Meaningful Clusters

Antonio Sanfilippo, Gus Calapristi, Vernon Crow, Beth Hetzler, Alan Turner

Pacific Northwest National Laboratory
902 Battelle Blvd, P.O. Box 999
Richland, WA 99352, USA
{antonio.sanfilippo,gus.calapristi,vern.crow,beth.hetzler,alan.turner}@pnl.gov

Abstract

We present an approach to the disambiguation of cluster labels that capitalizes on the notion of semantic similarity to assign WordNet senses to cluster labels. The approach provides interesting insights on how document clustering can provide the basis for developing a novel approach to word sense disambiguation.

Introduction

One of the long-standing issues in document clustering concerns the identification of clusters' meaning. A widespread practice consists in displaying a selection of prominent terms within each cluster. These are then presented to the user as labels providing an indication of semantic content for each cluster as a whole.

Cluster labels can be helpful in clarifying the meaning of clusters. However, the utility of a cluster label is severely limited when the word it represents is polysemous. For example, WordNet[1] gives 33 senses for the word "drive": 12 as a noun and 21 as a verb. Given enough time, a user may be able to select the correct sense for a cluster label such as "drive" by comparison with the remaining labels in the cluster and direct inspection of the cluster file(s) in which the label occurs. However, users do not usually have the time or disposition to carry out such a meaning-discovery task.

The goal of this paper is to present an approach to the disambiguation of cluster labels that capitalizes on the notion of semantic similarity to assign WordNet senses to cluster labels.

Background

A substantial amount of work has been done using WordNet and EuroWordNet as the basis for word sense disambiguation within the Senseval framework (see www.senseval.org). While the results obtained so far are encouraging, the approaches proposed are not appropriate for the disambiguation of cluster labels. First, the best accuracy these approaches can offer (nearly 0.70 f-measure for supervised methods) does not offer the kind of performance needed for a practical application of cluster label disambiguation. Secondly, the disambiguation algorithms proposed for both supervised and unsupervised methods require several hundred or even thousand usage occurrences of the same word to achieve such results. This is not a requirement that can be enforced in a real-world clustering application. In the data we used to evaluate our approach, for example, the most frequent cluster label word had 38 occurrences and most cluster labels had less than 10 occurrences. We therefore decided to explore a solution which would specifically address the task at hand and dispense with the requirements of existing word sense disambiguation approaches.

[1] http://www.cogsci.princeton.edu/~wn/.

The Approach

The hypothesis we set out to investigate was that a calculation of semantic similarity between a cluster label and each of a set of representative terms within the cluster would provide a reliable indication of the intended word sense for the label in the cluster. If feasible, such an approach would present a variety of advantages over conventional supervised and unsupervised machine learning approaches to word sense disambiguation as

- No training suites and preliminary domain specific language modeling would be required
- The disambiguation task would focus on cluster labels and would therefore require less effort while being specifically tailored to the problems under consideration

Semantic Similarity

Semantic similarity has attracted considerable interest in the last 10-15 years and we direct the reader to Budanitsky (1999) for an extensive survey. In this study, we concentrate on hybrid approaches that use information theoretic measures derived from corpus statistics in combination with the hierarchical structure of a semantic network. An example of such an approach is given by Resnik (1995) who defines semantic similarity between two WordNet synonym sets c_1 c_2 as the informa-

tion content of the *least shared common superordinate* synonym set (*lscs*) of c_1 c_2, as shown in (1) where p(c) is the probability of encountering instances of synonym c in a specific corpus.

(1) sim(c1,c2) = -log p(lscs(c1,c2))

Jiang and Conrath (1997) provide a refinement of Resnik's measure that factors in the relative distance from a synonym set to the least common shared superordinate by calculating the conditional probability of encountering instances of the subordinate synonym set in a corpus given the parent synonym set:

(2) sim(c1,c2) =
2*log p(lscs(c1,c2)) – (log p(c1) + log p(c2))

Lin (1998) introduces a slight modification to Jiang's and Conrath's measure:

(3) sim(c1,c2) =
2*log p(lscs(c1,c2)) / (log p(c1) + log p(c2))

Overall, Jiang's and Conrath's measure seems to outperform other approaches (Budanitsky and Hirst, 2001). Our study corroborates this trend.

Using Semantic Similarity to Disambiguate Cluster Labels

Our disambiguation approach relies on semantic similarity measurements between the label for a given cluster and each of a set of salient terms within the cluster. Both cluster labels and salient terms were obtained through the feature selection algorithm of the IN-SPIRE clustering tool we set out to work with (Wise et al., 1995).

IN-SPIRE cluster labels are selected from a cohort of *major terms* that are used as vector features for cluster modeling. The relevance of a major term in a given cluster is calibrated, or may also be determined *in absentia*, by the presence of *minor terms* that are known to co-occur with the major term in question. Major and minor terms are determined through a feature selection procedure that uses a weighting measure similar to TF*IDF: the first 200 terms with higher weight are selected as major terms, the next 2000 as minor terms.

Each cluster label in IN-SPIRE is therefore associated with a number of minor terms, as indicated in Table 1. Our hypothesis was that finding the word sense under which each cluster label is most similar to each of the minor terms with which it co-occurs, would give us a reliable indication of the prominent word sense for the cluster label.

We chose SemCor[2] as the document collection to test the hypothesis. SemCor contains 352 documents selected from the Brown corpus, where most content words have been manually tagged with a WordNet sense. Our idea was to remove word sense annotations from SemCor, cluster the resulting data with IN-SPIRE, disambiguate cluster labels and then compare the disambiguation results with the original SemCor word sense annotations.

Cluster ID	1
Cluster label	Protein
Minor Terms found with cluster label	body, cell, color, contain, cut, green, liver, normal, result, section, stain, study, wall, white

Table 1: Sample association of cluster labels and minor terms.

Before clustering the SemCor documents, we also removed all other tags and only kept lemmas and punctuation. We then developed a printing facility for IN-SPIRE which for each cluster label would output a *cluster label record* consisting of:

- The cluster ID
- The cluster label
- The filenames of the documents within the cluster
- The list of minor terms found in association with the cluster label

In selecting the data for evaluating the approach, we chose the following thresholds:

- Each cluster would have to contain at least 5 documents
- Each cluster label would have to occur in at least 3 documents within the cluster
- Each minor term would have to co-occur at least 3 times with the cluster label.

These thresholds were found to offer the best balance between amount of data considered for the experiment and goodness of results.

For each cluster label record, we created a *disambiguation hypothesis construct* (see Table 2) consisting of

- The set of semantic similarity record structures for each of its co-occurring minor terms given by the similarity measure chosen, including:
  - The cluster label word and the co-occurring minor term
  - The appropriate part of speech for the cluster label word and the co-occurring minor term
  - The WordNet sense assigned to the cluster label
  - The similarity score (when score > 0)
- The reference filenames.

Disambiguation hypotheses were obtained by deriving semantic similarity scores for each pair of cluster label and co-occurring minor terms using the implementation of Resnik's, Jiang's and Conrath's, and Lin's

[2] http://www.cs.unt.edu/~rada/downloads.html.

measure made available by Patwardhan and Pedersen[3] for WordNet 1.7.1. Since we used the lemmatized version of the SemCor corpus for clustering, all cluster labels and minor terms were already in dictionary form. In addition, we made reference to the original SemCor files to obtain information about part of speech for cluster labels and minor terms.

Cluster ID	1
Disambiguation Hypotheses	tissue#n#1 cell#n#2 0.073 tissue#n#1 cell#n#1 0.072 tissue#n#2 cell#n#2 0.058 tissue#n#2 cell#n#1 0.057 … … … tissue#n#1 liver#n#1 0.114 tissue#n#2 liver#n#2 0.061 tissue#n#1 liver#n#2 0.055 … … …
Filenames	br-j08, br-j12, br-j14, …

Table 2: Example of disambiguation hypothesis construct

For each disambiguation hypothesis, our algorithm selects the cluster label and part of speech with the lowest word sense number that has the highest similarity score. For example, for the two disambiguation hypotheses shown in Table 2, the algorithm would select *tissue#n#1 0.073* and *tissue#n#1 0.114* as the best hypotheses with reference to the minor terms *cell* and *liver*.

Intuitively, the higher the similarity score between a cluster label and its co-occurring minor term, the higher the likelihood that the two words are more indicative of the meaning of the cluster. However, our results show that this is a tendency that is best normalized in terms of word sense frequency. Favoring lower word sense numbers is just a way of carrying out such a normalization, as lower sense numbers in WordNet denote word senses that have higher rate of occurrence.

Whenever no similarity results are available to make an informed choice, the cluster label is assigned sense number 1 by default. About 39% of the final disambiguation hypotheses were obtained by default. Whenever a word is unambiguous, e.g. as the noun "data" which only has one sense in WordNet, no similarity scores are derived and the score is artificially set at 100. Unambiguous words accounted for 5.9% of the cluster labels in our evaluation set.

The score of the selected disambiguation hypotheses, e.g. *tissue#n#1 0.073* and *tissue#n#1 0.114*, are then summed together for all cluster labels bearing the same sense number. The cluster label sense number that has the highest cumulative score is chosen as the disambiguation selection for the cluster label.

[3] http://www.d.umn.edu/~tpederse/similarity.html.

Evaluation

The output of the cluster label disambiguation algorithm described in the previous section is a plurality of *disambiguated cluster label records*. Each such record provides information about cluster ID, the part of speech and sense number of the disambiguated cluster label, and all the files which constitute the cluster --- some of which will contain occurrences of the cluster label word:

Cluster ID	1
Disambiguated cluster label	tissue#n#1
Filenames	br-j08, br-j12, br-j14, …

Table 3: Example of a disambiguated cluster label record.

Disambiguated cluster label records contain all the information that is needed to calculate precision and recall with reference to the original SemCor corpus. To facilitate the evaluation, we created *gold standard records* from the SemCor corpus consisting of words corresponding to cluster labels with the part of speech and sense number for each file name:

cluster_label#filename#POS#sense
tissue#br-e23#n#2
tissue#br-e25#n#1
...

Table 4: Example of gold standard record.

The evaluation corpus consisted of 352 SemCor files grouped into 18 clusters. Each cluster had several cluster labels. In our evaluation we focused on cluster labels which were either nouns or verbs. In all, there were 271 cluster label words which accounted for 181 homographs. The total number of word sense occurrences in the gold standard for the 271 cluster labels was 791.

We ran two distinct tests: *by-cluster* and *by-file*.

The by-cluster test was intended to evaluate how good the word sense disambiguation algorithm was at choosing a correct sense for the cluster label. The requirement for this test was that the sense chosen by the algorithm should occur in at least one of the files within the cluster.

The by-file test was intended to evaluate how good the word sense disambiguation algorithm was at choosing all correct senses of the cluster label for all files in each cluster. The requirement for this test was considerably more stringent as it required that the word sense chosen by the algorithm for the cluster label match all occurrences of the corresponding word and part of speech in the cluster files.

Precision, recall and f-measure were calculated in the usual fashion:

- Precision = true positives/true positives+false positives
- Recall = true positives/true positives+false negatives
- F-measure = 2*precision*recall/(precision+recall)

In the by-cluster test, a true positive obtains when the sense chosen by the algorithm for the cluster label occurs in at least one of the files within the cluster. A true negative obtains when none of the senses in the gold standard files which correspond to the files in a cluster are found. A false positive obtains when a sense chosen by the algorithm does not occur in any of the cluster's files.

In the by-file test, a true positive obtains when the sense chosen by the algorithm for the cluster label with reference to a specific file occurs in that file. A true negative obtains whenever a sense in any of the gold standard files which correspond to the files in a cluster is not found. A false positive obtains when the sense chosen by the algorithm does not occur in any of the cluster's files.

For each test, we ran two scenarios. Each scenario includes results for three similarity measures: Resnik's, Jiang's and Conrath's, and Lin's.

In the first scenario, disambiguated cluster label records (see Table 3) were obtained by selecting the lowest word sense with the highest similarity score, as discussed in 3.2. Results are shown in Tables 5 and 6. Both in the by-cluster and the by-file test, the Jiang & Conrath similarity measure significantly outperforms the other two. These results are in keeping with previous findings (Bundanitsky and Hirst, 2001). The difference in F-measure between the by-cluster and the by-file tests indicates the increased difficulty of the task.

	Resnik	Lin	Jiang & Conrath
Precision	0.664	0.681	1
Recall	0.940	0.940	0.900
F-measure	0.778	0.790	0.947

Table 5: Results for the *by-cluster* test in scenario I.

	Resnik	Lin	Jiang & Conrath
Precision	0.570	0.583	0.733
Recall	0.796	0.796	0.724
F-measure	0.664	0.673	0.729

Table 6: Results for the *by-file* test in scenario I.

In the second scenario, disambiguated cluster label records were obtained by selecting the word sense with the highest similarity score. Results are shown in Tables 7 and 8. We evaluated this scenario as a way of corroborating our intuition that the use of semantic similarity alone for the disambiguation of cluster labels is not enough. Choosing the lowest (most common) word sense number is crucial to steer the disambiguation process in the right direction.

	Resnik	Lin	Jiang & Conrath
Precision	0.482	0.596	0.614
Recall	0.826	0.911	0.940
F-measure	0.609	0.721	0.743

Table 7: Results for the *by-cluster* test in scenario II.

	Resnik	Lin	Jiang & Conrath
Precision	0.443	0.528	0.543
Recall	0.705	0.782	0.812
F-measure	0.544	0.630	0.651

Table 8: Results for the *by-file* test in scenario II.

Conclusions

We have presented an approach to the disambiguation of cluster labels which uses semantic similarity between a cluster label and its co-occurring terms in the cluster to discard bad word sense candidates and relies on word sense frequency for sense selection. This approach achieves excellent results in the identification of a cluster label sense per cluster and fares well in the disambiguation of cluster labels in all their occurrences, opening the way to a novel promising approach which leverages document clustering in word sense disambiguation.

References

Budanitsky, A. 1999. *Lexical Semantic Relatedness and its Application in Natural Language Processing*. Technical report CSRG-390, Department of Computer Science, University of Toronto.

Budanitsky, A. and Hirst, G.. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. *Workshop on WordNet and Other Lexical Resources, NAACL*, Pittsburgh.

Hirst, G. and D. St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In Fellbaum 1998, pp. 305–332.

Jiang J. and D. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of International Conference on Research in Computational Linguistics, Taiwan.

Lin, D. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*,Madison, WI.

Philip Resnik. 1995. Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, Montreal.

Wise, J. A., J. J. Thomas, et al. 1995. Visualizing the non-visual: Spatial analysis and interaction with information from text documents. *IEEE Information Visualization*. IEEE Press, Los Alamitos, CA.