best results that can be obtained using publicly available data.

7. References

J. Atserias, L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, and P. Vossen. 2004. The meaning multilingual central repository. In *In Proceedings of GWC*, Brno, Czech Republic.

S. Brin and L. Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7).

J. Daude, L. Padro, and G. Rigau. 2000. Mapping WordNets using structural information. In *38th Anual Meeting of the Association for Computational Linguistics (ACL'2000)*, Hong Kong.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

S. M. Harabagiu and D. I. Moldovan. 2000. Enriching the wordnet taxonomy with contextual knowledge acquired from text. pages 301–333.

M. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26, New York, NY, USA. ACM.

D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. 2004. Finding predominant word senses in untagged text. In

ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, page 279, Morristown, NJ, USA. Association for Computational Linguistics.

R. Mihalcea. 2005. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In *Proceedings of HLT05*, Morristown, NJ, USA.

G.A. Miller, C. Leacock, R. Tengi, and R.Bunker. 1993. A semantic concordance. In *Proc. of the ARPA HLT workshop*.

R. Navigli and M. Lapata. 2007. Graph connectivity measures for unsupervised word sense disambiguation. In *IJCAI*.

R. Navigli and P. Velardi. 2005. Structural semantic interconnections: A knowledge-based approach to word sense disambiguation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(7):1075–1086.

R. Sinha and R. Mihalcea. 2007. Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. In *Proceedings of the IEEE International Conference on Semantic Computing (ICSC 2007)*, Irvine, CA, USA.

B. Snyder and M. Palmer. 2004. The english all-words task. In *Proc. of SENSEVAL*.