Towards Building Contextual Representations of Word Senses Using Statistical Models

Claudia Leacock,[1] Geoffrey Towell[2] and Ellen Voorhees[2]

[1]*Princeton University*
leacock@clarity.princeton.edu
[2]*Siemens Corporate Research, Inc.*
towell@learning.scr.siemens.com, ellen@learning.scr.siemens.com

Abstract

Automatic corpus-based sense resolution, or sense disambiguation, techniques tend to focus either on very local context or on topical context. Both components are needed for word sense resolution. A contextual representation of a word sense consists of topical context and local context. Our goal is to construct contextual representations by automatically extracting topical and local information from textual corpora. We review an experiment evaluating three statistical classifiers that automatically extract topical context. An experiment designed to examine human subject performance with similar input is described. Finally, we investigate a method for automatically extracting local context from a corpus. Preliminary results show improved performance.

1 Contextual Representations

The goal of automatic sense resolution is to acquire a *contextual representation* of word senses. A contextual representation, as defined by Miller and Charles [7], is a characterization of the linguistic contexts in which a word can be used. We look at two components of contextual representations that can be automatically extracted from textual corpora using statistical methods. These are *topical context* and *local context.*

Topical context is comprised of substantive words that are likely to co-occur with a given sense of a target word. If, for example, the polysemous word *line* occurs in a sentence with *poetry* and *write,* it is probably being used to express a different sense of *line* than if it occurred with *stand* and *wait.* Topical context is relatively insensitive to the order of words or their grammatical inflections; the focus is on the meanings of the open-class words that are used together in the same sentences.

Local context includes information on word order, distance and syntactic structure. For example, *a line from* does not suggest the same sense as *in line for.* Order and inflection are critical clues for local information, which is not restricted to open-class words.

In the next section, we briefly review an experiment using three statistical classifiers designed for sense resolution, and show that they are effective in extracting topical context. Section 3 describes an experiment that was performed to establish the upper bound of performance for these classifiers. Section 4 presents some techniques that we are developing to extract local context.

2 Acquiring Topical Context

Of the two types of context features, topical ones seem easier to identify. The idea is simple: for any topic there is a sub-vocabulary of terms that are appropriate for discussing

it. The task is to identify the topic, then to select that sense of the polysemous word that best fits the topic. For example, if the topic is writing, then *sheet* probably refers to a piece of paper; if the topic is sleeping, then it probably refers to bed linen; if the topic is sailing, it could refer to a sail; and so on.

Instead of using topics to discover senses, one can use senses to discover topics. That is to say, if the senses are known in advance for a textual corpus, it is possible to search for words that are likely to co-occur with each sense. This strategy requires two steps. It is necessary (1) to partition a sizeable number of occurrences of a polysemous word according to its senses, and then (2) to use the resulting sets of instances to search for co-occurring words that are diagnostic of each sense. That was the strategy followed with considerable success by Gale, Church, and Yarowsky [1], who used a bilingual corpus for (1), and a Bayesian decision system for (2).

To understand this and other statistical systems better, we posed a very specific problem: given a set of contexts, each containing the noun *line* in a known sense, construct a classifier that selects the correct sense of *line* for new contexts. To see how the degree of polysemy affects performance, we ran three- and six-sense tasks. A full description of the three-sense task is reported in Voorhees, *et. al.* [11], and the six-sense task in Leacock, *et. al* [5]. These experiments are reviewed briefly below.

We tested three corpus-based statistical sense resolution methods which attempt to infer the correct sense of a polysemous word by using knowledge about patterns of word co-occurrences. The first technique, developed by Gale *et. al.* [1] at AT&T Bell Laboratories, is based on Bayesian decision theory, the second is based on neural network with back propagation [9], and the third is based on content vectors as used in information retrieval [10]. The only information used by the three classifiers is co-occurrence of character strings in the contexts. They use no other cues, such as syntactic tags or word order, nor do they require any augmentation of the training and testing data that is not fully automatic. The Bayesian classifier uses all of the information in the sentence except word order. That is, it uses punctuation, upper/lower case distinctions, and inflectional endings. The other two classifiers remove punctuation and convert all characters to lower case. In addition, they remove a list of *stop words*, a set of about 570 very high frequency words that includes most function words as well as some content words. The remaining strings are *stemmed*: suffixes are removed to conflate across morphological distinctions. For example, the strings *computer(s), computing, computation(al)*, etc. are conflated to the stem *comput*.

2.1 Methodology

The training and testing contexts were taken from the 1987-89 *Wall Street Journal* corpus and from the APHB corpus.[1] Sentences containing *line(s)* and *Line(s)* were extracted and manually assigned a single sense from WordNet.[2] Sentences with proper names containing *Line*, such as *Japan Air Lines*, were removed from the set of sentences. Sentences containing collocations that have a single sense in WordNet, such as *product line* and *line of products*, were also excluded since the collocations are not ambiguous.

[1] The 25 million word corpus, obtained from the American Printing House for the Blind, is archived at IBM's T.J. Watson Research Center; it consists of stories and articles from books and general circulation magazines.

[2] WordNet is a lexical database developed by George Miller and his colleagues at Princeton University [6].

Typically, experiments have used a fixed number of words or characters on either side of the target word as the context. In these experiments, we used linguistic units – sentences – instead. Since the target word is often used anaphorically to refer back to the previous sentence, as in:

> That was the last time Bell ever talked on the *phone*. He couldn't get his wife off the *line*.

we chose to use two-sentence contexts: the sentence containing *line* and the preceding sentence. However, if the sentence containing *line* was the first sentence in the article, then the context consists of one sentence. If the preceding sentence also contained *line* in the same sense, then an additional preceding sentence was added to the context, creating contexts three or more sentences long. The average size of the training and testing contexts was 44.5 words.

The sense resolution task used the following six senses of the noun *line*:

1. a *product*: ... a new *line* of midsized cars ...
2. a *formation* of people or things: People waited patiently in long *lines* ...
3. spoken or written *text*: One winning *line* from that speech ...
4. a thin, flexible object; *cord*: With a *line* tied to his foot, ...
5. an abstract *division*: ... the Amish draw no *line* between work and religion and life.
6. a telephone *connection*: One key to WordPerfect's growth was its toll-free help *line*

The classifiers were run three times each on randomly selected training sets. The set of contexts for each sense was randomly permuted, with each permutation corresponding to one *trial*. For each trial, the first 200 contexts of each sense were selected as training contexts. The next 149 contexts were selected as test contexts. The remaining contexts were not used in that trial. The 200 training contexts for each sense were combined to form a final training set of size 1200. The final test set contained the 149 test contexts from each sense, for a total of 894 contexts. To test the effect that the number of training examples has on classifier performance, smaller training sets of 50 and 100 contexts were extracted from the 200 context training set.

2.2 Results

All of the classifiers performed best with the largest number (200) of training contexts, and the percent correct results reported here are averaged over the three trials with 200 training contexts. On the six-sense task, the Bayesian classifier averaged 71% correct answers, the content vector classifier 72%, and the neural networks 76%. None of these differences are statistically significant due to the limited sample size of three trials.

The ten most heavily weighted tokens for each sense for each classifier appear in Table 1. The words on the list seem, for the most part, indicative of the target sense and are reasonable indicators of topical context. However, there are some consistent differences among the methods. For example, while the Bayesian method is sensitive to proper nouns, the neural network appears to have no such preference. To test the hypothesis that the methods have different response patterns, we performed the χ^2 test for correlated proportions. This test measures how consistently the methods treat individual test contexts by determining whether the classifiers are making the same classification errors in each of the senses.

Product			Formation		
Bayesian	Vector	Network	Bayesian	Vector	Network
Chrysler	comput	comput	night	wait	wait
workstations	ibm	sell	checkout	long	long
Digital	produc	minicomput	wait	checkout	stand
introduced	corp	model	gasoline	park	checkout
models	sale	introduc	outside	mr	park
IBM	model	extend	waiting	airport	hour
Compaq	sell	acquir	food	shop	form
sell	introduc	launch	hours	count	short
agreement	brand	continu	long	peopl	custom
computers	mainframe	quak	driver	canad	shop
Text			**Cord**		
Bayesian	Vector	Network	Bayesian	Vector	Network
Biden	speech	familiar	fish	fish	hap
ad	writ	writ	fishing	boat	fish
Bush	mr	ad	bow	wat	wash
opening	bush	rememb	deck	hook	pull
famous	ad	deliv	sea	wash	boat
Dole	speak	fame	boat	float	rope
speech	read	speak	water	men	break
Dukakis	dukak	funny	clothes	dive	hook
funny	biden	movie	fastened	cage	exercis
speeches	poem	read	ship	rod	cry
Division			**Phone**		
Bayesian	Vector	Network	Bayesian	Vector	Network
blurred	draw	draw	phones	telephon	telephon
walking	fine	priv	toll	phon	phon
crossed	blur	hug	porn	call	dead
ethics	cross	blur	Bellsouth	access	cheer
narrow	walk	cross	gab	dial	hear
fine	narrow	fine	telephone	gab	henderson
class	mr	thin	Bell	bell	minut
between	tread	funct	billion	servic	call
walk	faction	genius	Pacific	toll	bill
draw	thin	narrow	calls	porn	silent

Table 1: Topical Context. The ten most heavily weighted tokens for each sense of *line* for the Bayesian, content vector and neural network classifiers.

The results of the χ^2 test for a three-sense resolution task (*product*, *formation* and *text*),[3] indicate that the response pattern of the content vector classifier is significantly different from the patterns of both the Bayesian and neural network classifiers, but the Bayesian response pattern is significantly different from the neural network pattern for the *product* sense only. In the six-sense disambiguation task, the χ^2 results indicate that the Bayesian and neural network classifiers' response patterns are not significantly different for any sense. The neural network and Bayesian classifiers' response patterns are significantly different from the content vector classifier only in the *formation* and *text* senses. Therefore, with the addition of three senses, the classifiers' response patterns appear to be converging.

A pilot two-sense distinction task (between *product* and *formation*) yielded over 90% correct answers.[4] In the three-sense distinction task, the three classifiers had a mean of 76% correct, yielding a sharp degradation with the addition of a third sense. Therefore, we hypothesized degree of polysemy to be a major factor for performance. We were surprised to find that in the six-sense task, all three classifiers degraded only slightly from the three-sense task, with a mean of 73% correct. Although the addition of three new senses to the task caused consistent degradation, the degradation is relatively slight. Hence, we conclude that some senses are harder to resolve than others, and it appears that overall accuracy is a function of the difficulty of the sense rather than being strictly a function of the degree of polysemy. The hardest sense for all three classifiers to learn was *text*, followed by *formation*, followed by *division*. The difficulty in training for the *product*, *phone*, and *cord* senses varied among the classifiers, but they were the three 'easiest' senses across the classifiers. To test our conclusion that the difficulty involved in learning individual senses is a greater factor for performance than degree of polysemy, we ran a three-way experiment on the three 'easy' senses. On this task, the content vector classifier achieved 90% accuracy and neural network classifier 92% accuracy.

The convergence of the response patterns for the three methods suggests that each of the classifiers is extracting as much data as is available in word co-occurrences in the training contexts. If this is the case, any technique that uses only word counts will not be significantly more accurate than the techniques tested here. Although the degree of polysemy does affect the difficulty of the sense resolution task, a greater factor for performance is the difficulty of resolving individual senses. From inspection of the contexts for the various senses, it appears that the senses of *line* that were easy to learn tend to be surrounded by a lot of topical context. With the senses that were hard to learn, the crucial disambiguating information tends to be very local, so that a greater proportion of the context is noise. Although it is recognized that local information is more reliable than distant information, the classifiers make no use of locality. Figure 1 shows some representative contexts for each sense of *line* used in the study. The *product*, *phone* and *cord* senses contain a lot of topical context, while the other senses have little or no information that is not very local.

The three classifiers are doing a good job finding topical context. However, simply knowing which words are likely to co-occur in the same sentences when a particular topic is under discussion is not sufficient for sense resolution.

[3] Training and test sets for these senses are identical to those in the six-sense resolution task.

[4] This task was only run with the content vector and neural network classifiers.

1. **text**: In a warmly received speech that seemingly sought to distance him from Reagan administration civil-rights policies, Mr. Bush outlined what he called a "positive civil-rights agenda," and promised to have "minority men and women of excellence as full-scale partners" during his presidency. One winning **line** from that speech: "Whenever racism rears its ugly head–Howard Beach, Forsyth County, wherever–we must be there to cut it off."

2. **formation**: On the way to work one morning, he stops at the building to tell Mr. Arkhipov: "Don't forget the drains today." Back in his office, the **line** of people waiting to see him has dwindled, so Mr. Goncharov stops in to see the mayor, Yuri Khivrich.

3. **division**: Thus, some families are probably buying take-out food from grocery stores–such as barbecued chicken–but aren't classifying it as such. The **line** between groceries and take-out food may have become blurred.

4. **cord**: Larry ignored the cries and came swooping in. The fisherman's nylon **line**, taut and glistening with drops of seawater, suddenly went slack as Larry's board rode over it.

5. **phone**: "Hello, Weaver," he said and then to put her on the defensive, "what's all the gabbing on the house phones? I couldn't get an open **line** to you."

6. **product**: International Business Machines Corp., seeking to raise the return on its massive research and development investments, said it will start charging more money to license its 32,000 patents around the world. In announcing the change, IBM also said that it's willing to license patents for its PS/2 **line** of personal computers.

Figure 1: Representative contexts for the six senses of *line* used in the study.

3 An Upper Bound For Classifier Performance

In an effort to establish an upper bound for performance on corpus-based statistical sense resolution methods, we decided to see how humans would perform on a sense resolution task using the same input that drives the statistical classifiers [4]. An experiment was designed to answer the following questions:

1. How do humans perform in a sense resolution task when given the same testing input as the statistical classifiers?

2. Are the contexts that are hard/easy for the statistical classifiers also hard/easy for people?

The three-sense task was replicated using human subjects. For each of the three senses of *line* (*product*, *text*, and *formation*), we selected 10 *easy* contexts (contexts that were correctly classified by the three statistical methods) and 10 *hard* contexts (contexts that were misclassified by the three methods), for a total of 60 contexts. These contexts were prepared in three formats: (1) a sentential form (as they originally appeared in the corpus), (2) a *long* list format (as was used by the Bayesian classifier), and (3) a *short* list format (as was used by the content vector and neural network classifiers). In order to mimic the fact that the classifiers do not use word order, collocations, or syntactic structure, the latter two contexts were presented to the subjects as word lists in reverse alphabetical order. 36 subjects each saw 60 contexts, 20 in each of the three formats, and were asked to choose the appropriate sense of *line*. The order in which the formats were presented was counter-balanced across subjects. No subject saw the same context twice. The subjects were Princeton undergraduates who were paid for their participation.

Human subjects performed almost perfectly on the sentential formats and had about a 32% error rate on the list formats. There was no significant difference between the two list formats – indicating that function words are of no use for sense resolution when word order is lost. They made significantly more errors on the contexts that were hard for the statistical classifiers, and fewer errors on the contexts that were easy for the classifiers. Not all the senses were equally difficult for human subjects: there were significantly fewer errors for the *product* sense of *line* than for the *text* and *formation* senses. Error rates for the subjects on the list formats were almost 50% for the hard contexts (contexts where the classifiers performed with 100% error), so subjects performed much better than the classifiers on these contexts. However, on the easy contexts, where the classifiers made no errors, the students showed an error rate of approximately 15%.

When subjects see the original sentences and therefore have access to all cues, both topical and local, they resolve the senses of *line* with 98% accuracy. When they are given the contexts in a list format, and are getting only topical cues, their performance drops to about 70% accuracy. Although their performance was significantly better than the classifiers (which all performed at 50% accuracy on this sample) human subjecrts are not able to disambiguate effectively using only topical context. From this result we conclude that in order to improve the performance of automatic classifiers, we need to incorporate local information into the statistical methods.

4 Acquiring Local Context

Kelly and Stone [3] pioneered research in finding local context by creating algorithms for automatic sense resolution. Over a period of seven years in the early 1970s, they (and some 30 students) hand coded sets of ordered rules for disambiguating 671 words. The rules include syntactic markers (part of speech, position within the sentence, punctuation, inflection), semantic markers and selectional restrictions, and words occurring within a specified distance before and/or after the target. An obvious shortcoming of this approach is the amount of work involved.

Recently there has been much interest in automatic and semi-automatic acquisition of local context (Hearst [2], Resnik [8], Yarowsky [13]). These systems are all plagued with the same problem, excellent precision but low recall. That is, if the local information that the methods learn is also present in a novel context, then that information is very reliable. However, quite frequently no local context match is found in a novel context. Given the sparseness of the local data, we hope to look for both local and topical context, and we have begun experimenting with various ways of acquiring the local context.

Local context can be derived from a variety of sources, including WordNet. The nouns in WordNet are organized in a hierarchical tree structure based on hypernomy/hyponomy. The hypernym of a noun is its superordinate, and the *is a kind of* relation exists between a noun and its hypernym. For example, *line* is a hypernym of *conga line*, which is to say that a *conga line* is a kind of *line*. Conversely, *conga line* is a hyponym of *line*. Polysemous words tend to have hyponyms that are monosemous collocations incorporating the polysemous word: *product line* is a monosemous hyponym of the merchandise sense of *line*; any occurrence of *product line* can be recognized immediately as an instance of that sense. Similarly, *phone line* is a hyponym of the telephone connection sense of *line*, *actor's line* is a hyponym of the text sense of *line*, *etc.* These collocational hyponyms provide a convenient starting point for the construction of local contexts for polysemous words.

We are also experimenting with template matching, suggested by Weiss as one approach to using local context to resolve word senses [12]. In template matching, specific word patterns recognized as being indicative of a particular sense (the templates) are used to select a sense when a template is contained in the novel context; otherwise word co-occurrence within the context (topical context) is used to select a sense. Weiss initially used templates that were created by hand, and later derived templates automatically from his dataset. Unfortunately, the datasets available to Weiss at the time were very small, and his results are inconclusive. We are investigating a similar approach using the *line* data: training contexts are used to both automatically extract indicative templates and create topical sense vectors.

To create the templates, the system extracts contiguous subsets of tokens including the target word and up to two tokens on either side of the target as candidate templates.[5] The system keeps a count of the number of times each candidate template occurs in all of the training contexts. A candidate is selected as a template if it occurs in at least n of the training contexts and one sense accounts for at least $m\%$ of its total occurrences. For example, Figure 2 shows the templates formed when this process is used on a training set of 200 contexts for each of six senses when $n = 10$ and $m = 75$. The candidate template *blurs the line* is not selected as a template with these parameter settings because it does not occur frequently enough in the training corpus; the candidate template *line of* is not

[5] In the template learning phase, tokens include punctuation and *stop words*. No stemming is performed and case distinctions are significant.

cord

his line

division

a fine line between, fine line between, a fine line, fine line
line between the, the line between, line between
draw the line, over the line

formation

a long line of, long line of, a long line, long line, long lines
in line for, wait in line, in line

phone

telephone lines
access lines
line was

product

a new line of, a new line, new line of, new line

Figure 2: Templates formed for a training set of 200 contexts for each of six senses when a template must occur at least 10 times and at least 75% of the occurrences must be for one sense. No templates were learned for the *text* sense.

selected because it appears too frequently in both the *product line* and *formation* contexts.

With the exception of *his line* (cord) and *line was* (phone), these templates readily suggest their corresponding sense. The $n = 10$ and $m = 75$ parameter settings are relatively stringent criteria for template formation, so not many templates are formed, but those templates that are formed tend to be highly indicative of the sense.

Preliminary results show template matching improves the performance of the content vector classifier. The six-sense experiment was repeated using a simple decision tree to incorporate the templates: The sense corresponding to the longest template contained in a test context was selected for that context; if the context contained no template, the sense chosen by the vector classifer was selected. The templates were automatically created from the same training set as was used to create the content vectors. To be selected as a template, a candidate had to appear at least 3 times for the training sets that included 50 of each sense, 5 times for the 100 each training sets, and 10 times for the 200 each training sets. In all cases, a single sense had to account for at least 75% of a candidate's occurrences. This hybrid approach was more accurate than the content vector classifier alone on each of the 9 trials. The average accuracy when trained using 200 contexts of each sense was 75% for the hybrid approach compared to 72% for the content vectors alone.

Other researchers have also suggested methods for incorporating local information into a classifier. Yarowsky found collocations[6] to be such powerful sense indicators that he suggests choosing a sense by matching on a set of collocations and choosing the most frequent sense if no collocation matches [13]. To resolve syntactic ambiguities, Resnik

[6]Yarowsky uses the term collocation to denote constructs similar to what we have called templates.

investigated four different methods for combining three sources of information [8]. The "backing off" strategy, in which the three sources of information were tried in order from most reliable to least reliable until some match was found (no resolution was done if no method matched), maintained high precision (81%) and produced substantially higher recall (95%) than any single method.

Our plans for incorporating templates into the content vector classifier include investigating the significance of the tradeoff between the reliability of the templates and the number of templates that are formed. When stringent criteria are used for template formation, and the templates are thought to be highly reliable sense indicators, the sense corresponding to a matched template will always be selected, and the sense vectors will be used only when no template match occurs. When the templates are thought to be less reliable, the choice of sense will be a function of the uniqueness of a matched template (if any) and the sense vector similarities. By varying the relative importance of a template match and sense vector similarity we will be able to incorporate different amounts of topical and local information into the template classifier.

5 Conclusion

The capacity to determine the intended sense of an ambiguous word is an important component of any general system for language understanding. We believe that, in order to accomplish this task, we need contextual representations of word senses containing both topical and local context. Initial experiments focused on methods that are able to extract topical context. These methods are effective, but topical context alone is not sufficient for sense resolution tasks. The human subject experiment shows that even people are not very good at resolving senses when given only topical context. Currently we are testing methods for learning local context for word senses. Preliminary results show that the addition of template matching on local context improves performance.

Acknowledgments

This work was supported in part by Grant No. N00014-91-1634 from the Defense Advanced Research Projects Agency, Information and Technology Office, by the Office of Naval Research, and by the James S. McDonnell Foundation. We are indebted to George A. Miller and Martin S. Chodorow for valuable comments on an earlier version of this paper.

References

[1] William Gale, Kenneth W. Church, and David Yarowsky. A method for disambiguating word senses in a large corpus. Statistical Research Report 104, AT&T Bell Laboratories, 1992.

[2] Marti A. Hearst. Noun homograph disambiguation using local context in large text corpora. In *Seventh Annual Conference of the UW Centre for the New OED and Text Research: Using Corpora*, pages 1–22, Oxford, 1991. UW Centre for the New OED and Text Research.

[3] Edward Kelly and Philip Stone. *Computer Recognition of English Word Senses.* North-Holland, Amsterdam, 1975.

[4] Claudia Leacock, Shari Landes, and Martin Chodorow. Comparison of sense resolution by statistical classifiers and human subjects. Cognitive Science Laboratory Report, Princeton University, in preparation.

[5] Claudia Leacock, Geoffrey Towell, and Ellen M. Voorhees. Corpus-based statistical sense resolution. In *Proceedings of the ARPA Workshop on Human Language Technology*, 1993.

[6] George Miller. Special Issue, WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4), 1990.

[7] George A. Miller and Walter G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1), 1991.

[8] Philip Resnik. Semantic classes and syntactic ambiguity. In *Proceedings of the ARPA Workshop on Human Language Technology*, 1993.

[9] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*, pages 318–363. MIT Press, Cambridge, 1986.

[10] G. Salton, A. Wong, and C.S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[11] Ellen M. Voorhees, Claudia Leacock, and Geoffrey Towell. Learning context to disambiguate word senses. In *Proceedings of the 3rd Computational Learning Theory and Natural Learning Systems Conference–1992*, Cambridge, to appear. MIT Press.

[12] Stephen Weiss. Learning to disambiguate. *Information Storage and Retrieval*, 9:33–41, 1973.

[13] David Yarowsky. One sense per collocation. In *Proceedings of the ARPA Workshop on Human Language Technology*, 1993.

THIS PAGE INTENTIONALLY LEFT BLANK