A New Framework for Unsupervised Semantic Discovery

Barry Schiffman
Department of Computer Science
Columbia University
New York, NY 10027
bschiff@cs.columbia.edu

Abstract

This paper presents a new framework for the unsupervised discovery of semantic information, using a divide-and-conquer approach to take advantage of contextual regularities and to avoid problems of polysemy and sublanguages. Multiple sets of documents are formed and analyzed to create multiple sets of frames. The overall procedure is wholly unsupervised and domain independent. The end result will be a collection of seets of semantic frames that will be useful in a wide range of applications, including question-answering, information extraction, summarization and text generation.

1 Introduction

As natural language processing research pushes into increasingly sophisticated areas, a corresponding need is growing for richer semantic resources. While considerable progress is being made, this paper specifically takes aim at addressing the difficulty that polysemy and sublanguages present and proposes a divide-and-conquer approach to sidestep the multiple meanings and capitalize on the way sublanguages develop around different topics.

The fundamental idea put forth here is to divide a large corpus into smaller, more coherent clusters and then to extract semantic information based on syntactic patterns from each one. The clusters are formed on the basis of an analysis of lexical co-occurrences within the documents of the corpus. Any tight cluster of correlated words can be used to search for and collect a corresponding set of documents. The extraction of deeper information is accomplished by familiar unsupervised strategies.

I demonstrate the idea here with an examination of a cluster of documents on judicial proceedings. A vector of terms is formed by examining co-occurrence statistics and then using a query to an open source search engine. The co-occurrence study (Schiffman and McKeown, 2004) was done previously on the AQUAINT corpus [1] Then, I searched all the Associated Press documents in the English Gigaword corpus, covering 10 years from 1994 through 2004. After removal of duplicate articles, I ended up with a cluster of 3,119 articles. A separate cluster of 9,531 randomly selected documents was collected for evaluation.

Lexical-syntactic regularities were then examined, using binomial log likelihood ratios to establish associations between recurring lexical entities and syntactic patterns, uncovering the roles played by the important types of entities for the topic, in this case, mostly people, judges, lawyers, juries and criminals. I use a dependency parser, and a named-entity recognizer in the analysis, and the end result tells us such things as:

- witnesses appear, and witness stands are taken
- juries acquit, and defendants are acquitted
- defendant plead, and pleas can be changed

This kind of information is the building block for further analysis. I determine who does what to

[1] Three years of English newswire used in the Advanced Question Answering for Intelligence program.

whom by the statistical strength of these repetitions. I contrast the strongest patterns in a topical cluster with those in general clusters to find what expressions truly belong to a topic.

It is important to emphasize that the selection of judicial subject is arbitrary. A large number of tight lexical clusters are evident from the table of co-occurrences. With sufficient resources, many more can be exploited

These are the terms used to forge the queries for our investigation, and they should likewise be capable of choosing the appropriate semantic resources for a given task on a given subject, for example to categorize questions in question answering or document sets in summarization tasks.

I take note that people have a difficult time understanding fragments of a discourse unless they can determine what the discourse is about. This is our fundamental intuition. I view language as a patchwork and that the patches must be learned one at a time.

I also note that our analysis requires well-behaved text to learn from. As researchers push into more unruly territory, like email and blogs, they will have a greater need for just the kind of knowledge I am developing.

In the next section, I will discuss previous work. I am building on a number of efforts. In section 3, I cover the construction of the document clusters. In section 4, I show how I extracted the information. Then in section 5, I examine the results of our method on the judicial cluster. And finally, in section 6, I discuss the next steps.

2 Related Work

A large body of work exists on the extraction of one kind of semantic relation or another from syntactic relations, using either patterns implemented as regular expressions or by parsers. In one of the key early papers on the subject, Pereira (1993) clustered verb-object pairs to determine similarity, using information theoretic measures of similarity. Hatzivassiloglou and McKeown (1993) group adjectives by their meaning, using syntactic patterns. Later, Lin (1998a) used a large number of syntactic relations that he found automatically with his rule-based parser Minipar (1998b), and determined similarity by computing the mutual information of pairs of words in various syntactic relationships, like subject-verb, verb-object, with each other. The statistics were drawn from a 64-million-word corpus of news, from the Wall Street Journal, the San Jose Mercury and the AP.

Riloff and Shepherd (1997) seed words to build up larger sets of semantically similar words in one or more categories. Riloff also experimented with determining which syntactic patterns fit information extraction tasks by comparing relevant and non-relevant corpora. Later Riloff and Jones (1999) adapted bootstrapping techniques to lexicon building targeted to information extraction.

Information extraction templates are also automatically created off-line by Filatova (2006) by finding the most frequent verbs in sets of documents relevant to some class of events.

In the same vein, researchers at Brown University (Caraballo and Charniak, 1999), (Berland and Charniak, 1999), (Caraballo, 1999) and (Roark and Charniak, 1998) focused on target constructions, in particular complex noun phrases, and searched for information not only on identifying classes of nouns, but also hypernyms, noun specificity and meronymy.

In a more recent effort, Snow, Jurafsky and Ng (2004) used Minipar to examine longer dependency paths and search for hypernym relations, incorporating WordNet (Miller et al., 1990) in a learning paradigm.

In these and many others, a particular semantic relation or a small set of them were defined and then searched in the corpus. Some methods input a particular pattern or at least some example patterns and built upon that.

And Pantel and Pennacchiotti (Pantel and Pennacchiotti, 2006; Pennacchiotti and Pantel, 2006) have pursued a generalized approach that will produce patterns and extract information on a large scale basis. They report success on very specific relations, such as birthdays.

3 Role of Context

Since context is often necessary for human understanding of language, it follows that an automated process would benefit as well, although I make no claim that an automated process will achieve human-

like understanding anytime soon. I show below that I can automatically select and build clusters of similar documents and then break the overall problem of extracting semantic information from free text into a large number of much smaller problems.

Our method begins with the automatic clustering of documents into coherent groups. Our goal was to do this without needing to manually define the clusters in advance. I made large study of document co-occurrences that had been compiled for other purposes, compiling a huge table of words that occurred at least 100 times in a corpus. As I said in section 1, I used the Associated Press portion of the Aquaint corpus, a 52 million-word collection of English newswire from 1998 through 2000. The statistic used to measure the strength of the co-occurrences, or association, between two words was the log likelihood ratio (LLR) (Dunning, 1993), the same that I used to extract related lexical-syntactic patterns.

$$\lambda = \frac{max_p(L(p, k_1, n_1)L(p, k_2, n_2)}{max_{p_1,p_2}L(p_1, k_1, n_1)L(p_2, k_2, n_2)}),$$

where

$$L(p, k, n) = p^k(1 - p)^{n-k}.$$

Table 1 shows the top 20 words associated with *pitcher*, *diplomat*, *recipe* and *trial*. The association relation here is not necessarily symmetrical and the fact that a trial is closely associated with a prosecutor does not absolutely imply that a prosecutor is closely associated with a trial, although in this case, the relation hold in both directions. The trial-to-prosecutor association measures 27,073 and the prosector-to-trial association is 26,687.

A number of interesting observations are clear from the table. While the four groups are generally coherent, the average scores vary widely, almost two magnitudes from recipe to trial. The choice of a threshold has been studied elsewhere (Inkpen and Hirst, 2001; Moore, 2004) The proper value is a guessing game that at best works sometimes and not others. As Dunning pointed out, the LLR has the desirable quality that $-2log\lambda$ is asymptotically χ^2 distributed. The likelihood ratio tests do not depend on the assumption of normality as do many other statistical tests, but the χ^2 critical values can be used with the degrees of freedom set at the difference in the number of parameters, here $df = 1$, which has a critical value of about 6.6, indicating that a candidate association is statistically significant with a p-value < 0.01. However, in practice, such an interpretation leads to low precision (Kiss and Strunk, 2002). A number of researchers have sought to scale the LLR to avoid some of its pitfalls by compensating for the tendency of the LLR to overvalue infrequent associations, and Moore (2004) explored a way to apply the metric to rare events. I view the scores as relative measures of association rather than absolute evidence of statistical significance, leaving aside the question of what to do with the large number of low-frequency pairs, many of which are meaningful.

Table 1 also suggestions a number of other qualities about the metric. I took the scores on part-of-speech tagged words rather than on the plain tokens or stemmed tokens because of the imbalance in the numbers of the tags. Verbs are generally outnumbered by nouns, and the bias of the LLR to rare events shows up by favoring verbs, in particular those that are frequent, like say, be, have, and make. In light of that fact, I discarded frequently used verbs when I formed the query vectors that are used to collect the document clusters.

From the tables it is also clear that single words often do not carry complete or specific topical associations. Consider the words associated with *pitcher*, from the game of baseball. Standing alone, neither season, nor start, nor run, nor manager indicate much about the topic at hand at all. Game, team, league and player begin to indicate a sports context, but any one of them alone would not be convincing, but taken together, with the first group, gives a reader a clear idea of what is being talked about. Inning and homer are, of course, baseball specific terms.

Using these lists to form a query, a larger collection of documents is searched to form a topical cluster. I used Lucene [2], a modifiable open source search engine. I altered the formula used to measure the similarity of documents to the query in the default version to eliminate $tf*idf$ weighting of search terms, since my queries specifically seek documents with most of the query term and they avoid any non-significant terms. In addition, I wanted the metric to

[2] http://lucene.apache.org/java/docs/index.html

emphasize inclusion of most terms.

I used the Associated Press portion of the English Gigaword corpus, which contains 1.6 million news documents spanning 1994 to 2004, covering most important global events during that period. The long span is necessary to avoid getting caught in the terminology of a few specific events. This tendency is evidence in the list of associations for the word *diplomat* in table 1. Iran and Albania are not intrisically linked to diplomacy, but were diplomatic issues at the turn of the 21st century.

In the same vein, duplication of articles can subvert the extraction process by counting multiple instances of whole documents often with only minor changes so that a sentence written once, on one day by one writer, can appear a dozen times in the corpus. To remove duplication, I used a procedure based on the near-duplicate detection method put forth by Yang and Callan (2006) that used all trigrams in the document set as keys in a hash. In the *trial* experiment, I imposed a draconian threshold of 0.4 similarity and reduced the initial cluster of 6,310 documents, to only 3,119. There were a number of important judicial proceding in the time frame, like the Yugoslav war crimes tribunal, and the articles were often repeated throughout the day by the news service as small details prompted a slightly different new version of the same coverage.

In this paper, I used the LLR in two ways. As described above, I am using it as a approximate guide to the topicality suggested by a word. The *trial* topic was selected because of the relatively high LLR scores for trial words and strong degree of symmetricality among them. In the future, I will develop a more sophisticated procedure for selecting topics and forming the query vectors. In section 4 I will make comparisons between scores derived from the target set of documents and those from a general set of documents.

4 Extraction

To extract candidate semantic structures, I used Dekang Lin's Minipar (1998b), a full coverage dependency parser that produces a list of words in some syntactic relation to one another, a head noun and its modifier, a preposition and its complement, and so forth. I extracted a number of extended paths to study, including a noun and its modifiers, a noun and a prepositional phrase postmodifier, and subject-verb-object triples, and focused on the later.

Subject	Subj-Verb	Count *	LLR
ENAMEX	17	53086	129.0887
court	46	3897	88.3637
jury	30	1629	80.5203
PRO	17	30413	47.8335
tribunal	15	1043	33.4651
TIMEX	2	9635	26.2386
juror	10	1100	14.6572
panel	6	519	11.1084
testimony	4	1232	0.7783
that	4	2861	0.6198
U.N. court	2	127	4.7553
evidence	8	1763	4.1071
prosecutor	2	3120	4.0428
judge	11	3174	2.7312
member	2	1326	0.1895
Judge	2	1140	0.0419

Table 2: This table show all the subjects of the instances of the verb convict where the frequency of the subject/verb pairs was greater than 1. The Subj-Verb column is the count of the co-occurrence, the Count * column show how often the particular noun appeared and the last column shows the LLR score for the pair.

Lin's parser handles long-distance attachments, so that subject arguments are often repeated in a sentence where there are relative clauses and non-finite clauses. However, there is no coreference of names and pronouns. In parallel, I ran the documents through BBN's Identifinder. and reduced all named entities to their major types, ENAMEX, TIMEX, NUMEX, and all pronouns to one type, which I label PRO. These often have high association scores, but will require further work.

Table 2 shows a breakdown of the subject-verb pairs with the verb convict, which appeared 814 times in the cluster. By its simple frequency, the verb convict is obviously important in a trial event, but the table gives a clear idea of the typical frame for the word. It tells us that courts, the juries, and tribunals are the agents of conviction in the criminal justice sense of the word. It tells us it is less likely

for prosecutors and judges to do the convicting – as in the case of the court system in the United States, where the AP is based.

Table 3 shows the corresponding chart of the objects of convict. here the distribution is much more diffuse, and more likely to employ a named entity. The obvious names that appear as plain nouns are failures of the parser and Identifinder to line up correctly. In addition some of the parsing errors ar evident, but it is clear if types were assigned to the nouns that these are mostly individuals who are convicted, producing a semantic verb frame for the verb convict. It is not hard to trace actual names of the people convicted from the parses and to get a clearer idea of who is convicted:

- The same jury earlier convicted Christian Longo.

  Christian Longo, 29, has pleaded guilty to killing his wife and ...

- The court Saturday convicted lawmaker Cesare Previti ...

  Cesare Previti, a lawyer who served as defense minister in Berlusconi's first government in 1994, had been accused of influencing judges ...

- Judge William Bristol listened to convicted murderer Jose Julian Santiago ...

  Santiago, 25, who is accused of shooting dead Zyrone Scrivens and his toddler nephew, Drequan ...

5 Evaluation

For a quick evaluation of the ability of the system to distinguish between legitimate topical associations both from general associations that would cut across large numbers of topics and from chance occurrences. The frames extracted from the general set should be consistent with general usage and not reflect any specialized usage.

To do this, I used the cluster of randomly selected documents and extracted the same syntactic relations and then examined the strength of association between the components, just as I did for the trial cluster. I chose a sample of 200 of the frames and divided them into four groups:

Object	Verb-Object	Count *	LLR
ENAMEX	129	53086	4.8324
PRO	67	30413	0.5317
man	24	2617	35.7638
defendant	18	2369	21.5544
officer	13	1322	20.7040
NUMEX	12	7493	0.7137
people	12	1824	11.8309
brother	10	585	25.3056
criminal	4	469	5.4488
woman	3	1327	0.0368
policeman	3	84	11.6804
mayor	3	173	7.6482
lawyer	4	3355	1.3482
journalist	2	280	2.1931
suspect	6	1093	4.4433
death	3	1156	0.1744
others	4	653	3.5329
soldier	6	509	11.3030
first person	7	44	48.5154
those	5	438	9.1478
official	4	1239	0.7612
accused	4	438	5.8734
firm	3	215	6.5032
agent	6	531	10.8807
commander	6	351	15.1602
someone	3	236	6.0249
doctor	3	342	4.2145
terrorist	3	363	3.9401
TIMEX	5	9635	15.6196
who	4	246	9.7440
person	7	440	16.8019
client	3	756	1.1111
businessman	3	149	8.4560
anyone	3	159	8.1027

Table 3: This table show all the objects of the verb convict where the frequency of the verb/object pairs was greater than 2. The Verb-Obj column is the count of the co-occurrence, the Count * column show how often the particular noun appeared and the last column shows the LLR score for the pair.

>, from the judicial cluster	4.418
<, from the random cluster	1.908
?, low scores from the judicial cluster	1.434
M, frames not found in the random cluster	3.920

Table 4: Results of a test of the systems ability to find valid frames associated with judicial proceedings.

- Frames where the LLR scores of the trial set exceeded the scores for the random set, designated as ">"
- Frames where the LLR scores of the random set exceeded the scores for the trial set, designated as "<"
- Frames where the LLR scores of the trials set were below the chi square critical value of 6.6, designated as "?"
- Frames where there were frames in the trial set, but no frames in the random set, designated as "M"

I asked 14 colleagues who are not related to this work to rate each of the 200 frames on how likely the frame was to have an association with a kind of judicial process, on a scale of 1 to 5, with 5 being the most likely. The sample was taken from the 4,000 frames that had a frequency of at least five.

Table 4 shows that there was a clear preference for the frames that chosen by the system operating on the trial set.

6 Future Work

The work here shows that it is possible to identify documents of arbitrary classes and to extract related semantic frames that apply to that class. By the same means, it should be possible to classify specific tasks. For example, the method would determine the appropriate class for a particular question or group of documents to be summarized, and then be able to apply the appropriate set of semantic frames to help in completing the task.

I've completed a sample evaluation and the found that humans reviewing the intermediate patterns confirmed that they are indeed associated with subject matter at hand. However, a task based evaluation will be more informative and one in question-answering and one in multidocument summarization are being planning for the near future.

Future work includes expanding the tests from the judicial cluster to several others, and to formalize the procedure for choosing the queries that are used to build the individual corpora. An interesting byproduct of the system would be to compile a large group of capsule descriptions of people and organizations found in the document clusters.

References

Matthew Berland and Eugene Charniak. 1999. Finding parts in very large corpora. Technical Report TR CS99-02, Brown University.

Sharon Caraballo and Eugene Charniak. 1999. Determining the specificity of nouns from text. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*.

Sharon Caraballo. 1999. Automatic acquisition of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, June.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Elena Filatova, Vasileios Hatzivassiloglou, and Kathleen McKeown. 2006. Automatic creation of domain templates. In *Proceedings of ACL-COLING, 2006*.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1993. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of the 31st Annual Meeting of the ACL*.

Diana Zaiu Inkpen and Graeme Hirst. 2001. Building a lexical knowledge-base of near-synonym differences. In *Proceedings of the NAACL workshop on WordNet and Other lexical Resources*.

Tibor Kiss and Jan Strunk. 2002. Scaled log likelihood ratios for the detection of abbreviations in text corpora. In *Proceedings of the 19th International Conference on Computational Linguistics*.

Dekang Lin. 1998a. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*.

Dekang Lin. 1998b. Dependency-based evaluation of minipar. In *Workshop on the Evaluation of Parsing Systems*.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography (special issue)*, 3(4):235–312.

Robert C. Moore. 2004. On log-likelihood-ratios and the significance of rare events. In *Proceedings of the 2004 Conference on Empirical Methods in NLP*.

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the Conference on Computational Linguistics/Association for Computational Linguistics*.

Marco Pennacchiotti and Patrick Pantel. 2006. A bootstrapping algorithm for automatically harvesting semantic relations. In *Proceedings in Inference in Computational Semantics*.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of english words. In *Proceedings of the 31st Annual Meeting of the ACL*.

Ellen Riloff and Rosie Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*. AAAI.

Ellen Riloff and Jessica Shepherd. 1997. A corpus-based approach for building semantic lexicons. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*.

Brian Roark and Eugene Charniak. 1998. Noun-phrasae co-occurrence statistics for semi-automatic semantic lexicon construction. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*.

Barry Schiffman and Kathleen R. McKeown. 2004. Machine learning and text segmentation in novelty detection. Technical Report CUCS-036-04, Columbia University.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. In *Proceedings of NIPS*.

Hui Yang and Jamie Callan. 2006. Near-duplicate detection by instance-level constrained clustering. In *Proceedings of the 29th annual international ACM SIGIR*.

Pitcher	Diplomat	Recipe	Trial
be vb 38922	say vb 10660	food nn 1122	say vb 92883
game nn 32133	be vb 5010	be vb 424	be vb 84357
say vb 19293	Iranian jj 3979	cheesecake nn 292	case nn 28173
season nn 16739	council nn 2652	say vb 280	prosecutor nn 27073
have vb 13564	Albanian nn 2423	egg nn 243	witness nn 25323
inning nn 13312	government nn 2295	cake nn 234	lawyer nn 24865
first jj 12700	country nn 2244	oven nn 179	court nn 24508
late jj 11237	ambassador nn 2123	make vb 178	judge nn 18316
get vb 8929	official nn 2094	beer nn 162	charge nn 17271
team nn 8761	military jj 2079	dish nn 161	jury nn 16165
start nn 8346	diplomatic jj 2030	cup nn 158	have vb 15492
manager nn 7569	Iranians nn 1928	have vb 157	year nn 14797
run nn 7185	force nn 1838	eat vb 151	testimony nn 14792
league nn 6724	embassy nn 1596	microwave nn 145	impeachment nn 14204
player nn 6664	afghan jj 1559	book nn 140	testify vb 13102
time nn 6476	ethnic jj 1448	butter nn 130	convict vb 12256
make vb 6323	border nn 1371	bacon nn 129	president nn 12064
hit vb 6314	weapon nn 1360	sausage nn 125	defense nn 11678
year nn 6303	inspector nn 1352	kitchen nn 125	time nn 11209
homer nn 6258	leader nn 1349	chocolate nn 118	tell vb 11173

Table 1: Sample of Log likelihood ratios. Each column show the top 20 words associated with the four nouns in the column header. In each row, the associated word, its part of speech tag and the log likelihood ratio score between them.