Automatically Discovering Word Senses

Patrick Pantel and **Dekang Lin**
Department of Computing Science
University of Alberta
Edmonton, Alberta T6G 2E8 Canada
{ppantel, lindek}@cs.ualberta.ca

Abstract

We will demonstrate the output of a distributional clustering algorithm called Clustering by Committee that automatically discovers word senses from text[1].

1 Introduction

Using word senses versus word forms is useful in many applications such as information retrieval (Voorhees 1998), machine translation (Hutchins and Sommers 1992), and question-answering (Pasca and Harabagiu 2001).

The Distributional Hypothesis (Harris 1985) states that words that occur in the same contexts tend to be similar. There have been many approaches to compute the similarity between words based on their distribution in a corpus (Hindle 1990; Landauer and Dumais 1997; Lin 1998). The output of these programs is a ranked list of similar words to each word. For example, Lin's approach outputs the following similar words for *wine* and *suit*:

```
wine: beer,    white    wine,    red    wine,
      Chardonnay,  champagne,  fruit,  food,
      coffee,   juice,   Cabernet,   cognac,
      vinegar, Pinot noir, milk, vodka,…
suit: lawsuit, jacket, shirt, pant, dress,
      case, sweater, coat, trouser, claim,
      business suit, blouse, skirt, litiga-
      tion, …
```

The similar words of *wine* represent the meaning of wine. However, the similar words of *suit* represent a mixture of its *clothing* and *litigation* senses. Such lists of similar words do not distinguish between the multiple senses of polysemous words.

We will demonstrate the output of a distributional clustering algorithm called Clustering by Committee (CBC) that discovers word senses automatically from text. Each cluster that a word belongs to corresponds to a sense of the word. The following is a sample output from our algorithm:

```
(suit
   0.39 (blouse, slack, legging, sweater)
   0.20 (lawsuit, allegation, case, charge)
)
(plant
   0.41 (plant, factory, facility,
         refinery)
   0.20 (shrub, ground cover, perennial,
         bulb)
)
(heart
   0.27  (kidney,   bone   marrow,   marrow,
         liver)
   0.17 (psyche, consciousness, soul, mind)
)
```

Each entry shows the clusters to which the headword belongs along with its similarity to the cluster. The lists of words are the top-4 most similar members to the cluster centroid. Each cluster corresponds to a sense of the headword.

2 Feature Representation

Following (Lin 1998), we represent each word by a feature vector. Each feature corresponds to a context in which the word occurs. For example, "sip __" is a verb-object context. If the word *wine* occurred in this context, the context is a feature of *wine*. These features are obtained by parsing a large corpus using Minipar (Lin 1994), a broad-coverage English parser. The value of the feature is the pointwise mutual information (Manning and Schütze 1999) between the feature and the word. Let c be a context and $F_c(w)$ be the frequency count of a word w occurring in context c. The pointwise mutual information, $mi_{w,c}$, between c and w is defined as:

[1] The demonstration is currently available online at www.cs.ualberta.ca/~lindek/demos/wordcluster.htm.

$$mi_{w,c} = \frac{\frac{F_c(w)}{N}}{\frac{\sum_i F_i(w)}{N} \times \frac{\sum_j F_c(j)}{N}}$$

where N is the total frequency counts of all words and their contexts. We compute the similarity between two words w_i and w_j using the *cosine coefficient* (Salton and McGill 1983) of their mutual information vectors:

$$sim(w_i, w_j) = \frac{\sum_c mi_{w_ic} \times mi_{w_jc}}{\sqrt{\sum_c mi_{w_ic}^2 \times \sum_c mi_{w_jc}^2}}$$

3 Clustering by Committee

CBC finds clusters by first discovering the underlying structure of the data. It does this by searching for sets of representative elements for each cluster, which we refer to as committees. The goal is to find committees that unambiguously describe the (unknown) target classes. By carefully choosing committee members, the features of the centroid tend to be the more typical features of the target class. For example, our system chose the following committee members to compute the centroid of the state cluster: Illinois, Michigan, Minnesota, Iowa, Wisconsin, Indiana, Nebraska and Vermont. States like Washington and New York are not part of the committee because they are polysemous. The centroid of a cluster is constructed by averaging the feature vectors of the committee members.

CBC consists of three phases. Phase I computes each element’s top-*k* similar elements. In Phase II, we do a first pass through the data and discover the committees. The goal is that we form tight committees (high intra-cluster similarity) that are dissimilar from one another (low inter-cluster similarity) and that cover the whole similarity space. The method is based on finding sub-clusters in the top-similar elements of every given element.

In the final phase of the algorithm, each word is assigned to its most similar clusters (represented by a committee). Suppose a word *w* is assigned to a cluster *c*. We then remove from *w* its features that intersect with the features in *c*. Intuitively, this removes the *c* sense from *w*, allowing CBC to discover the less frequent senses of a word and to avoid discovering duplicate senses. The word *w* is then assigned to its next most similar cluster and the process is repeated.

4 Conclusion

We will demonstrate the senses discovered by CBC for 54,685 words on the 3GB ACQUAINT corpus. CBC discovered 24,497 polysemous words.

References

Harris, Z. 1985. Distributional structure. In: Katz, J. J. (ed.) *The Philosophy of Linguistics. New York*: Oxford University Press. pp. 26–47.

Hindle, D. 1990. Noun classification from predicate-argument structures. In *Proceedings of ACL-90*. pp. 268–275. Pittsburgh, PA.

Hutchins, J. and Sommers, H. 1992. *Introduction to Machine Translation*. Academic Press.

Landauer, T. K., and Dumais, S. T. 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

Lin, D. 1994. Principar - an efficient, broad-coverage, principle-based parser. In *Proceedings of COLING-94*. pp. 42–48. Kyoto, Japan.

Lin, D. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING/ACL-98*. pp. 768–774. Montreal, Canada.

Manning, C. D. and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

Pasca, M. and Harabagiu, S. 2001. The informative role of WordNet in Open-Domain Question Answering. In *Proceedings of NAACL-01 Workshop on WordNet and Other Lexical Resources*. pp. 138–143. Pittsburgh, PA.

Salton, G. and McGill, M. J. 1983. *Introduction to Modern Information Retrieval*. McGraw Hill.

Voorhees, E. M. 1998. Using WordNet for text retrieval. In *WordNet: An Electronic Lexical Database*, edited by C. Fellbaum. pp. 285–303. MIT Press.