Bootstrapping Large Sense Tagged Corpora

Rada F. MIHALCEA

University of Texas at Dallas
Richardson, Texas, 75083-0688
rada@utdallas.edu

Abstract

The performance of Word Sense Disambiguation systems largely depends on the availability of sense tagged corpora. Since the semantic annotations are usually done by humans, the size of such corpora is limited to a handful of tagged texts. This paper proposes a generation algorithm that may be used to automatically create large sense tagged corpora. The approach is evaluated through comparative sense disambiguation experiments performed on data provided during the SENSEVAL-2 *English all words* and *English lexical sample* tasks.

1. Introduction

The availability of semantically tagged corpora is crucial for creating successful Word Sense Disambiguation (WSD) systems. The tagging process is usually done by humans, and therefore is highly expensive, thereby limiting the size of such corpora to a handful of tagged texts.

This paper describes an algorithm for the automatic generation of large sense tagged corpora. This algorithm was used to create *GenCor*, a semantically tagged corpus that was successfully employed in a system that participated in the SENSEVAL-2 competition, with an excellent performance during the *English all words* task. The dictionary used to accomplish the corpus sense tagging task is WordNet (Miller, 1995), which is at the same time the dictionary employed during the *English* tasks in SENSEVAL-2.

The generation algorithm is iterative, and follows the principles of a bootstrapping algorithm. We start with a set of seeds that are used (1) to extract text snippets from the Web, which are then added to the sense tagged corpus, and (2) to identify other instances of ambiguous words that can be accurately sense tagged. The newly tagged words are added to the set of seeds and the generation process continues.

To create the set of seeds, we make use of available sense tagged corpora, namely SemCor (Miller et al., 1993), and tagged examples that can be automatically extracted from WordNet. Seeds may also be extracted from additional sense tagged examples, as for instance the training data provided for the lexical sample tasks during SENSEVAL-1 and SENSEVAL-2. The seeds are merely formed as multiple word units that include an ambiguous word, such that the expression itself places a constraint over the possible meaning for the word of interest.

The generation algorithm is evaluated in two ways. First, this algorithm was used to create the *GenCor* corpus, which was employed during the *English all words* task, with significant improvement measured over the baseline performance. Secondly, we use the algorithm to build corpora for a subset of randomly selected words from the *English lexical sample* task. The performance achieved with the generated corpora is compared with the performance obtained when manually tagged data is employed. While the test benchmarks and tagging algorithms are the same in both experiments, the precision achieved with automatically tagged data is comparable, and sometimes better than the precision obtained with manually tagged data, at a production cost significantly lower.

The work presented in this paper relates to work previously reported in (Yarowsky, 1995), where few tagged seeds are used to train a decision list, which is then employed to tag new unlabeled instances. An efficient method for automatic generation of sense tagged corpora was also presented in (Mihalcea and Moldovan, 1999). The generation algorithm described in this paper combines these two previous approaches. Seeds are generated using either unambiguous words that are semantically related to a given word sense, or phrases that uniquely identify the sense of a word. Moreover, the algorithm is evaluated in the context of a disambiguation task performed on data provided during the SENSEVAL-2 *English all words* task, respectively a subset of the words from the *English lexical sample* task.

The paper is organized as follows. First, we give a short overview of the main approaches considered so far in WSD, and illustrate the need for large sense tagged corpora as an essential resource in developing accurate WSD algorithms. Next, we present the generation algorithm, able to create very large sets of tagged examples starting with few seeds. The algorithm is illustrated with an ambiguous word from the *English lexical sample* task, namely the noun *channel*. The generation methodology is evaluated (1) within the framework of the *English all words* task, and (2) with comparative experiments performed on randomly selected nouns from the *English lexical sample* task.

2. Word Sense Disambiguation

The task of WSD consists in assigning the most appropriate meaning to a polysemous word within a given context. A large range of applications, including machine translation, knowledge acquisition, information retrieval, information extraction, and others, require knowledge about word meanings, and therefore WSD algorithms represent a necessary step in all these applications. Starting with SENSEVAL-1 in 1999 (Kilgarriff and Palmer, 2000), WSD has received growing attention from the Natural Language Processing community, and motivates a continuously increasing number of researchers to develop systems and to try to find solutions to this challenging problem.

Most of the efforts in the WSD field have been concentrated so far towards *supervised* learning algorithms, and these are the methods that usually achieve the best performance at the cost of low recall. Each sense tagged occurrence of a particular word is transformed into a feature vector, suitable for an automatic learning process. Two main decisions need to be made when designing such a system: the set of features to be used and the learning algorithm. Commonly used features include surrounding words and their part of speech(Bruce and Wiebe, 1994), context keywords (Ng and Lee, 1996) or context bigrams (Pedersen, 2001), various syntactic properties (Fellbaum et al., 2001) etc. As for the learning methodology, a large range of algorithms have been employed, including neural networks (Leacock et al., 1998), decision trees (Pedersen, 2001), memory based learning (Veenstra et al., 2000) and others. Lately, classifiers tailored to the behavior of each word (Mihalcea and Moldovan, 2002), and combinations of heterogeneous classifiers (Tolga Ilhan et al., 2001) were proved as useful and efficient techniques for WSD.

The main weakness of these methods is the lack of widely available semantically tagged corpora. The disambiguation accuracy is strongly affected by the size of the training corpus. Given the high cost associated with the process of creating sense tagged data, all the attempts made so far were limited to labeling examples for few preselected words.

One of the first large scale hand tagging efforts is reported in (Bruce and Wiebe, 1994), where 2,476 usages of *interest* were manually assigned with sense tags from the Longman Dictionary of Contemporary English (LDOCE). This corpus was used in various experiments, with accuracies ranging from 75% to 90%, depending on the algorithm and features employed.

For the LEXAS system, described in (Ng and Lee, 1996), the high accuracy is due in part to the use of 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. The authors mention that approximatively one man-year of effort was spent in tagging the data set.

Lately, the SENSEVAL competitions provide a good environment for the development of supervised WSD systems, making freely available large amounts of sense tagged data for about 100 words. During SENSEVAL-1 in 1999, data for 35 words was made available adding up to about 20,000 examples tagged with respect to Hector dictionary. The size of the tagged corpus increased with SENSEVAL-2 in 2001, when 13,000 additional examples were released for 73 polysemous words. This time, the semantic annotations were performed with respect to WordNet.

2.1. WSD Evaluation System

The corpora generated with the methodology presented in this paper were evaluated using the SMUls/SMUaw disambiguation system, which is presented in great detail in (Mihalcea and Moldovan, 2002). Shortly, the system consists of two important modules. The first module uses pattern learning that relies on machine readable dictionaries and sense tagged corpora to tag all words in open text. The second module is triggered only for words with large training data, and it uses an instance based learning algorithm with automatic feature selection. The two modules are preceded by a preprocessing phase that includes compound concept identification, followed by a default phase that assigns the most frequent sense as a last resort, when no other previous methods could be applied.

During the preprocessing stage, SGML tags are eliminated, the text is tokenized, part of speech tags are assigned using Brill tagger (Brill, 1995), and Named Entities (NE) are identified with an *in-house* implementation of an NE recognizer. To identify collocations, we determine sequences of words that form compound concepts, based on WordNet definitions.

In the second step, patterns are learned from WordNet, SemCor and GenCor, which includes automatically labeled examples built with the methodology described in this paper.

The third step consists of a learning mechanism with automatic feature selection. This step is initiated only for words with a sufficiently large number of examples, as it was the case with the words in the SENSEVAL lexical sample tasks. It is important to mention that training and testing corpora are extracted for each ambiguous word. This means that examples pertaining to the multi-word *"dress down"* are separated from the examples for the single word *"dress"*.

The *SMUls/SMUaw* system achieved an excellent performance at SENSEVAL, and was ranked first in both *English all words* and *English lexical sample* tasks. A precision of 69% was measured in the *all words* task, respectively 63.8% for the *lexical sample* task. We make use of this system to evaluate the quality of the automatically generated corpora.

3. Generation Algorithm

The generation algorithm comprises three main steps:

Step 1. Create a set of seeds, consisting of

1.1 Sense tagged examples in SemCor

1.2 Sense tagged examples extracted from WordNet

1.3 Sense tagged examples created with the principles described in (Mihalcea and Moldovan, 1999). Currently, we only use monosemous synonyms, hypernyms or hyponyms for any given word.

1.4 Additional hand tagged examples, if available.

Step 2. Search the Web using queries formed with the seed expressions. Add to the generated corpus a maximum of N text passages containing the seed expressions.

Step 3. Disambiguate words in a small text snippet surrounding the seed expressions using the main ideas of the algorithm in (Mihalcea and Moldovan, 2000). Add examples formed with the disambiguated words to the seed set. Go back to step 2.

The sequences of words formed during step 1 have to obey the following constraints: (1) contain at least two open class words (2) at least one of the two open class words is semantically tagged (3) the words are part of a noun phrase or are involved in a verb-object or verb-subject relation.

Step 2 is the source of new sense tagged examples, and step 3 is the source of new seeds for the generation algorithm. Shortly, the algorithm employed at this stage disambiguates words based on their relation with already tagged words. The relations considered between words are (1) the identity relation; (2) the synonymy relation; (3) the hypernymy, hyponymy and sibling relations. This disambiguation procedure was proved highly accurate in the experiments reported in (Mihalcea and Moldovan, 2000). During the process of building *GenCor*, all three semantic relations are enabled. This is because we want *GenCor* to cover as many ambiguous words as possible, since this is a corpus employed for the disambiguation of *all words* in open text. On the other side, when the corpus is generated for one pre-selected ambiguous word, as in the example presented in the following section, or in the experiments involving words from the *lexical sample* task, the identity relation is the only relation employed to identify new seeds. The reason for this decision is the fact that in this case we are only interested in instances containing the pre-selected ambiguous word, and do not want to extract seeds pertaining to related words.

Regarding the corpora used as a resource for creating the initial seeds, both SemCor and WordNet had to undergo some transformations that would make them suitable for our sense tagging task. First of all, SemCor (Miller et al., 1993) was available only for earlier versions of WordNet. We had therefore to process this corpus and map the WordNet 1.6 senses to their corresponding senses in WordNet 1.7.[1] Secondly, we transformed the examples in WordNet definitions such that they may be used as a source of sense tagged examples. The main idea in creating sense tagged examples out of WordNet is very simple. It is based on the underlying assumption that each example in a gloss pertains to a word belonging to the current synset, thereby allowing us to assign the correct sense to at least one word in each example. For instance, the example given for *mother#4* is *"necessity is the mother of invention"*, where the word *mother* can be tagged with its appropriate sense.

4. A Walk Through Example

To illustrate the algorithm behavior and performance, let us consider an example, where a sense tagged corpus is generated for a highly ambiguous word. To this end, we use the noun *channel*, which has seven different senses defined in WordNet, and is one of the words provided during the SENSEVAL-2 *English lexical sample* task. The seven WordNet senses defined for this noun are listed in Table 1.

To initiate the generation algorithm, one starting seed is identified for each sense. As mentioned in the previous section, the seeds consist of multi word expressions that have the property of uniquely identifying the sense of the ambiguous word considered. Usually, these multi word expressions consist of noun phrases or verb noun constructs, depending on the nature of the ambiguous word.

Table 2 lists the starting seeds for *channel*. These seeds are used to extract text snippets from the Web. Usually, there are thousands of pages retrieved for each seed. For this experiment, the number of passages extracted for a certain seed was limited to a maximum of 100. Larger corpora can be build by raising this threshold to higher values, and further experiments will investigate the effect of very large corpora over the disambiguation accuracy.

1. {channel, transmission channel} - (a path over which electrical signals can pass)
2. {conduit, channel} - (a passage for water (or other fluids))
3. {groove, channel} - (a long narrow furrow cut either by a natural process (such as erosion) or by a tool (as e.g. a groove in a phonograph record))
4. {channel, sound} - (a relatively narrow body of water linking two larger bodies; ”the ship went aground in the channel”)
5. {channel, communication channel, line} - ((often plural) a means of communication or access; ”it must go through official channels”; ”lines of communication were set up between the two firms”)
6. {duct, canal, channel} - (a bodily passage or tube conveying a secretion or other substance)
7. {channel, television channel, TV channel} - (a television station and its programs; ”a satellite TV channel”; ”surfing through the channels”; ”they offer more than one hundred channels”)

Table 1: WordNet senses for the noun *channel*

channel#1:	“fiber optic channel”
channel#2:	“river channel”
channel#3:	“channels in the surface”
channel#4:	“water channel”
channel#5:	“channel of expression”
channel#6:	“calcium channel”
channel#7:	“sports channel”

Table 2: Starting seeds for the noun *channel*

The corpus build using only the initial set of seeds consists of 393 examples. Next, from the examples extracted for each sense, we attempt to identify occurrences of the same word *channel*, in a small text window around the original seed, but belonging to a multi word unit different than the seed itself. The newly identified expressions are added to the seed set for the particular word sense considered. The additional seeds extracted for various senses of *channel* are listed in Table 3.

Note that for each sense, only a maximum of 100 text snippets, as obtained with the original seed, are used as a possible source for additional seeds. A larger threshold will lead to a larger search space, and therefore a possibly larger set of seeds. This search space limitation is also the reason for no additional seeds found for *channel#3*. There were no multi word expressions found in the closed vicinity of “channels in the surface”, other than the seed itself.

The new seeds are employed to extract additional passages of text that will be appended to the set of examples

[1] SemCor 1.6 is available for download from the WordNet site http://www.cogsci.princeton.edu/˜wn/. SemCor 1.7 can be downloaded from http://www.seas.smu.edu/˜rada/semcor

channel#1:	“optical fiber channel”, “channel telephone” “transmission channel”, “channel banks” “channel service unit”, “channel interface” “channel fiber optic”, “multiplex channel” “mainframe channel”, “optic channel” “fiber optic sensor channel”
channel#2	“channel catfishing”, “channel catfish” “Channel morphology”, “Navigation Channel”, “Channel Hydrogeomorphology” “channel confluences”, “channel sand” “channel fishes”, “shipping channel”, “channel bed topography”
channel#4	“lagoon channel”, “Songkhla Channel” “Port Everglades Channel”, “Bay Channel” “Anclote River South Channel” “SANTA CRUZ CHANNEL” “Nassau Bay channel”
channel#5	“channel of activation” “channel of representation”
channel#6	“channel disorder”, “channel blocker” “channel genes”
channel#7	“CPB Channel”, “channel guide” “channel combiner”, “Malayalam channel”

Table 3: Additional seeds extracted for the seven senses of *channel*

already found for each word sense.

To evaluate the quality of the automatically generated corpus, we train our semantic tagger on this corpus and tag the test instances provided during SENSEVAL-2 for the word *channel*. The same semantic tagger and the same test instances are used in a second comparative experiment, where the system is trained on the hand tagged data provided during SENSEVAL-2.

Table 4 shows the number of examples acquired for each sense, using first the original set of seeds, and gradually adding examples extracted with the newly learned seeds. For each such training set, we determine the disambiguation accuracy on the test set. The figures in Table 4 are to be compared with the accuracies of 34.1% (fine grained scoring), respectively 47.7% (coarse grained scoring), obtained when the system is trained on hand tagged data. The last row in the table indicates the baseline accuracy, measured for the case when all instances in the test set are tagged by default with the most frequent sense.

The test set used during these experiments contains 52 instances. The original set provided during SENSEVAL-2 consisted of 73 instances. As mentioned earlier, multi word expressions are identified and separated from the training and test sets during the preprocessing stage. That is, all examples containing “Channel” tagged as a proper noun, or the collocations “television channel”, “Bristol channel”, “English Channel”, “Channel Tunnel”, and others, were eliminated from the test set, shrinking therefore its size from 73 to 52 examples. Similarly, the hand tagged training set went down from 138 examples to 78 examples, and the generated corpus changed from 2,218 examples to the final set of 1,851 examples.

It is interesting to observe that the fine grained precision drops after new examples are added for channel#6 and channel#7. Nevertheless, the best coarse grained precision is obtained when the system is trained on the entire set of 1,851 generated examples.

5. Results

The generation algorithm was evaluated in two ways.

First, a corpus was automatically generated starting with seeds containing the ambiguous words in the texts provided during the SENSEVAL-2 *English all words* task. The starting seeds were extracted automatically from SemCor and/or the tagged examples from WordNet. The size of the corpus generated for this task, which we referred to as *GenCor*, is about 160,000 examples. To evaluate the effect of *GenCor* on this tagging task, we first determine a baseline precision. That is, if all words are tagged with their most frequent sense, the precision in this task is 63.9%. Next, if SemCor and WordNet corpora are both used in the pattern learning procedure, the precision raises to 65.1%. Finally, when GenCor is used in addition to these two corpora, a precision of 69.3% is attained, therefore a significant improvement resulting from the use of this generated corpus.

Additionally, the generation algorithm was employed to create sense tagged corpora for randomly selected nouns from the *English lexical sample* task. For each such word, two comparative experiments are performed. In one experiment, the system is trained on the automatically generated corpus. In the second experiment, the system is trained on the hand tagged data provided during SENSEVAL. As mentioned earlier, test data and system parameters are the same in both experiments. Table 5 shows comparative results obtained during these experiments. For each word, the table indicates fine grained and coarse grained precisions, for the case when training is performed on hand tagged data, respectively for the case when the training is done using the generated corpus. From these experiments, it results that the precision achieved with the generated corpus is comparable, and sometimes better than the precision achieved with hand tagged corpora.

6. Conclusions

This paper has introduced a bootstrapping algorithm for generating large sense tagged corpora. The algorithm starts with a set of seeds that is used (1) to extract text snippets from the Web that are added to the sense tagged corpus, and (2) to identify new seeds, which may be subsequently used in the corpus generation process. The approach was evaluated through comparative sense disambiguation experiments performed on data provided during the SENSEVAL-2 *English all words* and *English lexical sample* tasks, with results that proved the usefulness of the automatically generated sense tagged corpora.

Training examples for sense							Training	Precision	
#1	#2	#3	#4	#5	#6	#7	size	(fine)	(coarse)
62	63	51	51	100	49	17	393	27.3%	40.9%
+425							818	34.1%	54.5%
	+421						1239	34.1%	59.1%
			+154				1393	36.4%	59.1%
				+79			1472	40.9%	63.6%
					+215		1687	34.1%	59.1%
						+164	1851	34.1%	65.9%
Hand tagged data							78	34.1%	47.7%
Baseline (most frequent sense)							-	22.7%	47.7%

Table 4: Fine and coarse grained disambiguation accuracy, obtained with the generated training corpus, for the noun *channel*.

		Hand tagged corpus			Automatically generated corpus		
	Test	Training	Precision		Training	Precision	
Word	size	size	(fine)	(coarse)	size	(fine)	(coarse)
art	52	123	65.4%	73.1%	265	73.1%	75.0%
chair	63	121	82.5%	84.1%	179	87.3%	87.3%
church	36	81	63.9%	63.9%	189	58.3%	58.3%
detention	24	46	87.5%	87.5%	163	83.3%	83.3%
nation	26	60	73.1%	73.1%	225	69.5%	69.5%

Table 5: Comparative results in semantic disambiguation, when training data is (1) hand tagged, and (2) automatically generated.

7. References

E. Brill. 1995. Transformation-based error driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–566, December.

R. Bruce and J. Wiebe. 1994. Word sense disambiguation using decomposable models. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, pages 139–146, LasCruces, NM, June.

C. Fellbaum, M. Palmer, H.T. Dang, L. Delfs, and S. Wolf. 2001. Manual and automatic semantic annotation with WordNet. In *WordNet and Other lexical resources: NAACL 2001 workshop*, pages 3–10, Pittsburgh, June.

A. Kilgarriff and M. Palmer, editors. 2000. *Computer and the Humanities. Special issue: SENSEVAL. Evaluating Word Sense Disambiguation programs*, volume 34, April.

C. Leacock, M. Chodorow, and G.A. Miller. 1998. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1):147–165.

R. Mihalcea and D.I. Moldovan. 1999. An automatic method for generating sense tagged corpora. In *Proceedings of AAAI-99*, pages 461–466, Orlando, FL, July.

R. Mihalcea and D.I. Moldovan. 2000. An iterative approach to word sense disambiguation. In *Proceedings of FLAIRS-2000*, pages 219–223, Orlando, FL, May.

R. Mihalcea and D. Moldovan. 2002. Word sense disambiguation using pattern learning and automatic feature selection. *Journal of Natural Language Engineering (submitted)*.

G. Miller, C. Leacock, T. Randee, and R. Bunker. 1993. A semantic concordance. In *Proceedings of the 3rd DARPA Workshop on Human Language Technology*, pages 303–308, Plainsboro, New Jersey.

G. Miller. 1995. Wordnet: A lexical database. *Communication of the ACM*, 38(11):39–41.

H.T. Ng and H.B. Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An examplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL-96)*, Santa Cruz.

T. Pedersen. 2001. A decision tree of bigrams is an accurate predictor of word sense. In *Proceedings of the North American Chapter of the Association for Compuatational Linguistics, NAACL 2001*, pages 79–86, Pittsburg, June.

H. Tolga Ilhan, S.D. Kamvar, D Klein, C.D. Manning, and K. Toutanova. 2001. Combining heterogeneous classifiers for word-sense disambiguation. In *Senseval 2001, ACL Workshop*, Toulouse, France, July.

J. Veenstra, A. van den Bosch, S. Buchholz, W. Daelemans, and J. Zavrel. 2000. Memory-based word sense disambiguation. *Computers and the Humanities*, 34:171–177.

D. Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL-95)*, pages 189–196, Cambridge, MA, 1995.