An Automatic Method for Generating 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

The unavailability of very large corpora with semantically disambiguated words is a major limitation in text processing research. For example, statistical methods for word sense disambiguation of free text are known to achieve high accuracy results when large corpora are available to develop context rules, to train and test them.

This paper presents a novel approach to automatically generate arbitrarily large corpora for word senses. The method is based on (1) the information provided in WordNet, used to formulate queries consisting of synonyms or definitions of word senses, and (2) the information gathered from Internet using existing search engines. The method was tested on 120 word senses and a precision of 91% was observed.

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 a small number of preselected words; the disambiguation process is based on the probability of occurrence of a particular sense in a given context The context is determined by the part of speech of surrounding words, keywords, syntactic relations, collocations.

Statistical methods for WSD consist usually of two phases:

1. a training phase, in which rules are acquired using various algorithms
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 disambiguation accuracy is strongly affected by the size of the corpora used in the disambiguation process. A larger corpora will enable the acquisition of a larger set of rules during the training phase, thus a higher 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 testing. 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. The information from WordNet is used to formulate a query consisting of synonyms or definitions of a word sense, while the Internet search engine extracts texts relevant to such queries.

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 the monosemous synonyms from the word synset, if such synonyms exist, or if not use the information provided by the gloss of that word sense. 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 synonyms or the definitions constitute the basis for creating a query which will be used for searching on 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 create example sentences for the usage of each sense of the word.

Background on resources.

The following resources have been used in developing and testing our method.

WordNet [1] is a Machine Readable Dictionary developed at Princeton University by a group led by George 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.

WordNet glosses The synsets of 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''` and two examples: `''they said nothing of great interest''` and `''primary colors can add interest to a room''`.
Some glosses 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, with the goal of increasing the quality of the information retrieved.

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.

[1]WordNet 1.6 has been used in our method.

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 finds 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.

The idea of using the definitions is based on the fact that, in order to identify possible examples in which a word with a given sense 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 of the senses of *interest* in SemCor is shown 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.

To augment 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, still it is much easier and less time consuming to check already tagged data then to start tagging from scratch. For the noun *interest*, a total of 5404 occurrences have been found using our method, thus significantly more than the 139 occurrences found in SemCor for the same word. The number of examples acquired for each of the senses of this noun are shown in Table 1 in the last column. Only 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.

The algorithm

The acquisition of sense tagged corpora for a particular word W, using our method, involves three main steps.

1. Preprocessing phase. For each sense $\#i$ of a word W, determine the synset from WordNet in which $W\#i$ is included. For each such synset:
   - Determine all the monosemous words included in the synset. A word is *monosemous* if it has exactly one sense defined in WordNet; a word having multiple senses is said to be *polysemous.*
   - Parse the gloss attached to the synset. This involves: (1) the separation of the gloss into its component parts (definitions, explanations, examples), (2) part of speech tagging and (3) syntactic tagging of the gloss definitions.
2. Search phase. For each sense $\#i$ of the word W, (1) form search phrases SP using, in ascending order of preference, one of the procedures 1 through 4, described below; then (2) search on Internet using the search phrases previously determined and gather documents; and (3) extract from these documents the sentences containing the search phrases.
3. Postprocessing phase. The sentences gathered during phase 2 will become examples for the usage of the original word $W\#i$. For this: (1) the part of speech of the search phrase SP within these examples is checked to be the same as the part of speech for $W\#i$; (2) the sentences in which SP has the same part of speech as $W\#i$ become valid examples by replacing SP with $W\#i$; the examples in which the part of speech of SP is different with respect to $W\#i$ are eliminated.

PREPROCESSING. During this phase: (1) For each sense $\#i$ of a word W, its monosemous synonyms are determined. For example, the adjective *large#3* belongs to the synset {`macroscopic, macroscopical, large`}. Both *macroscopic* and *macroscopical* have only one sense defined in WordNet, thus they will be picked as monosemous synonyms of *large#3*; (2) The gloss from the synset of $W\#i$ is parsed.

The input to the 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''`

SEARCH. In order to determine one or more search phrases for a sense $\#i$ of a 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. If such a synonym exists, this will constitute the search phrase.

Rationale. The context of a word is determined by the sense of that word. In the case of monosemous words, the context does not depend anymore on the sense of the word and is determined only by the word as a lexical string.

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 then using the definition from the glosses (*Procedure 2.*). Based on these empirical observations, we restricted the patterns for *Procedure 1* to synonymy relations.

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.

Rationale. The role of a dictionary is to give definitions which uniquely identify the meaning of the words. Thus, the definition is specific enough to determine the context in which a particular word could appear.

Example. The verb *produce#5* has the definition `(bring onto the market or release, as of an intellectual creation)`. 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 strengthened by concatenating the words from the current synset, using the AND search-operator.

Rationale. Using a query formed with the NEAR operator increases the number of hits but reduces 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).

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.*).

Rationale. If the search phrase determined during the previous procedure does not give any hits, the query can be relaxed by replacing the NEAR operator with the AND operator. Again, a reinforcement is achieved by appending to the query the words from the synset.

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).*

Searching on Internet with the queries from *Procedures 1-4*, several documents will be found. From these texts, only those sentences containing the search phrases SP, formed by monosemous synonyms or definitions of *W#i*, will be extracted.

POSTPROCESSING. The examples gathered during the search phase contain SP phrases, which have to have the same part of speech functionality as the original word *W*. If SP consists of a single word, then part of speech tagging (Brill 1992) will be enough to check if SP has the same functionality as the original word *W*. If SP consists of a phrase, then a further syntactic parsing is needed to determine if SP is a noun, verb, adjective or adverb phrase and whether or not it has the same functionality as *W*. Those examples containing SP with a different part of speech / syntactic tag with respect to the original word will be eliminated. In the remaining collection of examples, SP will be replaced with the original word, labeled with the appropriate sense number, i.e. *W#i*.

An example

Let us consider the acquisition of sentences for the different meanings of the noun *interest*. As defined in WordNet 1.6, *interest* is a common word, with a polysemy count of 7. The synset and the associated gloss for each sense of *interest* are presented in Figure 1.

In Table 2, we present the search phrases 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 percentage of amount AND interest
5	pastime
6	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*

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 ffor 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*

Using the (AltaVista) search-engine, 5404 sentences have been extracted for the various senses of the noun *interest*, using the search phrases from Table 2. From these, 70 examples were manually checked, out of which 67 were considered correct based on human judgment, thus an accuracy of 95.7% with respect to manually tagged data. Some of these examples are presented in Figure 2.

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

The algorithm presented here was tested on 20 polysemous words. The set consists of 7 nouns: *interest*, *report*, *company*, *school*, *problem*, *pressure*, *mind*; 7 verbs: *produce*, *remember*, *write*, *speak*, *indicate*, *believe*, *happen*; 3 adjectives: *small*, *large*, *partial* and 3 adverbs: *clearly*, *mostly*, *presently*. Overall, the 20 words have 120 word senses. The algorithm was applied to each of these senses and example contexts were acquired. Since experiments were performed for the purpose of testing the efficiency of our method, rather then for acquiring large corpora, we retained only a maximum of 10 examples for each sense of a word, from the top ranked documents. The correctness of the results was checked manually.

Table 3 presents the polysemy of each word, the total number of examples found in the SemCor corpus, the total number of examples acquired using our method, the number of examples that were manually checked and the number of examples which were considered to be correct, based on human judgment.

Word	Poly-semy count	Exam-ples 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
problem	3	199	710	23	23
pressure	5	101	2067	50	45
mind	7	113	7000	70	56
produce	7	148	4982	67	60
remember	8	166	3573	67	57
write	8	285	2914	69	67
speak	4	147	4279	40	39
indicate	5	183	4135	50	47
believe	5	215	3009	36	33
happen	5	189	5000	50	46
small	14	192	10954	107	92
large	8	129	5107	80	66
partial	3	1	598	23	18
clearly	4	48	4031	29	28
mostly	2	12	2000	20	20
presently	2	8	2000	20	20
Total	120	2582	80741	1080	978

Table 3: Results obtained for example contexts gathered for 20 words

As it results from this table, for the 120 different meanings considered, a total of 1081 examples have been automatically acquired and then manually checked. Out of these 1081 examples, 981 proved to be correct, leading to an accuracy of 91% 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 20 words, 80,741 examples have been acquired using this method, over thirty times more than the 2,582 examples found in SemCor for these words. Even though the corpora might be noisy, still it is much easier and less time consuming to check for correctness an already existing tagged corpora, then to start tagging free text from scratch.

Discussion. In almost all the cases considered, a large number of example sentences were found as a result of the Internet search. For some cases, though, only a few sentences have been retrieved. For example, the word *believe#5* belongs to the synset {`believe`} and it has the definition `(credit with veracity)`. Searching on the Internet with the query created based on our method, i.e. *credit with veracity*, or with a variant of this query *credit NEAR veracity AND believe*, resulted in only 4 hits. For such cases, a refinement of our method is needed, which considers also the hypernyms and hyponyms of a synset, together with their gloss definitions.

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

Related work

Several approaches have been proposed so far for the automatic acquisition of training and testing materials. In (Gale, Church et al., 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 as a source for seed collocations. Given an ambiguous word W, with its different meanings W#i, the algorithm presented in (Yarowsky 1995) identifies example sentences for W#i based on the words occurring in its context; the set of words likely to appear in W#i context is built iteratively. On the other hand, our method tries to locate example sentences for W#i by identifying words or expressions similar in meaning with W#i and which uniquely identify the sense #i of the word W.

In (Leacock, Chodorow et al., 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 exhibits high accuracy results for WSD with 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*.

Our approach tries to overcome these limitations (1) by using other useful information in WordNet for a particular word, i.e. the word definitions provided by glosses and (2) by using a very large corpora, consisting of texts electronically stored on the Internet. The unique identification of a word is provided either by its monosemous relatives, as they are defined in (Leacock, Chodorow et al., 1998), or by its definition.

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 available on the World Wide Web. The system has been tested on a total of 120 different word meanings and 80,741 context examples for these words have been acquired. Out of these, 1,081 examples were checked against human judgment which resulted in a 91% accuracy.

There is no basic limitation on the size of the corpus acquired for each word, other than the need to check the results of the Internet search, and filter out the inappropriate texts. Further work is needed to automate this verification. We plan to use 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, *Proc. 3rd 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, June 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 Examplar-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 Association of Computational Linguistics, 1995.