Word Sense Disambiguation as an Integer Linear Programming Problem

Vicky Panagiotopoulou[1], Iraklis Varlamis[2], Ion Androutsopoulos[1], and George Tsatsaronis[3]

[1] Department of Informatics, Athens University of Economics and Business, Greece
[2] Department of Informatics and Telematics, Harokopio University, Athens, Greece
[3] Biotechnology Center (BIOTEC), Technische Universität Dresden, Germany

Abstract. We present an integer linear programming model of word sense disambiguation. Given a sentence, an inventory of possible senses per word, and a sense relatedness measure, the model assigns to the sentences's word occurrences the senses that maximize the total pairwise sense relatedness. Experimental results show that our model, with two unsupervised sense relatedness measures, compares well against two other prominent unsupervised word sense disambiguation methods.

1 Introduction

Word sense disambiguation (WSD) aims to identify the correct sense of each word occurrence in a given sentence or other text span [11]. When the possible senses per word are known, supervised learning methods currently achieve the best results [3, 7], but they require manually sense-tagged corpora as training data. Constructing such corpora is costly; and the performance of supervised WSD methods may degrade on texts of different topics or genres than those of the training data [1]. Here we focus on *unsupervised* methods, meaning methods that do not require sense-tagged corpora [2, 5, 10]. We assume, however, that the possible word senses are known, unlike other unsupervised methods that also discover the inventory of possible senses [17].

Many state of the art unsupervised WSD methods construct a large semantic graph for each input sentence. There are nodes for all the possible senses of the sentence's words, but also for all the possible senses of words that a thesaurus, typically WordNet, shows as related to the sentence's words. The graph's edges correspond to lexical relations (e.g., hyponymy, synonymy) retrieved from the thesaurus and they may be weighted (e.g., depending on the types of the lexical relations). Algorithms like PageRank or activation propagation are then used to select the most active node (sense) of each word [2, 18, 19].

By contrast, we model WSD as an integer linear programming (ILP) problem, where the goal is to select exactly one possible sense of each word in the input sentence, so as to maximize the total pairwise relatedness between the selected senses. Our model can also be seen as operating on a graph of word senses, but the graph includes nodes only for the possible senses of the words in the input sentence, not other related words; hence, it is much smaller compared to the graphs of previous methods. Furthermore, the (weighted) edges of our graphs do not necessarily correspond to single lexical relations of a thesaurus; they represent the scores of a sense relatedness measure. Any pairwise sense relatedness (or similarity) measure can be used, including measures that consider all the possible paths (not single lexical relations) in WordNet connecting the two senses [20], or statistical distributional similarity measures. It is unclear how measures of this kind could be used with previous graph-based WSD approaches, where the graph's edges correspond to single lexical relations of a thesaurus.

To our knowledge, our model is the first ILP formulation of WSD. Although ILP is NP-hard, efficient solvers are available, and in practice our method is faster than implementations of other unsupervised WSD methods, because of its much smaller graphs. A major advantage of our ILP model is that it can be used with any sense relatedness measure. As a starting point, we test it with (i) *SR* [20], a measure that considers all the possible WordNet paths between two senses, and (ii) a Lesk-like [5] measure that computes the similarity between the WordNet glosses of two senses using pointwise mutual information (PMI) [6, 22] and word co-occurrence statistics from a large corpus without sense tags. With these two measures, our overall method is unsupervised. It is also possible to use statistical sense relatedness measures estimated from sense-tagged corpora, turning our method into a supervised one, but we reserve this for future work.

Section 2 below introduces our ILP model; Section 3 defines the two sense relatedness measures we adopted; Section 4 presents experimental results against two other prominent unsupervised WSD methods; and Section 5 concludes.

2 Our ILP model of Word Sense Disambiguation

Let $w_1, \ldots, w_n$ be the word occurrences of an input sentence; s_{ij} denotes the j-th possible sense of w_i, and $rel(s_{ij}, s_{i'j'})$ the relatedness between senses s_{ij} and $s_{i'j'}$. The goal is to select exactly one of the possible senses s_{ij} of each w_i, so that the total pairwise relatedness of the selected senses will be maximum. For each sense s_{ij}, a binary variable a_{ij} indicates if the sense is active, i.e., if it has been selected ($a_{ij} = 1$) or not ($a_{ij} = 0$). A first approach would be to maximize the objective function (1) below, where we require $i < i'$ assuming that the relatedness measure is symmetric, subject to the constraints (2). The last constraint ensures that exactly one sense s_{ij} is active for each w_i.

$$\text{maximize} \quad \sum_{i,j,i,'j',i<i'} rel(s_{ij}, s_{i'j'}) \cdot a_{ij} \cdot a_{i'j'} \tag{1}$$

$$\text{subject to} \quad a_{ij} \in \{0, 1\},\ \forall i, j \quad \text{and} \quad \sum_j a_{ij} = 1,\ \forall i. \tag{2}$$

The objective (1) is *quadratic*, because it is the weighted sum of products of variable pairs ($a_{ij} \cdot a_{i'j'}$). To obtain an ILP problem, we introduce a binary variable $\delta_{ij,i'j'}$ for each pair of senses $s_{ij}, s_{i'j'}$ with $i \neq i'$. Figure 1 illustrates the new formulation of the problem. Each one of the large circles, hereafter called *clouds*, contains the possible senses (small circles) of a particular word occurrence w_i. There must be exactly one active sense in each cloud. Each $\delta_{ij,i'j'}$ variable shows if the edge that connects two senses s_{ij} and $s_{i'j'}$ from different clouds is active ($\delta_{ij,i'j'} = 1$) or not ($\delta_{ij,i'j'} = 0$). We want an edge to be active if and only if both of the senses it connects are active ($a_{ij} = a_{i'j'} = 1$).

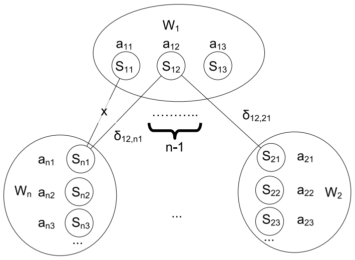

Fig. 1. Illustration of our ILP model of word sense disambiguation.

The problem can now be formulated as follows, where $i, i' \in \{1, \dots, n\}$, $i \neq i'$, and $(i, j), (i', j')$ range over the indices of the possible s_{ij} and $s_{i'j'}$, respectively.

$$\text{maximize} \quad \sum_{i,j,i,'j',i<i'} rel(s_{ij}, s_{i'j'}) \cdot \delta_{ij,i'j'} \tag{3}$$

$$\text{such that} \quad a_{ij} \in \{0,1\},\ \forall i, j \quad \text{and} \quad \sum_j a_{ij} = 1,\ \forall i \tag{4}$$

$$\delta_{ij,i'j'} \in \{0,1\} \quad \text{and} \quad \delta_{ij,i'j'} = \delta_{i'j',ij},\ \forall i, j, i', j' \tag{5}$$

$$\text{and} \sum_{j'} \delta_{ij,i'j'} = a_{ij},\ \forall i, j, i'. \tag{6}$$

The second constraint of (5) reflects the fact that the edges (and their activations) are not directed. Constraint (6) can be understood by considering separately the possible values of a_{ij}:

- If $a_{ij} = 0$ (s_{ij} is inactive), $\sum_{j'} \delta_{ij,i'j'} = 0,\ \forall i'$, i.e., all the edges that connect s_{ij} to the senses $s_{i'j'}$ of each other word (cloud) $w_{i'}$ are inactive, enforcing the requirement that any edge connecting an inactive sense must be inactive.
- If $a_{ij} = 1$ (s_{ij} is active), then $\sum_{j'} \delta_{ij,i'j'} = 1,\ \forall i'$, i.e., there is exactly one active edge connecting s_{ij} to the senses $s_{i'j'}$ of each other word (cloud) $w_{i'}$. The active edge from s_{ij} connects to the (single) active sense in the cloud of $w_{i'}$, because if it connected to a non-active sense in that cloud, the edge would have to be inactive, as in the previous case. Hence, the active edge connects two active senses (from different clouds), as required.

An advantage of ILP solvers is that they guarantee finding an optimal solution, if one exists. As already noted, ILP is NP-hard, but efficient solvers are available, and they are very fast when the number of variables and constraints is reasonably small, as in our case.[4] We also implemented a pruning variant of our ILP method, which removes from the graph of Fig. 1 any sense s_{ij} whose WordNet gloss contains none of the other word occurrences being disambiguated; the pruning is not applied to the senses of word occurrences that would be left without any sense after the pruning. This pruning significantly reduces the number of candidate senses and, consequently, the execution time of our method.

3 Relatedness Measures

Lexical relatedness measures can be classified in three categories: (i) measures based on dictionaries, thesauri, ontologies, or Wikipedia hyperlinks, collectively called knowledge-based measures [4, 14]; (ii) corpus-based measures, which use word or sense co-occurrence statistics, like PMI and χ^2 [6, 22]; and (iii) hybrid measures [16, 9]. Some measures are actually intended to assess the relatedness between words (or phrases), not word senses, but they can often be modified to work with senses. Other measures are intended to measure similarity, not relatedness, though the distinction is not always clear. The first measure that we adopt, *SR*, uses WordNet and belongs in the first category. The second measure is a hybrid one, since it uses both word co-occurrence statistics (to compute PMI scores) and WordNet's glosses.

SR [20] requires a hierarchical thesaurus O with lexical relations, in our case WordNet, and a weighting scheme for lexical relations. Given a pair of senses s_1, s_2 and a path (sequence) $P = \langle p_1, \ldots, p_l \rangle$ of senses connecting $s_1 = p_1$ to $s_2 = p_l$ via lexical relations, P's "semantic compactness" (SCM) is defined as below; $w_{i \to i+1}$ are the weights of the lexical relations (sense to sense transitions) of P.[5] The "semantic path elaboration" (SPE) of P is also defined below; d_i is the depth of p_i in O's hierarchy, and d_{max} is the maximum depth of O.

$$SCM(P) = \prod_{i=1}^{l-1} w_{i \to i+1} \qquad SPE(P) = \prod_{i=1}^{l} \frac{2 d_i d_{i+1}}{d_i + d_{i+1}} \cdot \frac{1}{d_{max}}$$

The semantic relatedness SR between s_1 and s_2 is defined below, where P ranges over all the paths connecting s_1 to s_2. If no such path exists, then $SR(s_1, s_2) = 0$.

$$SR(s_1, s_2) = \max_{P = \langle s_1, \ldots, s_2 \rangle} \{SCM(P) \cdot SPE(P)\}$$

Instead of $SR(s_1, s_2)$, we use $e^{SR(s_1, s_2)}$, which leads to slightly better results.

For two words w_1, w_2, their PMI score is $PMI(w_1, w_2) = \log \frac{P(w_1, w_2)}{P(w_1) \cdot P(w_2)}$, where $P(w_1, w_2)$ is the probability of w_1, w_2 co-occurring (e.g., in the same

[4] We use LP_SOLVE; see `http://lpsolve.sourceforge.net/`.

[5] P is a path on WordNet's entire graph, not the graph of Fig. 1 that we construct for each sentence. A Web service implementation of *SR* with precomputed *SR* scores for all WordNet senses is available; consult `http://omiotis.hua.gr/`.

sentence). If w_1, w_2 are independent, their PMI score is zero. If w_1, w_2 always co-occur, the score is maximum, equal to $-\log P(w_1) = -\log P(w_2)$. With sense-tagged corpora, the PMI score of two senses s_1, s_2 can be estimated similarly. In this paper, however, where we do not use sense-tagged corpora, we use the WordNet glosses $g(s_1)$ and $g(s_2)$ of s_1 and s_2, and the PMI scores of all word pairs w_1, w_2 from $g(s_1)$ and $g(s_2)$, respectively, excluding stop-words:

$$PMI(s_1, s_2) = \frac{\sum_{w_1 \in g(s_1),\ w_2 \in g(s_2)} PMI(w_1, w_2)}{|g(s_1)| \cdot |g(s_2)|}$$

Here $|g(s)|$ is the length of $g(s)$ in words. The intuition is that if s_1 and s_2 are related, the words that are used in their glosses will also co-occur frequently. We use an untagged corpus of approx. 953 million tokens to estimate $PMI(w_1, w_2)$.

4 Experimental Evaluation

We call ILP-SR-FULL and ILP-SR-PRUN the versions of our ILP method with and without sense pruning when the *SR* measure is used, and ILP-PMI-FULL and ILP-PMI-PRUN the versions with the PMI-based measure. We experimented with the widely used Senseval 2 and 3 datasets, whose word occurrences are tagged with the correct WordNet senses. Both datasets have training and test parts.

We compare our ILP approach against two other prominent unsupervised WSD methods, both of which construct a large semantic graph for each input sentence. The graph has nodes not only for all the possible senses of the sentence's words, but also for all the possible senses of the words that WordNet shows as related to the sentence's words. The edges of the graph correspond to single lexical relations of WordNet. (Recall that, by contrast, our ILP approach constructs a much smaller graph for each sentence, which only contains nodes for the possible senses of the sentence's words; and the edges of our graph do not necessarily correspond to single WordNet lexical relations.) The first method we compare against, Spreading Activation Network (SAN), consequently applies a spreading activation to the semantic graph, and eventually retains the most active sense (node) of each word of the input sentence. We use the SAN method of Tsatsaronis et al. [21], which is an improved version and, hence, representative of several other SAN methods for WSD going back to Quillian [15]. The second method we compare against, hereafter called PR, applies PageRank on the semantic graph and retains the most highly ranked sense (node) of each word in the input sentence. PageRank was first used in WSD by Mihalcea et al. [8], but with different improvements it has also been used by others [2, 19]. We use the PR method that was recently evaluated by Tsatsaronis et al. [19]. The SAN and PR methods were chosen because they are well-known and implementations of both were available to us, unlike other unsupervised WSD methods [18, 12, 2].

When they cannot disambiguate (at all, or with high confidence) a word occurrence, many unsupervised WSD methods resort to the *first-sense heuristic*, which selects the first sense of each word, as listed in WordNet. Te first sense is the most common one, based on frequencies from sense-tagged corpora; hence,

Senseval 2	Noun				Verb				Adjective				All			
Method	**C**	**P**	**R**	**F**	**C**	**P**	**R**	**F**	**C**	**P**	**R**	**F**	**C**	**P**	**R**	**F**
san	72.2	27.8	20.0	23.3	71.1	19.6	13.9	16.3	72.4	39.6	28.7	33.3	71.9	27.9	20.0	23.3
pr	72.2	45.5	32.8	38.1	71.1	30.0	21.3	24.9	72.4	38.8	28.1	32.6	71.9	39.4	28.4	33.0
ilp-sr-full	99.6	38.6	38.4	38.5	99.6	25.0	24.9	24.9	92.8	37.4	34.7	36.0	98.1	34.2	33.5	**33.8**
ilp-sr-prun	99.6	38.6	38.4	38.5	99.6	24.6	24.5	24.5	92.8	37.7	35.0	36.3	98.1	34.1	34.4	**33.8**
ilp-pmi-full	99.6	27.9	27.7	27.8	98.9	23.4	23.2	23.3	100.0	37.9	37.9	37.9	99.5	28.6	28.4	28.5
ilp-pmi-prun	99.6	28.6	28.5	28.6	98.9	24.7	24.5	24.6	100.0	43.5	43.5	43.5	99.5	30.5	30.4	30.5
Senseval 3	Noun				Verb				Adjective				All			
Method	**C**	**P**	**R**	**F**	**C**	**P**	**R**	**F**	**C**	**P**	**R**	**F**	**C**	**P**	**R**	**F**
san	97.9	30.6	29.9	30.2	94.2	28.8	27.1	27.9	94.9	37.8	35.9	36.8	95.8	31.0	29.7	30.4
pr	97.9	38.3	37.5	37.9	94.2	39.6	37.3	38.4	94.9	40.5	38.4	39.4	95.8	39.2	37.6	**38.4**
ilp-sr-full	99.9	32.3	32.2	32.3	98.0	25.8	25.3	25.6	97.0	38.3	37.1	37.7	98.6	30.6	30.2	30.4
ilp-sr-prun	99.9	32.0	31.9	32.0	98.0	25.8	25.3	25.6	97.0	38.7	37.5	38.1	98.6	30.5	30.1	30.3
ilp-pmi-full	96.7	30.2	29.2	29.7	94.1	18.1	17.1	17.6	96.9	39.4	38.2	38.8	95.7	26.9	25.8	26.3
ilp-pmi-prun	96.7	27.3	26.4	26.8	94.1	19.3	18.2	18.7	96.9	39.0	37.8	38.4	95.7	26.1	24.9	25.5

Table 1. Coverage (C), precision (P), recall (R), and F_1-measure (F) of WSD methods on the Senseval 2 and 3 datasets, *polysemous words only*, excluding adverbs, *without using the first-sense heuristic*. The results are percentages.

the heuristic is actually a *supervised* baseline. Unfortunately, the heuristic on its own outperforms all existing unsupervised WSD methods.[6] Hence, the experimental results of most unsupervised WSD methods, including ours, can be drastically improved by frequently invoking the first-sense heuristic, even for randomly selected word occurrences. We, therefore, believe that unsupervised WSD methods should not be allowed to use the heuristic in evaluations.

Table 1 lists the results of our experiments. We follow common practice and exclude adverbs. We consider only polysemous words, i.e., we ignore words with only one possible meaning (trivial cases), which is why the results may appear to be lower than results published elsewhere; the first-sense heuristic is also not used. Since all six methods may fail to assign a sense to some word occurrences, we show results in terms of coverage (percentage of word occurrences assigned senses), precision (correctly assigned senses over total assigned senses), recall (correctly assigned senses over word occurrences to be assigned senses), and F_1 measure.[7] A reasonable upper bound is human interannotator agreement [11]. For fine-grained sense inventories, like WordNet's, interannotator agreement is between 67% and 80% [13]. A random baseline, assigning senses randomly with uniform probability, achieves approx. 20% and 14% accuracy on Senseval 2 and 3, respectively, counting both monosemous and polysemous words.

On the Senseval 2 dataset, the coverage of our ILP method (with both measures, with and without sense pruning) was significantly higher than that of SAN

[6] When both monosemous and polysemous words are considered, the first-sense heuristic achieves 63.7% and 61.3% accuracy on the Senseval 2 and 3 datasets, respectively, with 100% coverage. At 100% coverage, precision and recall are equal to accuracy.

[7] We do not assign a sense to a word occurrence when the relatedness of all of its senses to all the senses of all the other word occurrences is zero.

and PR. In terms of F_1-measure, ILP-SR-FULL performed overall better than SAN and PR, outperforming SAN by a wide margin. Our method performed worse with the PMI-based measure (ILP-PMI-FULL) than with *SR* on the Senseval 2 dataset, though it still outperformed SAN, but not PR. The pruned versions (ILP-SR-PRUN, ILP-PMI-PRUN) performed as well as or better than the corresponding unpruned ones (ILP-SR-FULL, ILP-PMI-FULL), indicating that sense pruning successfully managed to remove mostly irrelevant senses. Sense pruning also leads to considerable improvements in execution time. The average execuation time per sentence (collectively on both datasets) was 82.81, 23.45, 81.46, 17.40 seconds for ILP-SR-FULL, ILP-SR-PRUN, ILP-PMI-FULL, ILP-PMI-PRUN, respectively. The corresponding times for SAN and PR were 101.38 and 91.92, i.e., our ILP methods are in practice faster than the implementations of SAN and PR we had available, even though the computational complexity of SAN and PR is polynomial.[8]

On the Senseval 3 dataset, the coverage of all ILP methods remains very high, with a small decline when the PMI-based measure is used. The F_1 scores of the ILP methods are now lower, compared to their respective scores in Senseval 2; this is due to the larger average polysemy of Senseval 3 (8.41 vs. 6.48 for polysemous words). Surprisingly, however, SAN and PR now perform better than in Senseval 2; and PR outperforms our ILP methods , with the overall difference between SAN and ILP-SR-FULL now being negligible. We can only speculate at this point that the improved performance of SAN and PR may be due to the higher polysemy of Senseval 3, which allows them to construct larger graphs, which in turn allows them to assign more reliable rank or activation scores to the nodes (senses). The coverage of SAN and PR is also now much higher, which may indicate that as their graphs become larger, it becomes easier for SAN and PR to construct connected graphs; both methods require a connected graph, in order to rank the nodes or spread the activation, respectively. Also, the pruned ILP methods now perform worse than the corresponding unpruned ones, indicating that sense pruning is less successful in discarding irrelevant senses as polysemy increases.

We aim to investigate the differences between the Senseval 2 and 3 results further in future work. For the moment, we conclude that our ILP approach seems to work better with lower polysemy. We believe, though, that our experiments against SAN and PR already show the potential of our ILP model.

5 Conclusions

We presented an ILP model of WSD, which can be used with off-the-shelf solvers and any sense relatedness measure. We experimented with *SR* and a hybrid PMI-based measure on the Senseval 2 and 3 datasets, against two well-known methods based on PageRank (PR) and Spreading Activation Networks (SAN). Overall, our ILP model performed better with *SR*. With that measure, it performed better than both PR and SAN on the Senseval 2 dataset, outperforming SAN by a wide

[8] The complexity of SAN is $O(n^2 \cdot k^{2l+3})$, and PR's is $O(n^2 \cdot k^{\frac{3}{2}l+3})$, where k is the maximum branching factor of the hierarchical thesaurus, l its height, and n the number of word occurrences to be disambiguated [19].

margin. By contrast, PR performed much better than our ILP methods on the Senseval 3 dataset, and the difference between SAN and our best ILP method was negligible. In practice, our ILP methods run faster than the PR and SAN implementations we had available. We hope that our ILP model will prove useful to others who may wish to experiment with different relatedness measures.

References

1. Agirre, E., Lopez de Lacalle, O.: Supervised domain adaption for word sense disambiguation. In: EACL (2009)
2. Agirre, E., Soroa, A.: Personalizing PageRank for word sense disambiguation. In: EACL (2009)
3. Florian, R., Cucerzan, S., Schafer, C., Yarowsky, D.: Combining classifiers for word sense disambiguation. Natural Language Engineering 8(4), 327–341 (2002)
4. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In: IJCAI (2007)
5. Lesk, M.: Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone. In: SIGDOC (1986)
6. Manning, C., Schutze, H.: Foundations of Statistical NLP. MIT Press (2000)
7. Mihalcea, R., Csomai, A.: SenseLearner: Word sense disambiguation for all words in unrestricted text. In: ACL (2005)
8. Mihalcea, R., Tarau, P., Figa, E.: PageRank on semantic networks with application to word sense disambiguation. In: COLING (2004)
9. Montoyo, A., Suarez, A., Rigau, G., Palomar, M.: Combining knowledge- and corpus-based word-sense-disambiguation methods. JAIR 23, 299–330 (2005)
10. Navigli, R.: Online word sense disambiguation with structural semantic interconnections. In: EACL (2006)
11. Navigli, R.: Word sense disambiguation: A survey. ACM Computing Surveys 41(2), 10:1–10:69 (2009)
12. Navigli, R., Lapata, M.: Graph connectivity measures for unsupervised word sense disambiguation. In: IJCAI. pp. 1683–1688 (2007)
13. Palmer, M., Dang, H., Fellbaum, C.: Making fine-grained and coarse-grained sense distinctions, both manually and automatically. NLE 13(2), 137–163 (2007)
14. Ponzetto, S., Strube, M.: Knowledge derived from Wikipedia for computing semantic relatedness. J. of Artificial Intelligence Research 30, 181–212 (2007)
15. Quillian, R.: The teachable language comprehender: a simulation program and theory of language. Communications of ACM 12(8), 459–476 (1969)
16. Resnik, P.: Using inform. content to evaluate semantic similarity. In: IJCAI (1995)
17. Schütze, H.: Automatic word sense discrimination. Computational Linguistics 24(1), 97–123 (1998)
18. Sinha, R., Mihalcea, R.: Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. In: IEEE ICSC (2007)
19. Tsatsaronis, G., Varlamis, I., Nørvåg, K.: An experimental study on unsupervised graph-based word sense disambiguation. In: CICLing (2010)
20. Tsatsaronis, G., Varlamis, I., Vazirgiannis, M.: Text relatedness based on a word thesaurus. JAIR 37, 1–39 (2010)
21. Tsatsaronis, G., Vazirgiannis, M., Androutsopoulos, I.: Word sense disambiguation with spreading activation networks generated from thesauri. In: IJCAI (2007)
22. Turney, M.: Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In: ECML (2001)