Category-Based Pseudowords

Preslav I. Nakov
EECS, UC Berkeley
Berkeley, CA 94720
nakov@cs.berkeley.edu

Marti A. Hearst
SIMS, UC Berkeley
Berkeley, CA 94720
hearst@sims.berkeley.edu

Abstract

A pseudoword is a composite comprised of two or more words chosen at random; the individual occurrences of the original words within a text are replaced by their conflation. Pseudowords are a useful mechanism for evaluating the impact of word sense ambiguity in many NLP applications. However, the standard method for constructing pseudowords has some drawbacks. Because the constituent words are chosen at random, the word contexts that surround pseudowords do not necessarily reflect the contexts that real ambiguous words occur in. This in turn leads to an optimistic upper bound on algorithm performance. To address these drawbacks, we propose the use of lexical categories to create more realistic pseudowords, and evaluate the results of different variations of this idea against the standard approach.

1 Introduction

In order to evaluate a word sense disambiguation (WSD) algorithm in a new language or domain, a sense-tagged evaluation corpus is needed, but this is expensive to produce manually. As an alternative, researchers often use *pseudowords*. To create a pseudoword, two or more randomly-chosen words (e.g., *banana* and *door*) are selected and their individual occurrences are replaced by their conflation (e.g., *banana-door*). Since their introduction (Gale et al., 1992; Schuetze, 1992), pseudowords have been accepted as an upper bound of the true accuracy of algorithms that assign word sense distinctions.

In most cases, constituent words are chosen entirely randomly. When used to evaluate a real WSD system on the SENSEVAL1 corpus, pseudowords were found to be optimistic in their estimations compared to real ambiguous words with the same distribution (Gaustad, 2001). Real ambiguous words often have senses that are similar in meaning, and thus difficult to distinguish (as measured by low inter-annotator agreement), while pseudowords, because they are randomly chosen, are highly likely to combine semantically distinct words. Another drawback is that the results produced using pseudowords are difficult to characterize in terms of the types of ambiguity they model.

To create more plausibly-motivated pseudoword pairings, we introduce the use of lexical category membership for pseudoword generation. The main idea is to take note of the relative frequencies at which pairs of lexical categories tend to represent real ambiguous words, and then use unambiguous words drawn from those categories to generate pseudowords. In the remainder of this paper we describe the category-based pseudoword generation process and evaluate the results against the standard methods and against a real word sense disambiguation task.

2 MeSH and Medline

In this paper we use the MeSH (Medical Subject Headings) lexical hierarchy[1], but the approach should be equally applicable to other domains using other thesauri and ontologies. In MeSH, each concept is assigned one or more alphanumeric descriptor codes corresponding to particular positions in the hierarchy. For example, A (Anatomy), A01 (Body Regions), A01.456 (Head), A01.456.505 (Face), A01.456.505.420 (Eye). *Eye* is ambiguous according to MeSH and has a second code: A09.371 (A09 represents Sense Organs).

In the studies reported here, we truncate the MeSH code at the first period. This allows for generalization over different words; e.g., for *eye*, we discriminate between senses represented by A01 and A09. This truncation reduces the average number of senses per token from 2.12 to 1.39, and the maximum number of ambiguity classes for a given word to 7; 71.18% of the tokens have a single class and 22.14% have two classes. From a collection of 180,226 abstracts from Medline 2003[2],

[1] http://www.nlm.nih.gov/mesh

[2] 235 MB of plain text, after XML removal, from files med-

Ambig. pair	Pair freq.	Class 1 freq.	Class 2 freq
{A11,A15}	16127	49350	3417
{A12,A15}	13662	7403	3417
{D12,D24}	12608	28805	17064
{E05,H01}	11753	17506	40744
{I01,N03}	6988	7721	11046
{A02,A10}	6834	4936	14083

Table 1: Most frequent ambiguous 2-category pairs.

w_1	w_2	pair	$\#w_1$	$\#w_2$
artifact	triton	{E05,H01}	55	40
humerus	mucus_memb.	{A02,A10}	51	38
lovastatin	palmitic_acid	{D04,D10}	35	54
child_abuse	Minnesota	{I01,Z01}	39	45
thumb	pupils	{A01,A09}	56	38
haptoglobin	hla_antigens	{D12,D24}	46	53

Table 2: Sample pseudowords.

training was done on 2/3 of the abstracts (120,150) and testing on the remaining 1/3 (60,076).

3 Pseudoword Generation

For the creation of pseudowrods with two-sense ambiguities, we first determined which ambiguous words fall into exactly two MeSH categories and built a list L of pairs (see Table 1). We then generated pseudowords with the following characteristics:

- The two possible pseudoword categories represent a pair that is really seen in the testing corpus and thus needs to be disambiguated;
- The number of pseudowords drawn from a particular pair is proportional to its frequency;
- Multi-word concepts can be used as pseudoword elements: e.g., *ion-exchange chromatography* and *long-term effects* can be conflated as *ion-exchange_chromatography_long-term_effects*
- Only unambiguous words are used as pseudoword constituents.

An important aspect of pseudoword creation is the relative frequencies of the underlying words. Since the standard baseline for a WSD algorithm is to always choose the most frequent sense, a baseline that is evaluated on words whose senses are evenly balanced will be expected to do more poorly than one tested against words that are heavily skewed towards one sense (Sanderson & van Rijsbergen, 1999).

In naturally occurring text, the more frequent sense for the two-sense distinction is reported to occur 92% of the time on average; this result has been found both on the CACM collection and on the WordNet SEMCOR sense-tagged corpus (Sanderson & van Rijsbergen, 1999). However, the challenge for WSD programs is to work on the harder cases, and the artificially constructed SENSEVAL1 corpus has more evenly distributed senses (Gaustad, 2001).

In these experiments, we explicitly compare pseudowords whose underlying word frequencies are even against those that are skewed. To generate pseudowords with more uniform underlying distributions, we first calculate the expected testing corpus frequency of those words w_i that have been unambiguously mapped to MeSH and whose class is used in at least one pair in L. In this collection the expected frequency was E = 45.21 with a standard deviation of 451.19. We then built a list W of all MeSH concepts mapped in the text that have a class used in a pair in L and whose frequency is in the interval [E/2;3E/2], i.e. [34;56]. This yields a list of concepts that could potentially be combined in 64,596 pseudowords for evaluation of the WSD algorithm performance over the classes in L.

We then generated a random subset of 1,000 pseudowords (88,758 instances) out of the possible 64,596 by applying the following importance sampling procedure:

1) Select a category pair c_1,c_2 from L by sampling from a multinomial distribution whose parameters are proportional to the frequencies of the elements of L.

2) Sample uniformly to draw two random distinct words w_1 and w_2 from W whose classes correspond to the classes selected in step 1).

3) If the word pair w_1,w_2 has been sampled already, go to step 1) and try again.

Table 2 shows a random selection of pseudowords generated by the algorithm. Note that the more unusual pairings come from the less frequent category pairs, whereas those in which word senses are closer in meaning are drawn from more common category pairs.

4 Results

For the experiments reported below, we trained an unsupervised Naive Bayes classifier using the categories as both targets and as context features. For example, an occurrence of the word *haptoglobin* in the context surrounding the word to be disambiguated will be replaced by its category label D12. Only unambiguous context words were used. The result of the disambiguation step is a category name, standing as a proxy for the word sense.

Table 3 reports accuracies for several experiments in terms of macroaverages (average over the individual accuracies for each pseudoword). *Baseline* refers to choos-

line03n0201.xml through med-line03n0209.xml.

CW	Base.	Pess.	Real.	Abbrev.	Opt.
10	53.24	62.93	64.60	70.37	71.35
20	53.24	66.80	68.90	73.83	76.36
40	53.24	69.92	73.28	76.46	80.03
300	53.24	72.79	75.34	77.99	81.88

Table 3: Accuracies (in %'s) of Baseline, Pessimistic, Realistic, Abbreviation, and Optimistic datasets for different context window (CW) sizes.

AAP:	acetaminophen	D02
	auricular_acupuncture	E02
GST:	general_systems_theory	H01
	glutathione_s-transferase	D08
ED:	eating_disorders	F03
	endogenous_depression	F03
	elemental_diet	J02

Table 4: Sample category mappings for abbreviations.

ing the most frequent sense[3]. *Pessimistic* refers to the evenly distributed category-based pseudowords, generated by requiring the word frequency to fall in the interval [E/2;3E/2]. In the column labeled *Realistic*, the requirement for evenly distributed senses is dropped, although the component words must have a frequency of at least 5. The column labeled *Optimistic* refers to the results when the pseudowords are generated the standard way: the words are selected at random rather than according to the category sets.

We expected the *Realistic* pseudowords to produce a better lower-bound estimate of the performance of a WSD algorithm on real word senses than *Optimistic*. To test this hypothesis we followed a method suggested by Liu et al. (2002) and evaluated the classifier on a set of 217 two-sense abbreviations (see Table 4).

Abbreviations are real ambiguous words, but they are also artificial in a sense. Many homonyms are similar in meaning as well as spelling because they derive etymologically from the same root. By contrast, similar spelling in abbreviations is often simply an accident of shared initial characters in compound nouns. Thus abbreviations occupy an intermediate position between entirely random pseudowords and standard real ambiguous words.

We extracted 98,841 unique abbreviation-expansion pairs[4] using code developed by Schwartz & Hearst (2003), and retained only those abbreviations whose expansions could be fully and unambiguously mapped to a single truncated MeSH category. The different expansions of each abbreviation were required to correspond to exactly two distinct categories (with overlap allowed when there were more than two expansions for a given abbreviation).

The question we wanted to explore is how well does the classifier do on category-based pseudowords versus abbreviations. As can be seen from Table 3, the accuracies for the abbreviations (evaluated on 332,020 instances) fall between the *Realistic* and *Optimistic* pseudowords, as expected.

5 Conclusions

We have shown that creating pseudowords based on distributions from lexical category co-occurrence can produce a more accurate lower-bound for WSD systems that use pseudowords than the standard approach. This method allows for the detailed study of a particular sense ambiguity set since many different pseudowords can be generated from one category pair. Additionally, this method provides a better-motivated basis for the grouping of words into pseudowords, since they more realistically model the meaning similarity patterns of real ambiguous words than do randomly paired words.

Acknowledgements Special thanks to Barbara Rosario for the discussions and valuable suggestions and to Ariel Schwartz for providing the abbreviation extraction code. This work was supported by a gift from Genentech and an ARDA Aquaint contact.

References

William A. Gale, Kenneth W. Church and David Yarowsky. 1992. *Work on statistical methods for word sense disambiguation.*, In R. Goldman et al. (Eds.), Working Notes of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language, 54-60.

Tanja Gaustad. 2001. *Statistical Corpus-Based Word Sense Disambiguation: Pseudowords vs. Real Ambiguous Words.*, Proc. 39th Annual Meeting of ACL (ACL/EACL 2001) - Student Research Workshop.

Hongfang Liu, Stephen B. Johnson and Carol Friedman. 2002. *Automatic Resolution of Ambiguous Terms Based on Machine Learning and Conceptual Relations in the UMLS*, JAMIA 2002.

Mark Sanderson and Keith van Rijsbergen. 1999. *The impact on retrieval effectiveness of skewed frequency distributions.*, TOIS 17(4): 440-465.

Hinrich Schuetze. 1992. *Context space.*, In R. Goldman et al. (Eds.), Working Notes of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language, 54-60.

Ariel Schwartz and Marti Hearst. 2003. *A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text.*, In Proceedings of the Pacific Symposium on Biocomputing (PSB 2003) Kauai, Jan 2003.

[3]The baseline is dependent on the (pseudo)words used. The one shown is the baseline for the abbreviations collection.

[4]From med-line03n0210.xml to med-line03n0229.xml.