Computers and the Humanities **34:** 115–120, 2000.
© 2000 *Kluwer Academic Publishers. Printed in the Netherlands.*

A Topical/Local Classifier for Word Sense Identification

MARTIN CHODOROW[1], CLAUDIA LEACOCK[2] and GEORGE A. MILLER[3]
[1]*Department of Psychology, Hunter College of CUNY, 695 Park Avenue, New York, NY 10021, USA (E-mail: mschc@cunyvm.cuny.edu);* [2]*Department of Cognitive and Instructional Science, Educational Testing Service, Princeton, NJ 08541, USA (E-mail: cleacock@ets.org);*
[3]*Cognitive Science Laboratory, Princeton University, 221 Nassau Street, Princeton, NJ 08542, USA (E-mail: geo@clarity.princeton.edu)*

Abstract. TLC is a supervised training (S) system that uses a Bayesian statistical model and features of a word's context to identify word sense. We describe the classifier's operation and how it can be configured to use only topical context cues, only local cues, or a combination of both. Our results on Senseval's final run are presented along with a comparison to the performance of the best S system and the average for S systems. We discuss ways to improve TLC by enriching its feature set and by substituting other decision procedures for the Bayesian model. Future development of supervised training classifiers will depend on the availability of tagged training data. TLC can assist in the hand-tagging effort by helping human taggers locate infrequent senses of polysemous words.

Key words: disambiguation, Senseval, Bayesian classifier

1. Introduction

Our goal in developing TLC (a Topical/Local Classifier) was to produce a generic classifier for word sense disambiguation that uses publicly available resources and a standard Bayesian statistical model. We designed it to be flexible enough to incorporate topical context, local context, or a combination of the two. Topical context is comprised of the substantive words within the sentence. Local context consists of all words within a narrow window around the target. The next section gives a brief description of TLC's design and describes how it was used in Senseval. (A more detailed account of TLC can be found in Leacock, et al., 1998). Section 3 focuses on our treatment of multiword expressions and proper names for Senseval. In Section 4 we discuss our Senseval results, which are presented as ets-pu in Kilgarriff and Rosenzweig (in this volume). In Section 5 we suggest some ways to improve performance. In the final section, we describe an application of TLC in manual sense tagging.

2. Overview of the Classifier

A word sense classifier can be thought of as comprising a feature set and a decision procedure. Operationally, it can be viewed as a sequence of processing stages. Here we describe TLC's operation, the features it extracts and the decision procedure it employs.

TLC's operation consists of preprocessing, training, and testing. (For Senseval, an extra preprocessing step was used to locate targets that are multiword expressions. This will be described in Section 3.) During preprocessing, example sentences are tagged with part-of-speech (Brill,1994) and each inflected open-class word found in WordNet (Fellbaum, 1998) is replaced with its base form. These steps permit TLC to normalize across morphological variants while preserving inflectional information in the tags.

Training consists simply of counting the frequencies of the various contextual features (the cues) in each sense. When given a test sentence containing the polysemous word, TLC uses a Bayesian approach to find the sense s_i which is the most probable given the cues c_j contained in a context window of ± k positions around the polysemous target. For each s_i, the probability is computed with Bayes' rule:

$$p(s_i \mid c_{-k}, \ldots, c_k) = \frac{p(c_{-k}, \ldots, c_k \mid s_i)p(s_i)}{p(c_{-k}, \ldots c_k)}$$

Since the term $p(c_{-k},\ldots,c_k) \mid s_i)$ is difficult to estimate because of the sparse data problem, we assume, as is often done, that the occurrence of each feature is conditionally independent of the others, so that the term can be replaced with:

$$p(c_{-k}, \ldots, c_k \mid s_i) = \prod_{j=-k}^{k} p(c_j \mid s_i)$$

We can estimate $p\ (c_j \mid s_i)$ from the training data, but the sparse data problem affects these probabilities too, and so TLC uses the Good-Turing formula (Good, 1953; Chiang, et al., 1995), to smooth the values of $p(c_j \mid s_i)$, and provide probabilities for cues that did not occur in the training. TLC actually uses the mean of the Good-Turing value and the training-derived value, an approach that has yielded consistently better performance than relying on the Good-Turing values alone.

There are four types of contextual features that TLC considers: (1) topical cues consisting of open-class words (nouns, verbs, adjectives and adverbs) found in the Senseval context; (2) local open-class words found within a narrow window around the target; (3) local closed-class items (non-open-class words, e.g., prepositions and determiners); (4) local part of speech tags. The local windows do not extend beyond a sentence boundary. Procedures for estimating $p(c_j \mid s_i)$ and $p(c_j)$ differ somewhat for the various feature types.

1. The counts for open-class words (common nouns, verbs, adjectives, and adverbs) from which the topical word probabilities are calculated are not sensi-

tive to position anywhere within a wide window covering the entire example (the “bag of words” method). By contrast, the local cue probabilities do take relative position into account.

2. For open-class words found in the three positions to the left of the target (i.e., j = –3, –2, –1), $p(c_j \mid s_i)$ is the probability that word c_j appears in any of these positions. This permits TLC to generalize over variations in the placement of premodifiers, for example. Similarly, we generalize over the three positions to the right of the target. The window size of ± 3 was chosen on empirical grounds (Leacock et al., 1998).
3. Local closed-class items include determiners, prepositions, pronouns, and punctuation. For this cue type, $p(c_j \mid s_i)$ is the probability that item c_j appears precisely at location j for sense s_i. Positions j = –2, –1, 1, 2 are used. The global probabilities, for example $p(the_{-1})$, are based on counts of closed-class items found at these positions relative to the nouns in a large textual corpus.
4. Finally, part of speech tags in the positions j = –2, –1, 0, 1, 2 are used. The probabilities for these tags are computed for specific positions (e.g., $p(DT_{-1} \mid s_i)$, $p(DT_{-1})$) in the same way as in (3) above.

When TLC is configured to use only topical information, feature type (1) is employed. When it is configured for local information, types (2), (3), and (4) are used. Finally, in combined mode, the set of cues contains all four types.

We determined which of the three configurations was best for each Senseval item by dividing the training materials into two subsets, one was used for training TLC, the remainder for evaluating the performance of each configuration. We then used the best configuration of TLC in Senseval’s final run. For twenty-four of the items, this was the combined classifier, for ten it was the local configuration, and for two, the topical configuration.

3. Multiword dictionary expressions and Proper Names

During the development of TLC (Leacock et al., 1998), collocations (called multiword expressions in Senseval) were not included in the training/testing corpus – for the simple reason that collocations are usually monosemous. For example, if “rubber band” had only one sense in WordNet, the term was not included in the training or testing corpora.

We emulated this filtering procedure for Senseval as follows. When a multiword expression appeared as a head word in the Hector dictionary, we automatically generated a regular expression to match morphological and other variants, and searched the Senseval final-run corpus for the regular expression. For example, to find instances of “rubber band”, we searched for “/rubber band[s]?/” in the test corpus, and assigned any matches to the “rubber band” sense of “band”. TLC was not subsequently trained on that sense of band. As a result, if the regular expression match failed, test examples could not be assigned the correct sense. This procedure

Table I. Comparison of TLC to best and average S system performance on trainable words (fine-grained scoring).

part of	**TLC**		**Best S System**		**Mean of S Systems**	
speech	precision	recall	precision	recall	precision	recall
All	.756	(.755)	.771	(.771)	.733	(.657)
Nouns	.806	(.806)	.850	(.850)	.789	(.787)
Verbs	.709	(.709)	.709	(.709)	.687	(.686)
Adjectives	.744	(.743)	.761	(.761)	.724	(.723)
Multi-word	.785	(.704)	.907	(.906)	.757	(.682)
Proper name	.811	(.360)	.937	(.937)	.758	(.480)

worked surprisingly well. About 25 regular expressions were generated, matching almost 7% of the test sentences. Of these, 84% were correctly identified.

Other multiword expressions in the Hector dictionary are often listed as *kinds* or as *idioms* within a Hector word sense. For example, “jazz band” and “rock band” are *kinds* of one sense of band. Again, regular expressions were used to locate and assign a sense to these collocations. However, since many other kinds of bands, like “rhythm and blues band”, are subsumed under the same sense but are not explicitly specified as a *kind*, the classifier was also trained in the usual way on that sense. This meant that even if the regular expression match failed, the correct sense might still be identified based on TLC’s cues.

In developing TLC we did not consider proper names, again because they are not polysemous. Proper name identification is a field unto itself, and our working assumption has been that a proper name filter would be applied to text prior to TLC’s operation. Since we do not have such a filter as part of TLC’s preprocessing, the proper names in Senseval were treated as separate senses, with training performed on each independently.

4. Results

We used TLC to assign senses to the 36 trainable words only. Features were extracted from the supervised training materials, but the definitions and example sentences provided in the Hector dictionary were not used. However, as described in Section 3, we did filter the collocations listed as *kinds* or *idioms* during preprocessing.

The results indicate that TLC’s precision increased with size of the training data (Pearson correlation coefficient $r = 0.33, p < 0.05$, two-tailed), but there was no significant effect of the number of senses ($r = -0.15, p > 0.10$). As expected for a Bayesian classifier, its performance was strongly affected by item entropy ($r = -0.63, p < 0.01$).

Table I shows the classifier’s performance over all trainable words when scored by the fine-grained method. It also lists the results by nouns, verbs, adjectives, multi-word expressions, and proper nouns. The data are taken from the final Senseval run and are designated ets-pu in the main summary tables. For purposes of comparison, Table I also gives performance figures for the best supervised training (S) system, as well as the mean for all S systems.

5. System Improvements

Most classifiers consist of two independent components: the statistical model and the set of features they manipulate. Ideally, these two should be prised apart and evaluated independently of one another.

For example, TLC’s Bayesian model assumes conditional independence of the features, which is clearly a false assumption. Other models for word sense disambiguation that do not assume independence are emerging, such as maximum entropy and TiMBL (Veenstra et al., this volume). It is quite possible that replacing the Bayesian model with one of these would improve the classifier’s overall performance.

It is also likely that the feature set TLC uses can be improved. For example, it currently uses the Penn Treebank part-of-speech tag sets. Recently, enriched tags that encode configurational information such as supertags (Joshi and Srinivas, 1994) are being developed, and might also improve the system’s performance.

6. A Sense-Tagging Application

Miller, et al. (submitted) are currently preparing a hand-tagged corpus for several hundred common words of English, as a resource for future development of statistical classifiers. Preparation of these materials is time-consuming and labor-intensive, in part because many words have secondary senses that are so infrequent that it is difficult to find examples, except by sifting through hundreds of cases of the primary sense. For instance, in every 100 occurrences of “bank”, 78 are likely to be examples of the “financial institution” sense, with the remaining 22 representing the other 8 senses. We wondered if TLC could perform a pre-screening function by flagging many examples of the primary sense, and in this way save the human taggers much time and effort. As an experiment, for each of eight words that have a single salient sense, we trained TLC on this sense and on the union of all the other senses of the word so that the classifier could score new examples in terms of “primary” sense and “other”. When we looked at examples that were classified as high probability primary, low probability other, there were very few misclassifications. This screening procedure should speed up the tagging process by allowing human taggers to concentrate their efforts on sentences in which a non-primary sense is more likely to be used. We hope that in this way, by assisting in the manual tagging of training corpora, TLC can contribute to the future development of all supervised training systems, including its own.

References

Brill, E. “Some advances in rule-based part of speech tagging” *Proceedings of the Twelfth National Conference on Artificial Intelligence*, Seattle: AAAI. 1994.

Chiang T-H., Y-C. Lin and K-Y Su. “Robust learning, smoothing, and parameter tying on syntactic ambiguity resolution”, *Computational Linguistics*, Vol. no. 21–3, 1995, pp. 321–349.

Fellbaum, C. (ed). *WordNet: An Electronic Lexical Database*, Cambridge: MIT Press. 1998.

Good, I. F. “The population frequencies of species and the estimation of population parameters”, *Biometrica*, Vol. no. 40, 1953, pp. 237–264.

Joshi, A.K., B. Srinivas. “Disambiguation of Super Parts of Speech (or Supertags): Almost parsing”, *Proceedings of COLING 1994*, 1994, pp. 154–160.

Leacock, C., M. Chodorow and G. A. Miller. “Using corpus statistics and WordNet relations for sense identification”, *Computational Linguistics*, Vol. no. 24-1, 1998, pp. 147–165.

Miller, G. A., R. Tengi and S. Landes (submitted for publication). “Matching the Tagging to the Task”.