Web-based Variant of the Lesk Approach to Word Sense Disambiguation

Miguel Ángel Ríos Gaona,[1] Alexander Gelbukh[2]

Center for Computing Research,
National Polytechnic Institute,
Mexico City, Mexico
[1] mriosb02@sagitario.cic.ipn.mx,
[2] www.gelbukh.com

Sivaji Bandyopadhyay

Computer Science & Engineering Department,
Jadavpur University
Kolkata 700 032
India
sivaji_cse_ju@yahoo.com

***Abstract*— Word Sense Disambiguation (WSD) is the task of selecting the meaning of a word based on the context in which the word occurs. The principal statistical WSD approaches are supervised and unsupervised learning. The Lesk method is an example of unsupervised disambiguation. We present a measure for sense assignment useful for the simple Lesk algorithm. We use word co-occurrences of the gloss and the context, which is statistical information retrieved from the Web. In the SemCor data our method always gives an answer. On the Senseval 2 data, our variant of the Lesk method outperformed some other Lesk-based methods.**

***Index Terms*—Natural Language Processing, Word Sense Disambiguation, Unsupervised disambiguation.**

1 Introduction

Word Sense Disambiguation (WSD) is the task of selecting the most appropriate meaning for a polysemous word, based on the context in which it occurs. For example, in the phrase *The bank down the street was robbed*, the word bank means a financial institution, while in *The city is on the Western bank of Jordan*, this word refers to the shore of a river. WSD is an internal task in the natural language processing (NLP) chain [25]. It is used in many applications such as machine translation and information retrieval. The problem of word sense disambiguation has been described as AI-complete, that is, a problem which can be solved only by first resolving all the difficult problems in artificial intelligence (AI), such as the representation of common sense and encyclopedic knowledge.

To address this task, different methods have been used, with various degrees of success. These methods can be classified depending on the type of knowledge they use to accomplish the task. The main statistical approaches to the WSD task are supervised and unsupervised disambiguation.

Supervised methods use a labeled training set to solve the task. They have been shown to be the most efficient ones [22]. However, the lack of large sense tagged corpora limits this kind of methods, and it is difficult and expensive to create such corpora manually.

Unsupervised methods are based on unlabeled corpora. This resolves the knowledge acquisition bottleneck, at the cost of low accuracy. These approaches often do not use any learning process; they only rely on a lexical resource, like WordNet [17], to carry out the WSD task.

An example of an unsupervised method is the original Lesk algorithm (OL) [15], which disambiguates polysemous words in (shorts) phrases. The definition, or gloss (from a dictionary), of each sense of an ambiguous word in a phrase is compared to the glosses of every other word in the phrase. Basically, the algorithm selects the set of senses such that their glosses have the largest number of words in common.

We show that to tackle the problem of knowledge acquisition bottleneck in supervised methods, the web can be used as a lexical resource.

The Web has become a source of data for NLP, and WSD is no an exception. The web is immense, free, and available at a mouse click. It contains hundreds of billions of words of text and can be used for all manner of language research. The simplest use of the web for language processing is spell-checking: is it spelled *speculater* or *speculator*? Google count for the former is 67 (usefully suggesting that the latter might have been intended) and for the latter, 82,000; the question is answered.

Web as a corpus for NLP research [25] has been used with success in many areas such as question answering [7], machine translation [12], anaphora resolution [6], paraphrasing [4], detection of malapropisms [3], collocation testing [5], translation [8], [9], and text representation [19], and dictionary building [10]. Many methods use the web to automatically generate sense tagged corpora [1], [16], [23].

This paper proposed a measure for sense number score assignment based on web statistics. It uses word co-occurrences of the gloss and the context, which is statistical information retrieved from web, instead of gloss overlaps.

The paper is structured as follows. An overview of the related work is presented in Section 2. Section 3 describes the proposed measure. Section 4 shows our experiments: first over the SemCor corpus and then a comparison with previous

978-0-7695-3933-1/09 $26.00 © 2009 IEEE
DOI 10.1109/MICAI.2009.41

results over the Senseval 2 corpus. Finally conclusions are drawn in Section 5.

2 Related work

Senseval, started in 1998 [13], tied to the evaluation of WSD systems, producing a set of benchmarks for evaluating WSD system performance, to establish the viability of WSD as a separately evaluable NLP task.

In the past versions of Senseval, exercises that were variants of the Lesk approach were considered as baseline approaches. In Senseval 1, most of the systems for disambiguating English words were outperformed by a Lesk variant, used as baseline. On the other hand, at Senseval 2, Lesk baselines were outperformed by most of the systems in the lexical sample task.

The Lesk-based baselines outperform the baseline that uses simpler algorithms such as random sense assignment, or an algorithm that always chooses the sense which has most training-corpus instances.

The simplified Lesk (SL) algorithm [13] chooses the sense of an ambiguous word *w* such that its gloss *g* has the greatest number of words in common with other words (the context of *w*) around the given word *w*:

```
For each sense s of w do
  weight(s) = sim(c,g(s))
s = argmax weight(s)
```

Here *c* is the context of the word *w* (in the simplest case, just a bag of words within a certain distance from *w*) and *g*(*s*) is the gloss associated with the sense *s*.

The Lesk-plus method [13] also considers a learning process, so it can be compared with supervised systems. For each word in the sentence containing the test item, it tests whether the word occurs in the dictionary entry or corpus instances for each candidate sense. For weighting of the sentences it uses the inverse document frequency (IDF) of a word, computed as log(*p*(*w*)), where *p*(*w*) is estimated as the fraction of dictionary "documents"—definitions or examples—which contain the word. Lesk-plus method does not explicitly represent the relative corpus frequencies of sense tags. Instead, it favors common tags because they have larger context sets, and an arbitrary word in a test-corpus sentence is more likely to occur in the context set of a more common training-corpus sense tag.

The original Lesk algorithm relies on glosses found in traditional dictionaries such as Oxford Advance Learner's dictionary. Banerjee and Pedersen [2] proposed a variant of the Lesk algorithm to take the advantage of the highly interconnected set of relations among synonyms that WordNet offers. This variant takes as back-off the glosses of words that are related to the words to be disambiguated. This back-off provides a richer source of information and improves accuracy. It outperforms the baseline methods in the Senseval 2 exercise.

Vasilescu *et al*. [24] proposed a set of different variants to the Lesk approach. The first variant, the score assigned to a candidate sense is the number of overlaps between the BOW of that sense and the BOW of the context. A second variant, called WHG (for weighted) also takes into account the length of the description for a given sense. According to Lesk, long descriptions can produce more overlaps than short ones, and thus dominate the decision making process.

Another variant multiplied the number of overlaps for a given candidate sense by the inverse of the logarithm of the description length for this sense. Other variant for weighting metrics were also proposed, taking into account the distance between a word in the context and the target word, or the frequency of the context word in the language, but that did not bring any significant difference.

3 Proposed measure

In Statistical NLP, one commonly receives as a corpus a certain amount of data from a certain domain of interest, without having any say in how it is constructed. In such cases, having more training data is normally more useful than any concerns of balance, and one should simply use all the text that is available. The problem of data sparseness, which is common for much corpus-based work, is especially severe for work in WSD. First, enormous amounts of text are required to ensure that all senses of a polysemous word are represented, given the vast disparity in frequency among senses.

We augment the Lesk approach with a measure for sense number assignment. The measure is based on the hypothesis of the high relationship between the gloss of a sense and the context of the word. We measure this relationship by finding the frequencies of co-occurrences between the gloss and the context, using the web as a corpus. We use the new measure applied to the Simple Lesk algorithm as follows:

```
For each word w to be tagged
 For each sense s of w
  g = gloss of sense s (bag of words)
  e = example of sense s (bag of words)
  d = g ∪ e
  dc = d ∪ c
  $f_g$  = web frequency of d
  $f_{gc}$ = web frequency of dc
  weight(s) = $f_{gc}/f_g$
s = argmax weight (s)
```

The web frequency is measured by a query to a web search engine. The weight is the probability of seeing the gloss of a sense in the context of the given word occurrence. The method chooses the sense which maximizes the weight.

If various senses have the same weight, then the sense is chosen by a back-off heuristic.

4 EXPERIMENTAL RESULTS

In this section firstly we show a brief description of the datasets used, second the experimental setting of the proposed measure and finally a comparison with previous results.

4.1 Data set

SemCor is a textual corpus in which words are syntactically and semantically tagged. The texts included in SemCor were extracted from the Brown corpus and then linked to senses in the WordNet lexicon. All the words in the corpus have been syntactically tagged using Brill's part of speech tagger; the semantically tagging was done manually for all the nouns, verbs, adjectives and adverbs, each of these words being associated with its correspondent WordNet sense. We show above an example of an entry in the SemCor corpus.

```
<wf cmd=done pos=VB lemma=say wnsn=1
lexsn=2:32:00::>said</wf>
```

The Senseval dataset consists of 4,328 instances each of which contains a sentence with a single target word to be disambiguated, and one or two surrounding sentences that provide additional context.

A task in Senseval consists of three types of data: 1) A sense inventory of word-to-sense mappings, with possibly extra information to explain, define, or distinguish the senses (e.g., WordNet); 2) A corpus of manually tagged text or samples of text that acts as the Gold Standard, and that is split into an optional training corpus and test corpus; and 3) An optional sense hierarchy or sense grouping to allow for fine or coarse grained sense distinctions to be used in scoring. The next XML is an example of an entry in Senseval.

```
<instance
    id="9:0@16@wsj/24/wsj_2444@wsj@en@on"
    docsrc="wsj">
<context>
Once metropolitan ...<head> asking </head> ...
</context>
</instance>
```

Senseval has two variants of the WSD task:

All words task participating systems have to disambiguate all words (open-class words) in a set of text, and Lexical sample task, first a sample of words is selected. Then for each sample word, a number of corpus instances are selected.

4.2 Experimental setting

In our preliminary experiments we aimed at the all words WSD task. For evaluation we used a subset of the first two tagged files of SemCor 1.6: the files br-a01 and br-a02. We used WordNet 2.1 as a sense repository. WordNet is a lexical database where each unique meaning of a word is represented by a synonym set or synset. Each synset has a gloss that defines the concept that it represents. For example, the words *car, auto, automobile*, and *motorcar* constitute a single synset that has the following gloss: *four-wheel motor vehicle, usually propelled by an internal combustion engine*. Many glosses have examples of usages associated with them, such as "*he needs a car to get to work*."

Context is the only means to identify the meaning of a polysemous word. Therefore, all work on WSD relies on the context of the target word to provide information to be used for its disambiguation. Most disambiguation work uses the local context of a word occurrence as a primary information source for WSD. Local or "micro" context is generally considered to be some small window of words surrounding a word occurrence in a text or discourse, from a few words of context to the entire sentence in which the target word appears.

Context is very often regarded as all words or characters falling within some window of the target, with no regard for distance, syntactic, or other relations. Yarowsky [27] examines different windows of micro-context, including 1-contexts, *k*-contexts, and words pairs at offsets –1 and –2; –1 and +1; +1 and +2, and sorts them using a log-likelihood ratio to find the most reliable evidence for disambiguation. Yarowsky makes the observation that the optimal value of k varies with the kind of ambiguity: he suggests that local ambiguities need only a window of $k = 3$ or 4.

We use the bag of words approach: here, context is considered as words in some window surrounding the target word, regarded as a group without consideration for their relationships to the target in terms of distance, grammatical relations. We take a symmetric window of ±3 words around the target word an optimal value to local ambiguities.

The web counts were collected using the Google[1] search engine. To construct the queries first we tokenize the sentence, then the target word is replaced by the gloss, and then we query the search engine with the obtained text string.

When our method cannot choose a sense number in the argmax function (e.g. two senses have the same weight), the sense is chosen randomly from the set of the top senses (i.e., those with the same maximal weight); in the sequel we refer to this as random top weight back-off.

We used precision and recall to score the system, although the metrics are not completely analogous to Information Retrieval evaluation. Recall (percentage of right answers on all instances in the test set) is the basic measurement of accuracy in this task, because it shows how many correct disambiguation results the system achieved overall. Precision (percentage of right answers in the set of answered instances) favors systems that are very accurate if only on a small subset of cases that the system chose to give answers to.

4.3 Comparison with other methods

Resnik and Yarowsky [20] have shown that it is difficult to compare WSD methods. The distinctions that make comparing

[1] http://www.google.com

methods difficult reside in the approach considered (supervised or unsupervised).

The result from the preliminary experiments over the SemCor subset obtained an accuracy of 47%. We only reported accuracy because of any word presented an equal weight of senses. If a system makes an assignment for every word, then precision and recall are the same, and can be called accuracy. Therefore the Web rarely presents data sparseness. Thus the method always gives an answer and it does not reach the back-off heuristic.

In Table 1 we present a comparison of the accuracy of our measure applied to the simple Lesk against variants of the original Lesk approach. This comparison was tested over the Senseval 2 data. The experiment had the same setting as the experiment over the SemCor subset.

TABLE 1
COMPARISON WITH SIMILAR METHODS

Method	Type	Back-off	Accuracy
Vasilescu *et al.* 2004	simplified	MFS	58%
Mihalcea and Tarau 2004	simplified	RS	47.27%
Our method	simplified	Random top senses	45%
Vasilescu *et al.* 2004	original	MFS	42%
Mihalcea and Tarau 2004	original	RS	35%
Banerjee and Pedersen 2002	original	Extended gloss overlap	31.7%

As it can bee seen from Table 1, the original Lesk (OL) algorithm method has a lower performance than the other ones and even than the baseline system. This observation is consistent with Litkowski [18] hypothesis that only about one third of the instances can rely on the Lesk-style information (gloss and example) in a disambiguation process. The simplified Lesk (SL) method, which only counts the overlaps between the description of a candidate sense and the words in the context, produces better results.

Our Lesk variant outperforms the OL of Banerjee and Pedersen [2] and the OL [21], [24] variants (back-off to random sense and most frequent sense). The SL of Mihalcea and Tarau [21] is better in performance than our method, with the help of the random sense heuristic. Finally, the SL of Vasilescu *et al.* [24] has the best accuracy. However, this method can be considered as a supervised method due to the most frequent sense heuristic (it is not clear what its performance would be with McCarthy *et al*. [14] unsupervised method for determining the predominant sense).

When a method can not make a judgment (i.e., no overlap between the gloss and the context in the simple Lesk) the judge is taken by the back-off heuristic. Most of these heuristics chose a random sense or uses information from a dictionary. So the most frequent sense is based in chose the first (or predominant) sense the heuristic assumes the availability of hand tagged data.

Therefore our method did not reach the back-off heuristic we present in Table 2 a comparison with the top three unsupervised methods of Senseval-2.

TABLE 2
COMPARISON WITH SENSEVAL-2 UNSUPERVISED METHODS

Method	Accuracy
Our method	45.0%
Senseval–First	40.2%
Senseval–Second	29.3%
Senseval–Third	24.7%
Original Lesk	18.3%

The Senseval–First, Senseval–Second, and Senseval–Third results are the top three most accurate fully automatic unsupervised systems in the Senseval-2 exercise. This class of systems can be compared to ours, since they require no human intervention and do not use any manually created training examples.

These results show that our approach was considerably more accurate than all of those systems. This method has the advantage of simplicity and the use of a very limited context window.

TABLE 3
COMPARISON WITH SEMEVAL-2007 UNSUPERVISED METHODS

Method	Accuracy
Radu Ion	52.7%
Davide Buscaldi	46.9%
Our method (Senseval-2)	45.0%
Sudip Kumar Naskar	40.2%

In Table 3 we present a comparison of our method (tested over Senseval-2) with the state of the art unsupervised systems in the SemEval-2007 [22]. Thus two of the methods outperform our method but the comparison is not so clear because our method was test over the Senseval-2 corpus. Thus we can see growing tendencies in the precision of the unsupervised approaches.

5 CONCLUSIONS

In this paper we used a variant of the Lesk algorithm for the WSD task. We proposed a new sense number weight measure based on web counts collected with a search engine.

We have shown that our variant outperforms some Lesk based methods and outperforms the top unsupervised methods of the Senseval-2 exercise. These results are significant because they are based on a very simple algorithm that relies on co-occurrences scores to the senses of a target word

We once more confirmed that the web could be used as a lexical resource for WSD.

In our future work we will explore the use of different context windows, as well as linguistically-motivated context windows (such as a syntactic unit) and test our method over the SemEval corpus.

ACKNOWLEDGEMENTS

Work was done under partial support of Mexican Government: SNI, PIFI-IPN, SIP-IPN 20091587, CONACYT 50206-H.

REFERENCES

[1] Agirre, E., Martinez, D. (2003): Exploring Automatic Word Sense Disambiguation with Decision Lists and the Web. In: Proc. of the COLING-2000.

[2] Banerjee, S. and Pedersen, T (2002). An adapted Lesk algorithm for word sense disambiguation using WordNet, CICLING 2002.

[3] Bolshakov, I.A., S.N. Galicia-Haro, A. Gelbukh: Detection and Correction of Malapropisms in Spanish by means of Internet Search. Lecture Notes in Artificial Intelligence N 3658, Springer, 2005, pp. 115–122.

[4] Bolshakov, I.A., A. Gelbukh: Synonymous Paraphrasing Using WordNet and Internet. Lecture Notes in Computer Science N 3136, Springer, 2004, pp. 312–323.

[5] Bolshakov, I.A., E.I. Bolshakova, A.P. Kotlyarov, A. Gelbukh: Various Criteria of Collocation Cohesion in Internet: Comparison of Resolving Power. CICLing 2008, Lecture Notes in Computer Science N 4919, Springer, 2008, 64-72

[6] Bunescu, R.(2003): Associative Anaphora Resolution: A Web-Based Approach. In: Proc. of the EACL-2003 Workshop on the Computational Treatment of Anaphora, Budapest, Hungary, April.

[7] Brill, E., Lin, J., Banko, M., Dumais, S., Ng, A.(2001): Data-intensive Question Answering. In: Proc. of the Tenth Text Retrieval Conference TREC-2001.

[8] Gelbukh, A., I.A. Bolshakov: Internet, a true friend of translator. International Journal of Translation, Vol. 15, No. 2, 2003, pp. 31–50.

[9] Gelbukh, A., I.A. Bolshakov: Internet, a true friend of translator: the Google wildcard operator. International Journal of Translation, Vol. 18, No. 1–2, 2006, pp. 41–48.

[10] Gelbukh, A., G. Sidorov, L. Chanona-Hernández. Compilation of a Spanish representative corpus. CICLing-2002, Lecture Notes in Computer Science N 2276, Springer, pp. 285–288.

[11] Gonzalo, J., Verdejo, F., Chugar, I.: The Web as a Resource for WSD. In: 1st MEANING Workshop, Spain.

[12] Grefenstette, G. (1999): The World Wide Web as a resource for example-based Machine Translation Tasks. In: Proc. of Aslib Conference on Translating and the Computer. London.

[13] Kilgarriff, A. et Rosenzweig, J. (2000). Framework and Results for English SENSEVAL, Computers and the Humanities, 34, (pp. 15-48).

[14] McCarthy, D., R. Koeling, J. Weeds, and J. Carroll, (2004) Finding predominant senses in untagged text. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. Barcelona, Spain. pp 280–287.

[15] Michael Lesk (1986). Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone, Proceedings of SIGDOC

[16] Mihalcea, R., Moldovan, D.I. (1999): An Automatic Method for Generating Sense Tagged Corpora. In: Proc. of the 16th National Conf. on Artificial Intelligence. AAAI Press.

[17] Miller, G. 1991. WordNet: An on-line lexical database. International Journal of Lexicography, 3(4).

[18] Litkowski, K. C. (2002). Sense Information for Disambiguation: Confluence of Supervised and Unsupervised Methods, Proceedings of the SIGLEX / SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions, Philadelphia.

[19] Pérez, D., J. Tecuapacho, H. Jiménez-Salazár, G. Sidorov: A term frequency range for text representation. In: Special issue "Advances in artificial intelligence and computer science", Journal "Research in computing science", vol. 20, 2006, pp. 113–118.

[20] Resnik, P., D. Yarowsky. (1997). A perspective on word sense disambiguation. In Proceedings if ACL Siglex Workshop on Tagging With Lexical Semantics, Why, What and How? Washington DC, April.

[21] Mihalcea, R., P. Tarau, E. Figa. (2004). PageRank on Semantic Networks with Application to Word Sense Disambiguation, COLING 2004.

[22] Sameer S. Pradhan, Edward Loper Dmitriy Dligach and Martha Palmer (2007). SemEval-2007 Task 17: English Lexical Sample, SRL and All Words". Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pp 87–92, Prague, June 2007. c2007 Association for Computational Linguistics

[23] Santamaria, C., Gonzalo, J., Verdejo, F. (2003): Automatic Association of WWW Directories to Word Senses. Computational Linguistics (2003), Vol. 3, Issue 3 – Special Issue on the Web as Corpus, 485–502.

[24] Vasilescu, F., P. Langlais, G. Lapalme. (2004).Evaluating variants of the Lesk approach for disambiguating words, LREC 2004.

[25] Volk, M. (2002): Using the Web as Corpus for Linguistic Research. In: Catcher of the Meaning. Pajusalu, R., Hennoste, T. (Eds.). Dept. of General Linguistics 3, University of Tartu, Germany.

[26] Wilks, Yorick and Stevenson, Mark. (1996). The grammar of sense: Is word sense tagging much more than part-of speech tagging? Technical Report CS-96-05, University of Sheffield, Sheffield, United Kingdom.

[27] Yarowsky, David (1993). One sense per collocation. Proceeding of ARPA Human Language Technology Workshop, Princeton, New Jersey, 266-271.