

Automatic Acquisition of Sense Tagged Corpora

Rada Mihalcea and Dan I. Moldovan

Department of Computer Science and Engineering
Southern Methodist University
Dallas, Texas, 75275-0122
{rada, moldovan}@seas.smu.edu

Abstract

An important problem in Natural Language Processing is identifying the correct sense of a word in a particular context. Thus far, statistical methods have been considered the best techniques in word sense disambiguation. Unfortunately, these methods produce high accuracy results only for a small number of preselected words. The reduced applicability of statistical methods is due basically to the lack of widely available semantically tagged corpora. In this paper we present a method which enables the automatic acquisition of sense tagged corpora. It is based on (1) the information provided in WordNet, particularly the word definitions found within the glosses and (2) the information gathered from Internet using existing search engines.

Introduction

Word Sense Disambiguation (WSD) is an open problem in Natural Language Processing. Its solution impacts other tasks such as discourse, reference resolution, coherence, inference and others.

Thus far, statistical methods have been considered the best techniques in WSD. They produce high accuracy results for small number of preselected words; the disambiguation process is based on the probability that a word could have a particular sense, given the context in which it occurs. The context is determined by the part of speech of encountering words, keywords, syntactic relations, collocations. These methods usually consist of two steps (1) a first training step, in which rules are acquired using various algorithms and (2) a testing phase in which the rules gathered in the first step are used to determine the most probable sense for a particular word. The weakness of these methods is the lack of widely available semantically tagged corpora.

The larger the corpora, the better the disambiguation accuracy. Typically, 1000-2500 occurrences of each word are manually tagged in order to create a corpus; from this, about 75% of the occurrences are used for the training phase and the remaining 25% are used for

the test phase. Although high accuracy can be achieved with these approaches, a huge amount of work is necessary to manually tag words to be disambiguated.

For the disambiguation of the noun *interest* with an accuracy of 78%, as reported in (Bruce and Wiebe, 1994), 2,476 usages of *interest* were manually assigned with sense tags from the Longman Dictionary of Contemporary English (LDOCE).

For the LEXAS system, described in (Ng and Lee 1996), the high accuracy is due in part to the use of a large corpora. For this system, 192,800 word occurrences have been manually tagged with senses from WordNet; the set consists of the 191 most frequently occurring nouns and verbs. As specified in their paper, approximatively one man-year of effort was spent in tagging the data set.

Thus, the sense tagging is done manually and creates serious impediments in applying statistic methods to word sense disambiguation.

In this paper, we present an automatic method for the acquisition of sense tagged corpora. It is based on (1) the information provided in WordNet, particularly the word definitions found within the glosses, and (2) information gathered from the Internet using existing search engines. Given a word for which corpora is to be acquired, we first determine the possible senses that the word might have based on the WordNet dictionary. Then, for each possible sense, we either determine a monosemous synonym from the word synset, if such a synonym exists, or extract and parse the gloss specified in WordNet, if a monosemous synonym does not exist. Each gloss contains a definition, which can be used as a more detailed explanation for each particular sense of the word we consider. The monosemous synonym or the definition will constitute the basis for creating a query which will be used for searching on the Internet. From the texts we gather, only those sentences containing the searching phrase will be selected. Further, the searching phrase will be replaced by the original word. In this way, we are creating example sentences for the usage of each sense of the word.

The idea of using the definitions is based on the fact that, in order to identify possible examples in which a particular sense of a word might occur, we need to

locate that particular meaning of the word within some text. The definitions provided in WordNet are specific enough to uniquely determine each sense of the word, thus searching for these definitions will enable us to find concrete examples.

To our knowledge, the only semantically tagged corpora with senses from WordNet is SemCor (Miller et al. 1994), which consists of files taken from the Brown corpus. In SemCor, all the nouns, verbs, adjectives and adverbs defined in WordNet are sense tagged. Although SemCor is a large collection of tagged data, the information provided by SemCor is not sufficient for the purpose of disambiguating words with statistical methods.

Consider, for example, the noun *interest*, which has 7 senses defined in WordNet. The number of occurrences in SemCor of the senses of *interest* are presented in Table 1

Sense number	No. of occurrences		Total occurrences	Automatic acquisition
	brown1	brown2		
1	33	25	58	246
2	15	6	21	545
3	7	25	32	895
4	5	9	14	1000
5	1	2	3	1000
6	0	7	7	718
7	0	4	4	1000
TOTAL	61	78	139	5404

Table 1: The number of occurrences of each sense of the noun *interest* in brown1 and brown2 concordance files from SemCor

The total of 139 occurrences of the noun *interest* is by far insufficient for creating rules leading to high accuracy disambiguation results.

For augmenting the data provided by SemCor, researchers have manually tagged other publicly available corpora, like for example The Wall Street Journal. We are proposing here a method for automatic acquisition of sense tagged corpora; even though this might be noisy, it is still much easier and less time consuming to check already tagged data then to start tagging from scratch. For the example considered, i.e. the noun *interest*, a total of 5404 occurrences have been found. The number of examples acquired for each of the senses of this noun are shown in Table 1 in the last column. A maximum of 1,000 examples can be acquired for each search phrase, due to a limitation imposed by the DEC-AltaVista that allows only the first 1,000 hits resulting from a search to be accessed.

Background on resources Several resources have been used in developing and testing our method. The first main step of extracting monosemous relatives or definitions for each sense of the considered word is performed based on the information provided in WordNet. The second step, i.e. fetching examples from the Internet, makes use of the AltaVista search engine.

WordNet¹ is a Machine Readable Dictionary developed at Princeton University by a group led by George

¹WordNet 1.6 has been used in our method.

Miller (Miller 1995) (Fellbaum 1998). WordNet covers the vast majority of nouns, verbs, adjectives and adverbs from the English language. It has a large network of 129,509 words, organized in 99,643 synonym sets, called *synsets*. There is a rich set of 299,711 relation links among words, between words and synsets, and between synsets.

The glosses in WordNet More than 95% of the synsets in WordNet have defining glosses. A gloss consists of a definition, comments and examples. For example, the gloss of the synset {*interest, interestingness*} is (*the power of attracting or holding one's interest (because it is unusual or exciting etc.); "they said nothing of great interest"; "primary colors can add interest to a room"*). It has a definition *the power of attracting or holding one's interest*, a comment *because it is unusual or exciting etc.* and two examples: *they said nothing of great interest* and *primary colors can add interest to a room*. Some glosses can contain multiple definitions or multiple comments.

AltaVista (AltaVista) is a search engine developed in 1995 by the Digital Equipment Corporation in its Palo Alto research labs. In choosing AltaVista for use in our system, we based our decision on the size of the Internet information that can be accessed through AltaVista (it has a growing index of over 160,000,000 unique World Wide Web pages) and on the possibility to create complex search queries using boolean operators (*AND, OR, NOT* and *NEAR*). This makes this search engine suitable for the development of software around it.

Previous work

Several approaches have been proposed so far for the automatic acquisition of training and testing materials.

In (Gale, Church and Yarowsky 1992), a bilingual French-English corpus is used. For an English word, the classification of contexts in which various senses of that word appear is done based on the different translations in French for the different word meanings. The problem with this approach is that aligned bilingual corpora is very rare; also, different senses of many polysemous words in English often translate to the same word in French, for such words being impossible to acquire examples with this method.

Another approach for creating training and testing materials is presented in (Yarowsky 1992). He is using Roget's categories to collect sentences from a corpus. For example, for the noun *crane* which appears in both Roget's categories *animal* and *tool*, he uses words in each category to extract contexts from *Grolier's Encyclopedia*.

(Yarowsky 1995) proposes the automatic augmentation of a small set of seed collocations to a larger set of training materials. He locates examples containing the seeds in the corpus and analyzes these to find new patterns; then, he retrieves examples containing these patterns. WordNet is suggested here as a source for seed collocations.

From (Leacock, Chodorow and Miller 1998), a method based on the monosemous words from WordNet is presented. For a given word, its monosemous lexical relatives provide a key for finding relevant training sentences in a corpus. An example given in their paper is the noun *suit* which is a polysemous word, but one sense of it has *business suit* as monosemous hyponym, and another has *legal proceeding* as a hypernym. By collecting examples containing *business suit* and *legal proceeding*, two sets of contexts for the senses of *suit* are built. Even this method proved to enable high accuracy results for WSD in respect to manually tagged materials, its applicability for a particular word *W* is limited by the existence of monosemous "relatives" (i.e. words semantically related to the word *W*) for the different senses of *W* and by the number of appearances of these monosemous "relatives" in the corpora. Restricting the semantic relations to synonyms, direct hyponyms and direct hypernyms, they found that about 64% of the words in WordNet have monosemous "relatives" in the 30-million-word corpus of the *San Jose Mercury News*. More than that, tests performed on a set of 1,100 words showed that only about 25% of word senses overall polysemous words have monosemous synonyms.

Our approach tries to overcome these limitations (1) by using other useful information that can be found in WordNet for a particular word, i.e. the word definitions provided the glosses and (2) by using a very large corpora, formed by the texts electronically stored on the Internet. An explanation uniquely identifying a word is provided either by its monosemous relatives, as they are defined in (Leacock, Chodorow and Miller 1998), or by its definition. Several procedures, shown later in this paper, are applied to determine such an explanation which will further constitute a search phrase. Based on this, several examples are automatically acquired from the World Wide Web, using an existing search engine.

Automatic acquisition of corpora

The method described in this paper enables the automatic acquisition of sentences as possible examples in which a particular sense of a word might occur; the word will be sense tagged in all these examples.

The basic idea is to determine a lexical phrase, formed by one or several words, which uniquely identifies the meaning of the word, and then find examples including this lexical phrase. Such a lexical phrase can be created either using monosemous synonyms of the word considered, or using the definition provided within the gloss attached to the WordNet synset in which the word occurs.

Applying this method on a particular word *W* involves three main steps:

1. For each sense $\#i$ of the word *W*, determine one or more search phrases using, in ascending order of preference, one of the procedures 1 through 4, described below.
2. Search on Internet using the search phrases de-

termined at step 1 and gather documents. From these documents, extract the sentences containing the search phrases.

3. Replace the search phrases in the collection of examples gathered in step 2 with the original word, labeled with the appropriate sense number, i.e. $W\#i$.

Procedures 2, 3 and 4 include a separate step in which the gloss attached to the word synset is parsed. The input to this parser is the gloss attached to the word synset. The output is a set of definitions, part of speech and syntactically tagged. Six steps are performed in order to parse the gloss.

Step 1. From each gloss, extract the definition part.

Step 2. Eliminate the explanatory part of the definition, such as words included in brackets, or phrases starting with *as of*, *of*, *as in*, *as for* etc.

Step 3. Part of speech tag the definition using Brill's tagger (Brill 1992).

Step 4. If the definition includes several phrases or sentences separated by semicolon, then each of these phrases can be considered as an independent definition.

Step 5. Syntactically parse the definitions, i.e. detect the noun phrases, verb phrases, preposition attachments (Srinivas 1997).

Step 6. Based on the parsing from the previous step and the position of the *or* conjunction, create definitions with maximum one verb phrase and one noun phrase. For example, the definition for *better* $\#1$ "*to make better in quality or more valuable*" will be separated into two definitions "*to make better in quality*" and "*to make more valuable*"

In order to determine one or more search phrases for a sense $\#i$ of the word *W*, denoted as $W\#i$, one of the following procedures will be applied, in ascending order. If a search on the Internet using the search phrases from *Procedure i* does not provide any hits, then *Procedure i+1* will be applied.

Procedure 1. Determine a monosemous synonym, from the $W\#i$ synset. A word is *monosemous* if it has exactly one sense defined in WordNet; a word having multiple senses is said to be *polysemous*. If such a synonym exists, this will constitute the search phrase. We performed several tests by considering also the direct hyponyms and direct hypernyms as possible relatives; the examples we gathered using such words proved to give less representative examples than using the definition from the glosses (*Procedure 2.*). Based on these empirical observations, we considered only the synonymy relations for this first rule.

Example. The noun *remember* $\#1$ has *recollect* as a monosemous synonym. Thus the search phrase for this word will be *recollect*.

Procedure 2. Parse the gloss, as explained above in this section. After the parse phase, we are left with a set of definitions, each of them constituting a search phrase.

Example. The verb *produce* $\#5$ has the definition (*bring onto the market or release, as of an intellectual cre-*

ation). The search phrase will be *bring onto the market* (the other possible definition *release* is eliminated, as being an ambiguous word).

Procedure 3. Parse the gloss. Replace the stop-words with the NEAR search-operator. The query will be straighten by concatenating the words from the current synset, using the AND search-operator. Using a query formed with the NEAR operator will reduce the precision of the search; for this, we reinforce the query with words from the synset. This is based on the idea of one sense per collocation, as presented in (Yarowsky 1993). As the definition is weaker with NEAR, the use of synonyms (i.e. the words from the synset) will reinforce it.

Example. The synset of *produce#6* is {*grow, raise, farm, produce*} and it has the definition (*cultivate by growing*). This will result in the following search phrase: *cultivate NEAR growing AND (grow OR raise OR farm OR produce)*.

Procedure 4. Parse the gloss. Keep only the head phrase, combined with the words from the synset using the AND operator, as in (*Procedure 3*).

Example. The synset of *company#5* is {*party, company*}, and the definition is (*band of people associated temporarily in some activity*). The search phrase for this noun will be: *band of people AND (party OR company)*.

An example

Consider, as an example, the acquisition of sentences for the different meanings of the noun *interest*. As defined in WordNet 1.6, *interest* is considered to be a common word, with a polysemy count of 7. The synsets and the associated glosses for each of the senses of *interest* are presented in Figure 1.

1. {*interest#1, involvement*} - (*a sense of concern with and curiosity about someone or something; "an interest in music"*)
2. {*interest#2, interestingness*} - (*the power of attracting or holding one's interest (because it is unusual or exciting etc.); "they said nothing of great interest"; "primary colors can add interest to a room"*)
3. {*sake, interest#3*} - (*a reason for wanting something done; "for your sake"; "died for the sake of this country"; "in the interest of safety"; "in the common interest"*)
4. {*interest#4*} - (*a fixed charge for borrowing money; usually a percentage of the amount borrowed; "how much interest do you pay on your mortgage?"*)
5. {*pastime, interest#5*} - (*a subject or pursuit that occupies one's time and thoughts (usually pleasantly); "sailing is her favorite pastime"; his main pastime is gambling"; "he counts reading among his interests"; "they criticized the boy for his limited interests"*)
6. {*interest#6, stake*} - (*a right or legal share of something; a financial involvement with something; "they have interests all over the world"; "a stake in the company's future"*)
7. {*interest#7, interest group*} - (*(usually plural) a social group whose members control some field of activity and who have common aims; "the iron interests stepped up production"*)

Figure 1: Synsets and associated glosses of the different senses of the noun *interest*

In Table 2, we present the search phrases we created for each of the senses of the noun *interest*, by applying

one of the Procedures 1-4.

Sense #	Search phrase
1	sense of concern AND (interest OR involvement)
2	interestingness
3	reason for wanting AND (interest OR sake)
4	fixed charge AND interest
5	percentage of amount AND interest
6	pastime
7	right share AND (interest OR stake) legal share AND (interest OR stake) financial involvement AND (interest OR stake)
7	interest group

Table 2: Search phrases for each sense of the noun *interest*

Using the (AltaVista) search-engine, 70 sentences have been extracted for the various senses of the noun *interest*, using the search phrases from Table 2. In Figure 2 we present some of these examples. All these examples have been manually checked: out of 70 sentences, 67 have been considered correct based on human judgment, thus a similarity of 95.7% with respect to manually tagged data.

1. I appreciate the genuine interest#1 which motivated you to write your message
2. The webmaster of this site warrants neither accuracy nor interest#2.
3. He forgives us not only for our interest#3, but for his own!
4. Interest coverage was 4.6x, and interest#4 coverage, including rents, was 3.6x.
5. As an interest#5, she enjoyed gardening and taking part in church activities.
6. Voted on issues, when they should have abstained because of direct and indirect personal interests#6 in the matters at hand.
7. The Adam Smith Society is a new interest#7 organized within the American Philosophical Association.

Figure 2: Context examples for various senses of the noun *interest*

Results

We tested our algorithm on a set of 10 words, randomly selected from a set of words considered to be common, based on their polysemy in WordNet. The set consists of 4 nouns: *interest, report, company, school*; 4 verbs: *produce, remember, write, speak*; 1 adjective: *small*, and 1 adverb: *clearly*. This led to a set of 75 different word meanings. For each of these words, the algorithm was applied and example contexts were acquired. As the tests were performed for the purpose of testing the efficiency of our method, rather than for acquiring large corpora, we consider only 10 examples for each sense of a word, from the top ranked documents. These we manually check for sense tagging correctness.

Table 3 presents the polysemy for each of the words, the total number of occurrences within SemCor (brown1, brown2 and brownv semantic concordances), the total number of examples acquired using our method, the examples out of this total which were manually checked and the number of examples which were considered to be correct, based on human judgment.

As it results from this table, for the 75 different meanings considered, a total of 658 examples have been automatically acquired and then manually checked. Out

Word	Poly-sense count	Examples in SemCor	Total # examples acquired	Examples manually checked	Correct examples
interest	7	139	5404	70	67
report	7	71	4196	70	63
company	9	90	6292	80	77
school	7	146	2490	59	54
produce	7	148	4982	67	60
remember	8	166	3573	67	57
write	8	285	2914	69	67
speak	4	147	4279	40	392
small	14	192	10954	107	92
clearly	4	48	4031	29	28
TOTAL	75	1432	49115	658	604

Table 3: Results obtained for example contexts gathered for 10 words

of these 658 examples, 604 proved to be correct, thus an accuracy of 92% such as the tag assigned with our method was the same as the tag assigned by human judgment.

Using this method, very large corpora can be generated. For the total of 10 words, 49,115 examples have been acquired using this method, over thirty times more than the 1,432 examples found in SemCor for these words. Even though this corpora is noisy, it is still much easier and less time consuming to check for correctness an already existing tagged corpora, then to start tagging free text from scratch.

An important observation which has to be made related to the number of examples which can be obtained is that this number does not always correlate with the frequency of senses, thus classifiers using such a corpora will have to establish prior probabilities.

Conclusion and further work

In this paper we presented a method which enables the automatic acquisition of sense tagged corpora, based on the information found in WordNet and on the very large collection of texts which can be found on the World Wide Web. The system has been tested on a total of 75 different word meanings and 658 context examples for these words have been acquired. The accuracy of 92% such as the tag assigned with our method was the same with the tag assigned by human judgment is encouraging.

Further work will include the use of this method for automatic acquisition of very large corpora which will be used to test word sense disambiguation accuracy.

References

- Digital Equipment Corporation. AltaVista Home Page. URL:<http://www.altavista.com>.
- Brill, E. 1992, A simple rule-based part of speech tagger, *Proceedings of the Third Conference on Applied Natural Language Conference*, ACL, Trento, Italy, 1992.
- Bruce, R. and Wiebe, J. 1994 Word Sense Disambiguation using Decomposable Models, *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, LasCruces, 1994.

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

Gale, W.; Church, K. and Yarowsky, D. 1992, One Sense per Discourse, *Proceedings of the DARPA Speech and Natural Language Workshop*, New York, 1992.

Leacock, C.; Chodorow, M. and Miller, G.A. 1998, Using Corpus Statistics and WordNet Relations for Sense Identification, *Computational Linguistics*, March 1998.

Miller, G.A.; Chodorow, M.; Landes, S.; Leacock, C. and Thomas, R.G. 1994, Using a semantic concordance for sense identification. *Proceedings of the ARPA Human Language Technology Workshop*, 240-243, 1994.

Miller, G.A. 1995, WordNet: A Lexical Database, *Communication of the ACM*, vol 38: No11, November 1995.

Ng, H.T. and Lee, H.B. 1996, Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach, *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL-96)*, Santa Cruz, 1996.

Srinivas, B. 1997, Performance Evaluation of Supertagging for Partial Parsing, *Proceedings of Fifth International Workshop on Parsing Technology*, Boston, USA, September 1997.

Yarowsky, D. 1992, Word-sense disambiguation using statistical models of Roget's categories trained on large corpora, *Proceedings of COLING-92*, Nantes, France, 1992.

Yarowsky, D. 1993, One sense per collocation, *Proceedings of ARPA Human Language Technology*, Princeton, 1993

Yarowsky, D. 1995, Unsupervised Word Sense Disambiguation rivaling Supervised Methods, *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL-95)*, Cambridge, 1995.