Using Online Knowledge Sources for Semantic Noun Clustering

Emily Jamison
Alias-i
181 North 11th St, #401
Brooklyn, NY 11211
jamison@ling.ohio-state.edu

Abstract

In this paper, we compare different sources of internet knowledge for automatic semantic noun clustering. Two knowledge sources are used: a search-engine-query Hearst-pattern (Hearst, 1992) hypernym generator based on (Kozareva et al., 2008) and (Evans, 2003) and the human-labeled Wikipedia page categories. To fully explore the open-domain flexibility of internet-knowledge-based clustering, six different datasets were clustered, including two samples of the CoNLL 2003 Named Entity dataset, three samples of intra-domain nouns, and a widely cross-domain list. Clustering was performed with the open source package Cluto[1]. The results show that while clustering performance varies across domains, the addition of Wikipedia information universally increases both coverage and F-measure.

1 Previous Research

Semantic clustering (e.g. the recognition that the Dutch Golden Age painter *Gerard Dou* has more in common with the Dutch Golden Age painter *Cornelis Saftleven* than with the Italian Renaissance painter *Antonello da Messina*) is used in information extraction tasks such as coreference resolution to provide similarity values between multiple items. Existing semantic dictionaries and hand-compiled lists may lack the coverage to handle large open domains or rapidly changing categories: Vieira and Poesio (2000) found that of antecedent/anaphoric coreferent pairs in the WSJ, only 56% in hyponymy relations were in WordNet as direct or inherited links.

Several named entity recognition shared tasks, such as CoNLL 2003 and BioCreAtIvE 2004, have focused community resources on the task of automatically identifying and categorizing named entities (NEs) and gene and protein names. However, these tasks use a fixed set of categories and a significant training set; the systems produced cannot be used with other categories for other purposes. The 2008 Concrete Nouns Categorization Task (Baroni et al., 2008) performed clustering on a handmade list of 44 "birds, ground animals, fruit trees, greens, tools, and vehicles." Evans (2003) clustered named entities using hypernyms gathered from the internet with a search engine. However, search engine hypernyms can have limited coverage of data within specific domains. In this paper, we add the use of Wikipedia categories as a knowledge source, and evaluate our algorithms on nouns (named entities and common nouns) from a number of domains.

2 Data Sets

Six different datasets[2] were clustered. Two datasets were subsets of the CoNLL 2003 Named Entity dataset; both were randomly chosen, category-balanced lists of 47 named entities, including persons, organizations, and locations. Since our knowledge sources do not use context, however, the method is particularly sensitive to abbreviations and typos; in one list these were removed. Three datasets were created from www.freebase.com, an online user-compiled database covering a wide variety of topics. These lists included a list of artists (22 persons, 3 categories), a list of professional sports teams (30 teams, 6 categories), and a list of generic medicines (30 medicines, 5 categories). The last dataset was a broadly inter-domain list with subcategories of automakers, languages, artists, sports

[1] Available at http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download

[2] Due to encoding problems with malformed HTML pages online, some nouns had to be excluded from the experiments.

Nouns	Internet hypernyms
Raphael	artists, painters, masters, centuries, others, renaissance, angels, contemporaries, architects, geniuses
Paolo Veronese	artists, painters, masters
Cornelis Saftleven	*None*
Ford	automakers, manufacturers, companies, brands, marques, competitors, oems, trucks, giants, manufactures
Toyota	automakers, rivals, manufacturers, competitors, brands, makers, companies, firms, imports, cars

Table 1: A sample of nouns from the mixed dataset, and their top ten internet hypernyms.

Nouns	Internet hypernyms
ampicillin	antibiotics, amino--penicillins, medicines, spectrum, penicillins, agents, together, lactams, medications, compounds
halothane	*None*
New York Mets	items, events, sports, gifts, team
Southern Redbacks	*None*
Chiefs	leaders, ancestors, authorities, structures, rulers, groups, individuals, leadership, figures, roles
Telugu	languages, scripts, circles, vernaculars, industries, bilinguals, films, requirements, circuits
Breton	languages, surrealists, french, patois, figures, artists, writers, era, france, walkers

Table 2: Continuation of a sample of nouns from the mixed dataset, and their top ten internet hypernyms.

teams, and medicines (40 common nouns and named entities, 5 categories).

3 Algorithms Used

Two different sources of knowledge are used for this study: a search-engine[3]-query Hearst-pattern hypernym generator based on (Kozareva et al., 2008) and the human-labeled Wikipedia page categories. Hearst (1992) showed that hyponymy information could be collected by using a series of hand-crafted frames to search a corpus (here, the internet). Kozareva et al. (2008) used a doubly-anchored Hearst frame to generate a list of class members from web searches. Evans (2003) performed clustering on named entities using their hypernyms from web searches. For the search engine hypernyms, we collected the 10 most frequent categories from 100 web results.

Tables 1 and 2 display a sample of nouns from the mixed dataset along with their internet hypernyms[4]. Some of the nouns have sets of hypernyms that provide a good deal of knowledge about the noun. For example, *Rafael*'s two most frequent hypernyms are *artist* and *painter*, properties that we expect to be most helpful in clustering *Raphael* with other artists in the mixed dataset. Even *Raphael*'s other hypernyms *masters*, *renaissance*, *angels*, *architects*, and *geniuses* may prove useful in clustering with the other artists and not with the medications, sports teams, languages, and car manufacturers. Although short in length, *Paolo Veronese*'s hypernym list of *artists*, *painters*, and *masters* still contains knowledge common with its cluster member *Raphael*.

However, not all nouns have helpful internet hypernym lists. The hypernym list for the sports team the *New York Mets* has nothing in common with the sports team Chiefs. Nouns with hypernym lists that have no overlap with the hypernym lists of other nouns cannot be clustered. Sports team *Southern Redbacks*, among others, cannot even be clustered regardless of the other nouns' hypernyms, because no *Southern Redbacks* internet hypernyms were found. Up to 53%[5] of nouns were unclusterable with only internet hypernyms.

For our second knowledge source, we collected the hand-created categories from the bottom of each noun's Wikipedia page[6]. If the noun had

[3]The Yahoo! Developer API that we used can be downloaded from http://developer.yahoo.com/

[4]Hypernyms are listed in descending frequency of occurrence. When a noun is not listed with 10 hypernyms, this is because its query phrase produced less than 10 unique words.

[5]See Table 4: Cleaned CoNLL subset clustering results

[6]The categories on a Wikipedia can be found as hyperlinks below the External Links section.

no Wikipedia page, we collected the categories from the bottom of the page that was the first result when searching the internet for the noun and the term "Wikipedia".[7] Wikipedia categories were used as clustering features, as had been the internet hypernyms.

The Wikipedia categories frequently added useful information when a noun had few or no internet hypernyms. For example, in Table 1 the sports team *Southern Redbacks* has no internet hypernyms. However, our method of collecting knowledge from Wikipedia finds the correct Wikipedia web page, and the Wikipedia categories for the *Southern Redbacks*: *Sport in South Australia* and *Australian first-class cricket teams*. Depending on which other nouns are also being clustered, this may enable the clustering of the *Southern Redbacks*.

Two other nouns in Tables 1 and 2 with no internet hypernyms also become clusterable by using their Wikipedia categories. *Cornelis Saftleven* has Wikipedia categories including *Dutch painters* and *Baroque painters*, *halothane* has Wikipedia categories including *Anesthetics* and *World Health Organization essential medicines*.

We created three algorithms to test these knowledge sources.

3.1 Evans-based Algorithm

The Evans (2003)-based algorithm used only the web search hypernyms. For clustering, the hypernyms were weighted by their search counts.

3.2 Wikipedia Algorithm

The second algorithm used only Wikipedia categories. The categories were weighted equally.

3.3 Combination Algorithm

The third algorithm used a combination of both Evans (2003)-based internet hypernyms and Wikipedia categories. For the third algorithm, hypernyms were weighted by their search counts and Wikipedia categories were weighted by roughly 50% of the mean weight for internet hypernyms. For Wikipedia category *j* of noun *i*, the weight $W_{i,j}$ is given by the following formula:

$$W_{i,j} = \left(\frac{\sum_k E_{k,i}}{0.5 N_i} \right) + 1$$

[7] If a noun's top "Wikipedia" + noun search result was not a Wikipedia page, then no categories would be collected from the Wikipedia knowledge source. However, this problem did not arise in our datasets.

Dataset	base	E	W	C
mixed nouns	20%	79%	73%	85%
painters	41%	30%	81%	77%
sports teams	17%	37%	83%	50%
medicines	20%	83%	60%	87%
CoNLL cleaned	34%	38%	64%	72%
CoNLL original	33%	42%	60%	60%

Table 3: Clustering results, as f-measures. E = Evans-based; W = Wikipedia.

Algorithm	Cov	Pur	Rec	F-m.
Baseline	100%	20%	20%	20%
Evans-based	70%	90%	70%	79%
Wikipedia	100%	73%	73%	73%
Combination	100%	85%	85%	85%

Table 4: Mixed nouns clustering results.

where $E_{k,i}$ is the weight of Evans internet hypernym *k* and N_i is the total number of internet hypernyms for noun *i*. Nouns were clustered with Cluto (Steinbach et al., 2000), using the categories as predicates, similarly to Evans (2003). The clustering algorithm was a k-ways algorithm ("RBR"), with a predefined number of clusters.

4 Evaluation

The summary of results from our evaluation is displayed in Table 3 as f-measures[8]. The baseline algorithm is a majority class baseline[9]. Coverage is the percent of nouns that were clustered, i.e., for which a result was returned. Non-clustered nouns are included in all results. For most datasets (mixed nouns, sports teams, medicines, and both CoNLL datasets), both the Evans-based algorithm significantly outperformed the majority class baseline. The Evans-based algorithm performed worse than the majority-class baseline on the 3-category painters dataset, which was the only dataset dealing entirely with historical knowledge (painters

[8] This corresponds with 'purity' in the 2008 Concrete Nouns Categorization Task

[9] All nouns are clustered as one cluster.

Algorithm	Cov	Pur	Rec	F-m.
Baseline	100%	34%	34%	34%
Evans-based	47%	59%	28%	38%
Wikipedia	100%	64%	64%	64%
Combination	100%	72%	72%	72%

Table 5: Cleaned CoNLL subset clustering results.

Algorithm	Cov	Pur	Rec	F-m.
Baseline	100%	33%	33%	33%
Evans-based	49%	64%	31%	42%
Wikipedia	100%	60%	60%	60%
Combination	100%	60%	60%	60%

Table 6: CoNLL original subset clustering results.

Algorithm	Cov	Pur	Rec	F-m.
Baseline	100%	41%	41%	41%
Evans-based	50%	45%	23%	30%
Wikipedia	100%	81%	81%	81%
Combination	100%	77%	77%	77%

Table 7: Painters clustering results with 3 clusters (based on the painters' nationalities).

from Italian, Dutch, and French eras of antiquity). We attribute this worse performance to sparse internet coverage of this historical period. The Wikipedia algorithm, however, outperformed the baseline algorithm on all datasets.

For all datasets, either the Wikipedia or the Combination algorithm outperformed the Evans-based and baseline algorithms. The Combination algorithm produced the highest f-measure on the lists with widely-varying categories (i.e. the CoNLL lists and the mixed dataset). The Wikipedia algorithm produced the highest f-measure on intra-topic lists (i.e. painters, sports teams, and medicines).

Tables 4, 5, 6, 7, 8, and 9 show the coverage, purity[10], recall, and f-measure results of the datasets in greater detail. None of the datasets were fully clusterable (i.e., 100% coverage) using the Evans-based algorithm. This is the result of a lack of phrases of the type " * such as *Noun*", either on the internet or accessible by search engine. All of the lists had 100% coverage when Wikipedia was used as a knowledge source. However, for topics lacking Wikipedia coverage, clusterability of less than 100% is to be expected.

[10] similar to precision; see (Baroni et al., 2008)

Algorithm	Cov	Pur	Rec	F-m.
Baseline	100%	17%	17%	17%
Evans-based	63%	47%	30%	37%
Wikipedia	100%	83%	83%	83%
Combination	100%	50%	50%	50%

Table 8: Sports teams clustering results.

Algorithm	Cov	Pur	Rec	F-m.
Baseline	100%	20%	20%	20%
Evans-based	93%	86%	80%	83%
Wikipedia	100%	60%	60%	60%
Combination	100%	87%	87%	87%

Table 9: Medicines clustering results.

5 Conclusions and Future Work

In this paper, we compared two different sources of internet knowledge for automatic semantic noun clustering: a search-engine-query Hearst-pattern (Hearst, 1992) hypernym generator based on (Kozareva et al., 2008) and (Evans, 2003) and the human-labeled Wikipedia page categories. Using a variety of datasets, we found that Wikipedia knowledge used either alone or in combination with internet hypernyms (Evans, 2003) universally increases both f-measure and coverage.

In future work on knowledge sources for noun clustering, we hope to evaluate contextual features, such as argument structure ("John Bunyan wrote *The Pilgrim's Progress*" and "Shakespeare wrote *Macbeth*"; Bunyan and Shakespeare both appear in similar argument position to *wrote*) as semantic knowledge sources.

Acknowledgments

The author wishes to thank Yannick Versley, as well as two anonymous reviewers, for their comments and advice on this project.

References

M. Baroni, S. Evert, A. Lenci (eds). 2008. *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics: Bridging the gap between semantic theory and computational simulations*.

R. Evans. 2003. Framework for Named Entity Recognition in the Open Domain. *Proc. of RANLP-2003*.

M. Hearst. 1992. Automatic Acquisition of hyponyms from large text corpora. *Proc. of the 14th conference on Computational Linguistics*.

Z. Kozareva, E. Reiloff, and E. Hovy. 2008. Semantic Class Learning from the Web with Hyponym Pattern Linkage Graphs. *Proc. of ACL-08: HLT*.

M. Steinbach, G. Karypis, and V. Kumar. 2000. A Comparison of Document Clustering Techniques. *KDD Workshop on Text Mining*.

R. Vieira and M. Poesio. 2000. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):539-593.