An Experiment in Semantic Tagging using Hidden Markov Model Tagging

Frédérique Segond, Anne Schiller, Gregory Grefenstette, Jean-Pierre Chanod
Rank Xerox Research Centre, 6 Chemin de Maupertuis , F-38240 Meylan, France
{Segond, Schiller, Grefenstette, Chanod}@grenoble.rxrc.xerox.com

Abstract

The same word can have many different meanings depending on the context in which it is used. Discovering the meaning of a word, given the text around it, has been an interesting problem for both the psychology and the artificial intelligence research communities. In this article, we present a series of experiments, using methods which have proven to be useful for eliminating part-of-speech ambiguity, to see if such simple methods can be used to resolve semantic ambiguities. Using a publicly available semantic lexicon, we find the Hidden Markov Models work surprising well at choosing the right semantic categories, once the sentence has been stripped of purely functional words.

1 Introduction

Any natural language processing system treating anything beyond the most restricted domains is confronted with the problem of distinguishing between uses of polysemous words. The idea behind semantically tagging words is that sense markings added to words may be used by some automatic process in order to choose the proper senses of words in a given context. For example, the word *bark* would receive at least two possible semantic tags and these tags along with the tags of other words in the surrounding context would allow the process to distinguish between the senses the *bark* of a tree, and the *bark* of a dog. (See [Dagan and Itai, 1994; Gale *et al.*, 1992a; Gale *et al* , 1992b; Ng and Lee, 1996; Wilks, 1996; Yarowski, 1992; Yarowski, 1995] for recent work on word sense disambiguation).

Semantic tagging is considered to be a much more difficult task than part-of-speech tagging. Despite this current thinking, we decided to perform an experiment to see how well words can be semantically disambiguated using techniques that have proven to be effective in part-of-speech tagging. We decided to use the 45 semantic tags available through the WordNet package. In this typology, the word *bark* has two a priori semantic tags: *bark* as a ``covering, natural covering, cover'' receives tag 20 (nouns denoting plants); and *bark* as ``noise, cry'' has tag 11 (nouns denoting natural events). This semantic tagset has two advantages: it is a reasonable size, so that statistical techniques that we are testing do not need an inordinate amount of training data; and secondly, a semantically tagged corpus is available that we can use for testing.

2 WordNet Semantic tags

Part-of-speech tagging is better understood than semantic tagging. For one thing, no consensus on semantic tags exists, contrary to the general consensus on the higher level part-of-speech tags. And it seems more likely that syntactic tags be generalizable over wider textual domains than semantic ones.

Despite this, the WordNet team has taken upon themselves to create a general semantic tagging scheme and to apply it on a large scale: every set of synonymous senses, *synsets*, are tagged with one of 45 tags as WordNet version 1.5[1]. In their schema, there are 3 tags for adjectives (relational adjectives, participial adjectives and all others), 1 tag for all adverbs, 26 tags for nouns (act, animal, man-made artifact, attributes, body parts, ..., substance, and time), and 15 tags for verbs (from grooming and dressing verbs, to verbs of weather).

These tags are assigned for the most general uses of words. For example, the noun *blood* is tagged as 07 (an attribute of people and objects), as 08 (body part) and as 14 (groupings of people and objects). *Blood* is not tagged as 27 (substance) or as 13 (food), though it might well be considered as such in certain contexts.

[1] Ftp-able at clarity princeton edu

3 HMM Tagging

We wanted to see how well these WordNet semantic tags could be disambiguated using the same well-understood techniques employed in statistical part-of-speech disambiguation. Part-of-speech disambiguation relies on the fact that certain sequences of parts of speech are more probable than others. Often, this probability is estimated from the frequency of sequences of tags in hand tagged texts.

In our experiments, we used the *Hidden Markov Model* (HMM) tagging method described in [Cutting *et al.*, 1992]. In this method, the probability of seeing a given tag depends on the ambiguity class of the word and on the ambiguity class of the words preceding it. An *ambiguity class* of a word is the set of words which each have exactly the same set of ambiguous tags. This class is used during the Xerox HMM tagging in place of more specific lexical (= word-based) probabilities. Lexical probabilities would more accurately inform the tagger with the frequency with which a certain word receives a certain tag, but acquiring this frequency requires much greater amounts of tagged text than is necessary with the ambiguity class method. The HMM training and tagging programs in our experiment [Wilkens and Kupiec, 1995] are based on *bigrams*, i.e. only the immediate context of a word is taken into account.

The use of this statistical disambiguation combines with the advantage of the limited number of WordNet tags so that training can be performed on a relatively small corpus.

4 Data Preparation and Tagger Training

In order to make a HMM for semantic tags we performed the following steps:

1. We derived a *lexicon* from the WordNet data files which contains all possible semantic tags for each noun, adjective, adverb and verb. Words having no semantic tags (determiners, prepositions, auxiliary verbs, etc.) are assigned their part of speech tags.

2. With version 1.5 of WordNet is delivered about one-fifth of the Brown corpus which has been semantically tagged by the WordNet team. From these 11,182 sentences, we constructed a *training corpus* and a *test corpus* of equal size, taking all even numbered sentences for the training corpus and all odd-numbered sentences for the test corpus. From both corpora, in order to use ``semantically relevant'' tokens for the HMM bigrams, we retained all nouns, verbs, adverbs, and adjectives and deleted all function words except prepositions, commas, final stops, personal pronouns and interrogative adverbs.

3. We computed a HMM model based on the training corpus, ran the resulting semantic tagger on an untagged version of test corpus and we compared the tags assigned by the semantic tagger to original tags in the test corpus.

5 Tagging Results

5.1 Test 1

As described above, the semantically tagged text provided by WordNet (C0) was transformed into a training corpus (C1).

(C0) The/DT Fulton_County_Grand_Jury/03 said/32 Friday/28 an/DT investigation/09 of/IN Atlanta/15 's/POS recent/00 primary_election/04 produced/39 ``/`` no/DT evidence/09 "/" that/IN any/DT irregularities/04 took_place/30 ./.

(C1) Fulton_County_Grand_Jury/03 said/32 Friday/28 investigation/09 of/IN Atlanta/15 recent/00 primary_election/04 produced/39 evidence/09 that/IN irregularities/04 took_place/30 ./.

The lexicon used for this experiment contains 3,282 different ambiguity classes made of 52 semantic tags (45 WordNet tags + 6 part-of-speech tags + 1 tag for non-lexicalized word-forms).

The training corpus consists of 75,000 tokens and covers about 72% of all possible ambiguity classes. The test corpus contains 90,000 tokens. 46% of the words are ambiguous, i.e. the lexicon provides at least two (and at most 15) different semantic tags for these words.

For the test corpus the overall accuracy was of 86% and the accuracy over ambiguous tokens of 71% correctly chosen WordNet semantic tags.

5.2 Test 2

In fact, the first experiment combined syntactic and semantic tagging, as the WordNet tags are classified by part-of-speech categories.

Thus we run a second experiment which applies semantic tagging *after* part-of-speech tagging. We simulated the part-of-speech tagging step by adding a syntactic category to the training and test corpus:

(C3) Fulton_County_Grand_Jury=NOUN/03

said=VERB/32 Friday=NOUN/28
investigation/09 of/IN Atlanta=NOUN/15
recent=ADJ/00 primary_election=NOUN/04
produced=VERB/39 evidence=NOUN/09
that/IN irregularities=NOUN/04
took_place=VERB/30 ./.

We modified the lexicon accordingly. For example, a single lexicon entry for *bark* was divided into two entries for the verb and for the noun reading:

(L1)	bark	{06, 11, 20, 30, 32, 35}
(L2)	bark=VERB	{30, 32, 35}
	bark=NOUN	{06, 11, 20}.

Using part-of-speech pre-tagging, the number of ambiguity classes decreases (1685) and only 27% of the word forms in the test corpus are ambiguous.

With this method, the accuracy over the entire text is of 89%. This improvement is mainly due to the lower overall ambiguity rate: part-of-speech pre-tagging solved the ``semantic" ambiguity for 40% of the ambiguous words in Test 1. The error rate for those words which remain ambiguous after part-of-speech disambiguation is almost identical (71% correctly chosen tags) for both test cases.

5.3 Test 3

For the part-of-speech tagging problem, it is known that assigning the most common part of speech for each lexical item gives a baseline of 90% accuracy [Brill, 1992]. In order to see what a similar baseline is for semantic tagging over part-of-speech tagged text, we performed the following experiment. From the training corpus, we calculated the most frequent semantic tag for each part-of-speech tagged lemma[2]. On the test corpus, we assigned the most frequent semantic tag to each known word, and for unknown nouns, verbs, adverbs, and adjectives, we assigned the most common semantic tag per part-of-speech. Capitalized unknown nouns were assigned the S03 tag. Non-semantically tagged words were considered correctly tagged. The result of this tagging resulted in a baseline of 81% of correctly chosen semantic tags over all words, worse than the two preceding tests.

6 Discussion and Conclusion

We found it surprising that the same statistical techniques that improve part-of-speech tag disambiguation from a baseline of 90% to 95-96% work almost as well with semantic tags once function words are removed from the text to be tagged. The HMM technique improved the baseline 81% to 89% correctly chosen semantic tags. These experiments show renewed promise for a statistical approach to the problem of sense disambiguation, with a relatively small training set.

Future plans include analyzing the kind of errors we get, to classify them. Starting from this classification we hope to be able to answer the following questions: what type of semantic tags should be used, should a non-binary HMM be used, and how much ambiguity can be resolved using local clues.

We also plan to consider reasonable applications for semantic tagging. One possibility would be to use semantic tagging in the framework of an intelligent on line dictionary lookup such as LocoLex [Bauer *et al*, 1995]. LocoLex is a tool that has been developed at RXRC and which looks up a word in a bilingual dictionary taking the syntactic context into account. For instance, in a sentence such as *They like to swim* the part of speech tagger in LocoLex determines that *like* is a verb and not a preposition. Accordingly, the dictionary lookup component provides the user with the translation for the verb only. LocoLex also detects multi-word expressions[3]. For instance, when *stuck* appears in the sentence *my own parents stuck together* the translation displayed after the user clicks on *stuck* is the one for the whole phrase *stick together* and not only for the word *stick*.

Currently LocoLex is purely syntactic and cannot distinguish between the different meanings of a noun like *bark*. If, in addition to the current syntactic tags, we had access to the semantic tags provided by WordNet for this word (natural event and plants) and we were able to include this label in the online dictionary, this would improve the bilingual dictionary access of Locolex even further.

Current bilingual dictionaries often include some semantic marking. For instance, in the OUP-Hachette English French dictionary, under *bark* we find the label Bot(anical) attached to one meaning and the collocator (of dog) associated with the other one. It is possible that some type of automated matching between these indications and the WordNet semantic tags[4] would allow the integration of a semantic tagger into LocoLex.

Using only existing dictionary labels might still not be completely satisfying for machine translation. Indeed looking back at the example *my own parents stuck together* even if we retrieved the multi-word expression meaning it will be difficult to decide which translation to choose with existing dictionary indications[5].

---

2 Ties were resolved by randomly choosing one of the semantic tags.

3 Multi-words expressions include idiomatic expression (*to sweep something under the rug*), phrasal verbs (*to spit up*), or compounds (*warning light*)

4 Or some other derived tag set.

5 Especially considering that WordNet provides only two senses of *stick together* S35 and S41.

For instance for *stick together* the Oxford-Hachette English French dictionary gives:

stick together
1. (become fixed to each other)
(pages) se coller
2. (Coll) (remain loyal)
se serrer les coudes (Fam) être solidaire
3. (Coll) (not separate)
rester ensemble

It appears clearly that using general dictionary labels would not be enough to choose the third meaning only. We would need to investigate further how to make better use of dictionary information such as collocators, etc.

Another interesting application we would like to examine is how useful semantic tagging could be in determining the genre or topic of a text. Here, an initial idea would be to just count the number of occurrence of a given semantic tag and from this to determine the topic or the genre of a given text. This could be useful in machine translation system to help, for instance, in choosing the appropriate lexicon (containing the specific terminology). Assuming that such dictionaries are less ambiguous, this could in return, improve the accuracy of the lexical semantic choice in automatic translation.

References

[Bauer *et al*, 1995] Daniel Bauer, Frédérique Segond, Annie Zaenen. LOCOLEX : the translation rolls off your tongue. *Proceedings of ACH-ALLC95*. Santa-Barbara, USA, July 1995.

[Brill, 1992] Eric Brill. A simple Rule-Bases Part of Speech Tagger. *Proceedings of ANLP-92*. Trento, Italy, 1992.

[Cutting *et al*, 1992] Doug Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun. A Practical Part-of-speech Tagger. *Proceedings of ANLP-92*. Trento, Italy, 1992.

[Dagan and Itai, 1994] Dagan I. and Itai A. Word Sense Disambiguation Using a Second Language Monolingual Corpus. *Computational Linguistics*, 20-4, 563-596, 1994.

[Gale *et al*, 1992a] Gale, Church and Yarowsky. A Method for Disambiguating Word Senses in a Corpus. *Computers and the Humanities* 26, 415-439, 1992.

[Gale *et al*, 1992b] Gale, Church and Yarowsky. Using Bilingual Materials to Develop Word Sense Disambiguation Methods. *Proceedings of TMI-92*. 1992.

[Ng and Lee,1996] Hwee Tou Ng and Hian Beng Lee. Integrating Multiple Knowledge sources to disambiguate word sense. *Proceedings of ACL 96*. 1996.

[Wilkens and Kupiec, 1995] Mike Wilkens and Julian Kupiec. Training Hidden Markov Models for Part-of-speech Tagging. Internal document, Xerox Corporation. 1995.

[Wilks, 1996] Wilks Y. Oral communication. Courmayeur, 1996.

[Yarowski, 1992] Yarowsky D. Word Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. *Proceedings of COLING-92*. 1992.

[Yarowski, 1995] Yarowsky D. Unsupervised Word Sense Disambiguation Methods Rivaling Supervised Methods. *Proceedings of ACL-95*. 1995.