Google & WordNet based Word Sense Disambiguation

Ioannis P. Klapaftis GIANNIS@CS.YORK.AC.UK
Suresh Manandhar SURESH@CS.YORK.AC.UK
Department of Computer Science, The University of York, York, UK, YO10 5DD

Abstract

This paper presents an unsupervised methodology for automatic disambiguation of noun terms found in domain specific unrestricted corpora. This method extends approaches of Fragos (Fragos et al., 2003) and others that use the WordNet (Miller, 1998) database in order to resolve semantic ambiguity. The method is evaluated by disambiguating the noun collection of SemCor 2.0. Parameter adjustment was performed using a supervised technique that significantly increased the accuracy.

1. Introduction

Gruber (Gruber, 1993) defined ontology as a formal, explicit specification of a shared conceptualisation, while Maedche and Staab (Maedche & Staab, 2001) defined the task of ontology learning as the construction of a controlled vocabulary with explicitly defined concepts or instances, explicitly defined relations between them and machine-processable semantics. Word sense disambiguation (WSD), which is part of semantic interpretation, is the process of associating an appropriate and unique concept identifier with each term found in documents (Missikoff et al., 2002). That way it is possible to automatically and explicitly define instances and have machine processable semantics.

In many ontology learning systems such as Dogma (Reinberger et al., 2003), Text-To-Onto (Maedche & Staab, 2000b), Ontolearn (Missikoff et al., 2002) the WSD task is an integral part of ontology learning. However, in the first two, WSD is carried out manually.

The methodology presented here uses Google to find contextually relevant terms, which in turn help in assigning the correct WordNet sense to each term under disambiguation.

Appearing in *Proceedings of the 22ndICML Workshop on Learning & Extending Ontologies*. Bonn, Germany, 2005.

2. Existing approaches to WSD for Ontology Learning

MO'K (Bisson et al., 2000) and ASIUM (Faure et al., 1998), (Faure & Poibeau, 2000) are tools that support the development of conceptual hierarchies from parsed corpora. Their papers do not describe how extracted ontology members are semantically disambiguated. Dogma (Reinberger et al., 2003) and Text-To-Onto (Maedche & Staab, 2000b), (Maedche & Staab, 2000a) are two other systems that aim to build a domain specific ontology as well. Their papers state that word sense disambiguation of ontology members is manually done by domain experts. Thus both of these approaches are semi-automatic.

OntoLearn (Navigli & Velardi, 2002), (Navigli et al., 2003) is a workbench aimed at the construction of domain specific ontologies. The Ontolearn workbench provides a semi-automatic method for WSD of compound nouns. They do not provide a method for WSD of single terms. We describe their approach briefly as it has some similarity with the proposed method.

Ontolearn defines the semantic interpretation $S(t)$ of a compound term $t = w_1 w_2 \dots w_n$, where each w_k is a noun term, as the union of the disambiguated WordNet synsets s_k for each w_k.

$$S(t) = \bigcup_{1}^{n} s_k, s_k \in Synset(w_k), w_k \in t.$$

For each w_k and synset s_k of w_k, their procedure creates a mini-semantic net, starting from s_k, using the WordNet relations hypernymy, hyponymy, synonymy, meronymy and others. For any pair w_k, w_{k+1} alternative pairs of mini-semantic nets are intersected. For each intersection i, *common semantic patterns* are identified, evaluated and stored in a vector. The best score vector determines the sense for w_{k+1}. The com-

mon semantic patterns are handcrafted. Their procedure requires the first word to be manually disambiguated.

3. Disambiguation Procedure

Our approach aims to automatically find the correct sense for each ontology candidate term by using WordNet and Google. It overcomes two deficiencies of Ontolearn (Navigli & Velardi, 2002): disambiguation only of compound nouns and the manual disambiguation of the first word in a compound term. Our experiments show that the proposed method achieves a high accuracy without requiring any manual intervention.

3.1. Stating the hypotheses

Agirre (Agirre et al., 2000), created topic signatures for automatic WSD. A topic signature is a list of words that appear within a window along with their respective frequencies (calculated from corpora), which is topically related to the word under disambiguation. Their hypothesis can be stated as follows:

Hypothesis 1: The meaning of a word can be discovered from words around it.

In our approach, *hypothesis 1* is further strengthened by assuming that surrounding words that determine the sense of a word have a semantic relationship with the word under disambiguation. These relations can be used, in order to determine the meaning of the word. This is stated as *hypothesis 2*.

Hypothesis 2: Semantically related words that impose constraints on each other are expected to be topically related.

The disambiguation procedure is based on the use of the following WordNet relations: hypernymy/hyponymy, meronymy, synonymy and holonymy. Consequently an essential assumption of our approach is the following.

Assumption: WordNet contains all the words needed to disambiguate ontology members.

3.2. Algorithm

Assume that we have a collection of domain specific documents C and a list of words W extracted from C. Our approach intends to disambiguate each term t_i found in W. Let S be the sentence that contains term t_i. The output of our method will be the discovery of the appropriate WordNet synset for each t_i.

If term t_i is single (consisting of a single token) then its semantic interpretation $S(t_i)$ is defined as:

$$S(t_i) = s_k, where\ t_i \in W, s_k \in Synset(t_i)$$

If term t_i is compound (consisting of more than one tokens) then its semantic interpretation is defined as:

$$S(t_i) = \bigcup_{i}^{n} s_k, where\ s_k \in Synset(w_k), w_k \in t_i, t_i \in W$$

The disambiguation of a single term t_i consists of the following steps:

Step1

a) Send the sentence S containing term t_i to Google and retrieve the first four documents.

b) Tokenise and perform POS tagging on all the retrieved texts. Let this annotated text be U.

c) Retrieve from WordNet all synsets s_k of term t_i.

Step2

a) For a synset s_k of t_i retrieve hypernym terms at a distance equal to 3 and store them in a list (hypernym list).

b) Repeat the same process for hyponym, synonym, meronym, and holonym terms to generate hyponym list, synonym list, meronym list and holonym list.

c) For each retrieved term in each list calculate its frequency in U and store it.

d) Calculate lists scores (as given in section 3.3).

e) Calculate normalized Total Sense Score. (TSS) (as given in section 3.3)

f) Repeat the same process for all the sense synsets of term t_i.

Step3

a) Compare the TSS of each sense.

b) Find the maximum TSS and assign the corresponding synset as the appropriate meaning of t_i.

The disambiguation of a compound term t_i involves the application of the above steps for each w_k, when $t_i = w_0 w_1 \dots w_n$, $k = 0 \dots n$.

3.3. Calculation of frequencies and weights

The purpose of this section is to present the method used to calculate the semantic relatedness between a sense of a term and a document collection, which in our case is denoted as *total sense score* (TSS). TSS is based on the calculation of the frequencies of hypernym, hyponym, meronym, holonym and synonym terms in the document collection retrieved from Google.

In steps 2.a, 2.b and 2.c we retrieved five lists of terms, the hypernym list, the hyponym list, the meronym list, the holonym list and the synonym list for a sense s_k of a term t_i. These lists contain the retrieved terms along with their respective frequencies in U. Essentially these lists contain terms that are semantically related to the word

under disambiguation. In our approach we make use of the hypothesis that these words are expected to be both semantically and topically related. That means that the set of words contained in the five lists is or contains a very small subset of the topic signature for sense s_k of term t_i.

If we compare a topic signature with a set containing the terms of our five lists, then we could say that a topic signature contains topically related terms and consequently captures a number of semantically related terms. On the other hand, based on our hypothesis, our set contains semantically related terms and consequently captures a number of topically related terms. Additionally while WSD using topic signatures includes the search of surrounding terms into the topic signatures, the application of our approach includes the search of the lists terms in the document collection.

Topic signatures were used to overcome three major shortcomings of WordNet. According to Aggire (Agirre et al., 2000) these include the lack of explicit links among semantic variant concepts with different part of speech, the lack of explicit relations between topically related concepts and the fact that many sense distinctions are unclear. We aim to overcome the first two fore mentioned problems by calculating the frequencies of lists terms in the document collection. Thus our intuition is that frequencies of terms that share semantic and topical relations with a term under disambiguation are indicative of the sense of that term. Consequently TSS increases as frequencies increase, which in turn increases our confidence on a correct disambiguation.

For example consider the following paragraph from (Gruber, 1993):

> Several technical problems stand in the way of shared, reusable **knowledge**-based software. Like conventional applications, **knowledge**-based systems are based on heterogeneous hardware platforms, programming languages, and network protocols. However, **knowledge**-based systems pose special requirements for interoperability. Such systems operate on and communicate using statements in a formal **knowledge** representation. They ask queries and give answers. They take **background knowledge as an input**.

Let us assume that we need to disambiguate the noun *input*. WordNet provides two senses for *input*.

> 1.input signal, input – (signal going...)
> 2.stimulation, stimulus, stimulant, input – (any stimulating...)

The correct sense of *input* in that case is the second one. By examining the above text it is possible to note that noun *knowledge* appears 5 times. *Knowledge* is a hypernym for the second sense of *input*. Thus its frequent appearance provides us with strong evidence that the second sense of *input* might be the correct one.

According to all the above we are able to define the following three scores:

3.3.1. Score of hypernym list.

Let A be the score of hypernym list:

$$A = \frac{\sum_{1}^{3} a[i] * HyperFrequencyDistance[i]}{\sum_{1}^{3} HyperFrequencyDistance[i]}$$

$HyperFrequencyDistance$ is a vector that contains the sum of the frequencies of the hypernym terms at a given distance i. The weight vector a contains the weights applied to each distance i. We have chosen $a = [1, 1/2, 1/4]$. The denominator term in A is a normalising term which ensures that $0 \leq A \leq 1$.

3.3.2. Score of hyponym list.

Let B be the score of hyponym list:

$$B = \frac{\sum_{1}^{3} a[i] * HypoFrequencyDistance[i]}{\sum_{1}^{3} HypoFrequencyDistance[i]}$$

$HypoFrequencyDistance$ is a vector that contains the sum of the frequencies of the hyponym terms at a given distance i. The weight is the same as in score A and $B \leq 1$.

3.3.3. Score of meronym, holonym and synonym list.

The scores of meronym, holonym and synonym list can be computed in one score C.

$$C = \frac{MeroFreq * c_1 + HoloFreq * c_2 + SynoFreq * c_3}{MeroFreq + HoloFreq + SynoFreq}$$

The above frequencies represent the sum of frequencies of the meronym, synonym and holonym terms. The weights $c1, c2, c3$ will be produced empirically during the evaluation and $C \leq 1$. Consequently the Total Sense Score (TSS) will be the following:

$$TSS = \frac{A + B + C}{3}$$

Since $A \leq 1$, $B \leq 1$, $C \leq 1$, it is obvious that $TSS \leq 1$.

4. Evaluation

Our disambiguation method was evaluated using the first 10 Semcor 2.0 files (Lande et al., 1998). Semcor files are manually disambiguated text corpora using WordNet senses. Experiments provided us with the values of the weights c_1, c_2, c_3 (Section 3.3) as $c_1 = 0.9, c_2 = 0.9, c_3 = 0.2$.

The evaluation procedure consists of the following steps:

- Get the i unannotated sentence of k file of Semcor 2.0.

- Begin disambiguation of sentence i.
- Compare system output with the Semcor 2.0 i annotated sentence of k file.
- Repeat the above steps for:
  $i = 1 \ldots NumberOfSentences(File_k)$, $k = 1 \ldots 10$

Files contained 5463 nouns. Our system managed to disambiguate 5153 out of 5463 nouns (94.32%). The accuracy of our approach was 58.90%, which means that our system disambiguated correctly 3218 out of 5463 nouns. The following table depicts the results of our system.

File	Nouns	Disambiguated	Accuracy(%)
br-a01	573	540	63.87
br-a02	611	590	60.22
br-a11	582	540	63.05
br-a12	570	545	57.89
br-a13	575	535	62.22
br-a14	542	516	51.10
br-a15	535	509	62.80
br-b13	505	472	52.67
br-b20	458	419	54.58
br-c01	512	487	58.59
Total	5463	5153	58.90

Table 1. Results from the first 10 files of Brown 1 Corpus

5. Parameters Adjustment using Semcor

In section 3.3 we mentioned that TSS increases as frequencies increase, which in turn increases our confidence on a correct disambiguation. Consequently WSD based on a high TSS is more likely to be correct than WSD based on a low TSS.

Based on this idea, we introduced a supervised technique, in order to improve the accuracy of our approach. This technique involves the calculation of a threshold that mirrors our certainty on a correct disambiguation. More precisely if scores of winning senses are below our threshold, then corresponding terms will be regarded as ambiguous and their semantic interpretation will be arbitrarily assigned to the most frequent sense of WordNet. Naturally scores of winning senses above our threshold, will not be regarded as ambiguous.

In order to calculate the particular threshold we created a diagram for the first 90 sentences of Semcor, showing the normalized distribution of incorrect disambiguations (number of incorrect disambiguations divided by the total number of disambiguations) on 20 equal intervals from 0 to 1.

What we were expecting was a diagram with a uniform behavior, which means that as score increases, the number of incorrect disambiguations decreases and the inverse. The behavior of *incorrect disambiguations* is not uniform and has slight anomalies. This is possibly caused by the fact, that we have used a small number of sentences (90) to generate the diagram.

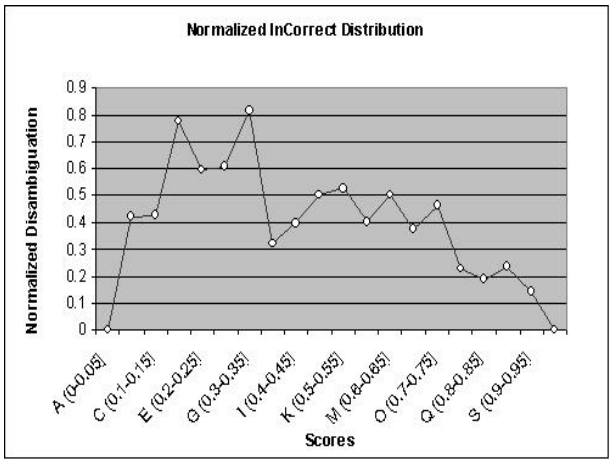

Figure 1. Distribution of incorrect disambiguations

However what we can clearly see is that scores from 0.05 to 0.35 cause a greater number of incorrect disambiguations than scores from 0.35 to 0.95. Consequently our threshold should be between 0.05 and 0.35. A very high point in the diagram exists when scores are between 0.15 and 0.20, while the highest point is found when scores are between 0.30 and 0.35. Thus we have set our threshold to 0.35 (highest point in the diagram).

The application of the threshold to the first 90 sentences increased the accuracy of our approach by 9%. We applied the same threshold to the rest of the evaluation sentences, adding to our unsupervised approach a supervised technique. The accuracy of our approach increased significantly. 383 more nouns were disambiguated correctly. Totally 3601 out of 5463 nouns (65.91%) had their sense correctly predicted.

6. Discussion and Conclusion

We have introduced an unsupervised word sense disambiguation approach, which assigns to each term under disambiguation the appropriate WordNet synset and has potential applications for ontology learning. In essence our algorithm is based on two hypotheses (section 3.1). Evaluation was performed on the first ten files of Semcor 2.0. Overall our approach achieved an accuracy of 58.90%. The application of a supervised technique significantly increased the overall accuracy of our system (65.91%).

In section 3.3 we mentioned three major shortcomings of WordNet. We aimed to overcome the first two by calculating frequencies of semantically related terms in the document collection. Our intuition was that these frequencies are indicative of the sense of the term under disambiguation. Our approach managed to handle the fore mentioned shortcomings and achieve a satisfactory result. However these problems were not solved and consequently future work in WSD would include the development of methods to retrieve and use terms that are not explicitly related in WordNet.

The third shortcoming i.e the fine-grainedness of WordNet's sense distinctions is inevitably inherited to our approach from the definition of semantic interpretation of a single or compound term t_i (section 3.2). To deal with this problem the desired level of sense distinction must be defined. However this is unclear and possibly depends on the domain.

Our approach is based on Google returned results. Thus we expect returned documents to be domain consistent. Experiments showed that we should restrict the number of retrieved results to 4 to avoid getting too much noise. However, when input sentences were quite small, the retrieved pages contained a lot of noise. That is why in files br-a14, br-b13 accuracy is quite low. The solution to that problem would be to perform WSD for a set of sentences based on Yarowsky's *one sense per discourse* (Yarowsky, 1995) i.e. if a target term appears more than once in input sentences, then its interpretation can and will be one WordNet sense, because according to Yarowsky (Yarowsky, 1995) the sense of a target word is highly consistent within any given document.

Acknowledgments

The first author is grateful to the *General Michael Arnaoutis* charitable foundation for its financial support and encouragement. Finally we are more than grateful to the research associate of our department, Dimitrios Kolovos, for his valuable ideas, especially relating to parameter adjustment, and help during the implementation and evaluation of our methodology.

References

Agirre, E., Ansa, O., Hovy, E., & Martinez, D. (2000). Enriching very large ontologies using the www. *Proceedings of the Ontology Learning Workshop, ECAI.*

Bisson, G., Nedellec, C., & Canamero, D. (2000). Designing clustering methods for ontology building: The MoK workbench. *Proceedings of the First Workshop on Ontology Learning (OL-2000) in conjunction with the 14th European Conference on Artificial Intelligence.*

Faure, D., Nedellec, C., & Rouveirol, C. (1998). *Acquisition of semantic knowledge using machine learning methods: The system ASIUM.* (Technical Report). Laboratoire de Recherche en Informatique, Inference and Learning Group, Universite Paris, Sud.

Faure, D., & Poibeau, T. (2000). First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX. *Proceedings of the ECAI2000 Workshop Ontology Learning, 2000..*

Fragos, K., Maistros, Y., & Skourlas., C. (2003). Word sense disambiguation using wordnet relations. *First Balkan Conference in Informatics, Thessaloniki.*

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5, (2): 199-220.*

Lande, S., Leacock, C., & Tengi, R. (1998). Wordnet, an electronic lexical database. *MIT Press, Cambridge MA 199-216.*

Maedche, A., & Staab, S. (2000a). Discovering conceptual relations from text. *ECAI-2000 -European Conference on Artificial Intelligence, Proceedings of the 13th European Conference on Artificial Intelligence, pages 321325. IOS Press, Amsterdam..*

Maedche, A., & Staab, S. (2000b). Mining ontologies from text. *Proc. 12th Intl Conf. Knowledge Eng. and Knowledge Management, Lecture Notes in Computer Science, vol. 1937, Springer Verlag, New York, pp. 189202..*

Maedche, A., & Staab, S. (2001). Ontologies learning for the semantic web. *IEEE Intelligent Systems and Their Applications, Vol. 16 (2), 72-79.*

Miller, G. (1998). Wordnet: A lexical database for english. *Comm. ACM, vol. 38, no. 11*, pp. 3941.

Missikoff, M., Navigli, R., & Velardi, P. (2002). Integrated approach to web ontology learning and engineering. *IEEE Computer Vol 35(11) pp.60-63.*

Navigli, R., & Velardi, P. (2002). Semantic interpretation of terminological strings. *Proc. 4th Intl Conf. Terminology and Knowledge Engineering (TKE 2002), New York* (pp. 325–353.). Springer-Verlag.

Navigli, R., Velardi, P., & Gangemi, A. (2003). Ontology learning and its application to automated terminology translation. *IEEE Intelligent Systems, v.18 n.1, p.22-31.*

Reinberger, M.-L., Spyns, P., Daelemans, W., & Meersman, R. (2003). Mining for lexons: applying unsupervised learning methods to create ontology bases. *In Proceedings ODBASE03, Springer-Verlag..*

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, Cambridge, Massachusetts* (pp. 189–196).