Noun Sense Induction Using Web Search Results

Goldee Udani, Shachi Dave, Anthony Davis and Tim Sibley
StreamSage Inc.
1016 16th St NW, Suite 200
Washington DC 20036
{udani, shachi, davis, sibley}@streamsage.com

ABSTRACT
This paper presents an algorithm for unsupervised noun sense induction, based on clustering of Web search results. The algorithm does not utilize labeled training instances or any other external knowledge source. Preliminary results on a small dataset show that this technique provides two advantages over other techniques in the literature: it detects real-world senses not found in dictionaries or other lexical resources, and it does not require that the number of word senses be specified in advance.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Clustering.*

General Terms
Algorithms, Experimentation

Keywords
Noun Sense Induction, Disambiguation, Clustering, Evaluation

1. INTRODUCTION
Word sense induction refers to the process of discovering different senses of an ambiguous word without prior information about the inventory of senses. Thus, it is an example of an unsupervised learning problem, which we distinguish from word sense disambiguation (WSD), the task of choosing the correct sense of an ambiguous word from among two or more predefined alternative senses. These senses usually come from a pre-compiled lexicon such as WordNet, a dictionary, or a thesaurus. However, many terms are frequently used in ways that these lexical resources cannot be expected to cover. For example, *patriot* is defined in the sense of *nationalist* in WordNet, whereas it has at least three other major senses {*Patriot act, Patriot missile* and the movie – *The Patriot*} on the Web. If major real-world senses are missing from the reference lexicon, disambiguation is doomed to fail when applied to Web corpora. In addition, many lexical resources contain overly general word senses, unsuited to WSD for search and retrieval.

Some researchers have already explored unsupervised techniques that use context surrounding the ambiguous word in a corpus. For example, [3] and [4] use Singular Value Decomposition as a dimensionality reduction technique for a high-dimensional word space consisting of local and global features. Yarowsky [5] uses a bootstrapping approach involving generalization from a small number of labeled instances. (The system discussed in [3] requires as input an upper bound on the number of word senses to induce.) Some recent work in unsupervised WSD uses Web search results [2], but in a procedure that uses WordNet to bootstrap the induction process. This technique is thus subject to the same weaknesses as other lexicon-based systems (in particular, a predetermined, and not necessarily optimal inventory of senses). This project's noun sense induction technique is based on Web search results, as they appear to be good indicators of the type of corpora available on the Web. Search engine results are thus likely to be better predictors of which senses users are likely to search on and encounter.

SIGIR'05, August 15–19, 2005, Salvador, Brazil.
ACM 1-59593-034-5/05/0008.

2. METHODOLOGY

2.1 Feature Selection
The system first fetches 500 Google search results (title and snippets only) for the ambiguous word. The search results are lemmatized and tagged using MontyLingua [1].

Previous research ([3] and [5]) demonstrates that local features such as unigrams and bigrams are often excellent disambiguators. Since search results consisting of one-or-two-line snippets and a title provide little context, we use bigrams to reduce the sparsity.

The procedure extracts noun and adjective unigrams, and bigrams consisting of two consecutive content words, one of which is the ambiguous word. Stop-words – prepositions, determiners, conjunctions and other very common words – are filtered out, as well as bigrams containing stop-words. After filtering, the system constructs a feature vector from the remaining unigrams and bigrams.

After some experimentation, we found that noun-noun and adjective-noun bigrams are the best indicators of context for an ambiguous word.

2.2 Clustering
In the first pass, to create base clusters, the algorithm combines all search results that have bigrams in common. (This step is based on the *one sense per collocation* criterion suggested by [5].) The weight assigned to each word feature in a base cluster is proportional to its TF-IDF measure, where the number of documents corresponds to the number of search results retrieved.

The algorithm then merges clusters by agglomerative clustering; it calculates the cosine similarities between all base cluster vectors, and keeps merging the most similar clusters until the similarity between the clusters drops below a particular threshold

(0.1). When two clusters are merged, their feature vectors are combined and their word weights are updated.

Finally, those clusters with weights that fall below the average total weight of all clusters are discarded, because they are generally either incoherent or too fine-grained. Each of the remaining clusters is assumed to represent a unique sense of the ambiguous word. The feature vector of each final cluster constitutes the context vector for that sense. The weight associated with each term in a word sense's context vector indicates the importance of that term in discriminating that sense.

3. EXPERIMENTS

3.1 Induced Senses from Context Vectors

We have tested this technique on a set of 25 ambiguous nouns. Five nouns with their dictionary senses and the senses induced by the system are shown in Table 1. The granularity of the senses varies, depending on how evenly the search engine results represent these senses. The senses discovered are listed in a descending order of their weights. Minor (italicized) senses in Table 1 are those constituting a relatively small proportion of the word's occurrences.

Table 1. Senses discovered for ambiguous nouns

Query	Dictionary Senses	Major/*Minor* Senses
Patriot	Nationalist	Patriot Act, Patriot Missile, Patriot (Movie), Patriot (Game)
Cell	Prison Cell, Biology Cell, Fuel Cell, Mobile, Spreadsheet Cell, Electrochemical Cell	Plant/Animal Cell, Cell Phone, Stem Cell, Fuel Cell
Crane	Construction crane, Bird	Construction Crane, Stephen Crane (Writer), Paper Crane (Origami), *Harold Hart Crane (Poet), Whooping Crane (Bird)*
Tank	Gas Tank, Combat Vehicle, Tank Top, Water tank	Storage Tank, Army Tank (Vehicle), Think Tank, Tank War (Video Game), *Tank Girl (Movie), Thomas the Tank Engine (Toy)*
Gate	Passageway, Logic Gate	Golden Gate (San Francisco), Lions Gate (Film Company), *City Gate*

3.2 Evaluation

To evaluate the resulting context vectors, we manually constructed a search query incorporating the ambiguous word and its most discriminating related word(s) for each major word sense found. We randomly chose 50 results from the first 500 search results for that query. (Some queries returned fewer than 500 search results.) These search results were then presented in random order to the disambiguation system.

System accuracy was defined as the percentage of search results that were correctly categorized by the system. The baseline accuracy was defined as the percentage of results correctly categorized by classifying all results into the most frequent sense. Table 2 shows the baseline accuracy and system accuracy for each word from Table 1.

Table 2. Baseline and system accuracies

Ambiguous word	Baseline Accuracy (%)	System Accuracy (%)
Patriot	25	93.5
Cell	25	81.0
Crane	33	80.6
Tank	25	78.5
Gate	50	95.0
Average	31.6	85.72

We believe that the comparatively low system accuracy for *cell* is because *stem cell* and *plant/animal cell* are essentially fine-grained senses of *biological cell*. These are closely related senses, hence they are easily confused by our system. Similarly, for *tank*, the *Tank War (video game)* and *army tank* senses share several context words.

4. CONCLUSION AND FUTURE WORK

We have presented a novel approach to noun sense induction using an on-the-fly, unsupervised clustering algorithm operating on Web search results. Our approach eliminates the need for any external knowledge source. Preliminary results for nouns show that the system is 85.7% accurate for disambiguating search results on the Web. We plan to run experiments on a larger set of nouns. One natural extension of this system is to use additional bigram templates, which should aid in inducing senses for ambiguous verbs and adjectives. This technique can also be extended to work on other languages for which reliable part-of-speech tagging, and morphological analysis are available. Another avenue would be to induce a human readable sense definition automatically.

5. REFERENCES

[1] Liu, H. *MontyLingua: Commonsense-Informed Natural Language Understanding Tools (2003).* Available at: http://web.media.mit.edu/~hugo/montylingua

[2] Martinez, D. and Agirre, E. The effect of bias on an automatically-built word sense corpus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluations (LREC-04)*(Lisbon, Portugal, May 24-30, 2004)

[3] Purandare, A. and Pedersen, T. SenseClusters - Finding Clusters that Represent Word Senses. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)* (San Jose, USA, July 25-29, 2004)

[4] Schütze, H. Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97-123, 1998.

[5] Yarowsky, D. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189-196, 1995.