Experiments in Automated Lexicon Building for Text Searching

Barry Schiffman and **Kathleen R. McKeown**
Department of Computer Science
Columbia University
New York, NY 10027, USA
{bschiff,kathy}@cs.columbia.edu

Abstract

This paper describes experiments in the automatic construction of lexicons that would be useful in searching large document collections for text fragments that address a specific information need, such as an answer to a question.

1 Introduction

In developing a system to find answers in text to user questions, we uncovered a major obstacle: Document sentences that contained answers did not often use the same expressions as the question. While answers in documents and questions use terms that are related to each other, a system that searches for answers based on the question wording will often fail. To address this problem, we developed techniques to automatically build a lexicon of associated terms that can be used to help find appropriate text segments.

The mismatch between question and document wording was brought home to us in an analysis of a testbed of question/answer pairs. We had a collection of newswire articles about the Clinton impeachment to use as a small-scale corpus for development of a system. We asked several people to pose questions about this well-known topic, but we did not make the corpus available to our contributors. We wanted to avoid questions that tracked the terminology in the corpus too closely to simulate questions to a real-world system. The result was a set of questions that used language that rarely matched the phrasing in the corpus. We had expected that we would be able to make most of these lexical connections with the help of Wordnet (Miller, 1990).

For example, consider a simple question about testimony: "Did Secret Service agents give testimony about Bill Clinton?" There is no reason to expect that the answer would appear baldly stated as "Secret Service agents did testify ..." What we need to know is what testimony is about, where it occurs, who gives it. The answer would be likely to be found in a passage mentioning juries, or prosecutors, like these found in our Clinton corpus:

> Starr immediately brought Secret Service employees before the grand jury for questioning.
>
> Prosecutors repeatedly asked Secret Service personnel to repeat gossip they may have heard.

Yet, the Wordnet synsets for "testimony" offer: "evidence, assertion, averment and asseveration," not a very helpful selection here. Wordnet hypernyms become general quickly: "declaration," "indication" and "information" are only one step up in the hierarchy. Following these does not lead us into a courtroom.

We asked our contributors for a second round of questions, but this time made the corpus available to them, explaining that we wanted to be sure the answers were contained in the collection of articles. The result was a set of questions that much more closely matched the wording in the corpus. This was, in fact, what the 1999 DARPA question-answering competition did in order to ensure that their questions could be answered (Singhal, 1999). The second question-answering conference adopted a new approach to gathering questions and verifying separately that they are answerable.

Our intuition is that if we can find the typical lexical neighborhoods of concepts, we can efficiently locate a concept described in a query or a question without needing to know the precise way the answer is phrased and without relying on a costly, hand-built concept hierarchy.

The example above illustrates the point. Testimony is given by witnesses, defendants, eyewitnesses. It is solicited by prosecutors, counsels, lawyers. It is heard by judges, juries at trials, hearings, and recorded in depositions and transcripts. What we wanted was a complete description of the world of testimony – the who, what, when and where of the word. Or, in other words, the "meta-aboutness" of terms.

To this end, we experimented using shallow linguistic techniques to gather and analyze word co-occurrence data in various configurations. Unlike previous collocation research, we were interested in an expansive set of relationships between words

rather than a specific relationship. More important, we felt that the information we needed could be derived from an analysis that crossed clause and sentence boundaries. We hypothesized that news articles would be coherent so that the sequences of sentences and clauses would be linked conceptually.

We examined the nouns in a number of configurations – paragraphs, sentences, clauses and sequences of clauses – and obtained the strongest results from configurations that count co-occurrences across the surface subjects of sequences of two to six clauses. Experiments with multi-clause configurations were generally more accurate in a variety of experiments.

In the next section, we briefly review related research. In section 3 we describe our experiments. In section 4, we discuss the problem of evaluation, and look ahead to future directions in the concluding sections.

2 Related Work

There has been a large body of work in the collection of co-occurrence data from a broad spectrum of perspectives, from information retrieval to the development of statistical methods for investigating word similarity and classification. Our efforts fall somewhere in the middle.

Compared with document retrieval tasks, we are more closely focused on the words themselves and on specific concepts than on document "aboutness." Jing and Croft (1994) examined words and phrases in paragraph units, and found that the association data improves retrieval performance. Callan (1994) compared paragraph units and fixed windows of text in examining passage-level retrieval.

In the question-answering context, Morton (1999) collected document co-occurrence statistics to uncover part-whole and synonymy relationships to use in a question-answering system. The key difference here was that co-occurrence was considered on a whole-document basis. Harabagiu and Maiorano (1999) argued that indexing in question answering should be based on paragraphs.

One recent approach to automatic lexicon building has used seed words to build up larger sets of semantically similar words in one or more categories (Riloff and Shepherd, 1997). In addition, Strzalkowski and Wang (1996) used a bootstrapping technique to identify types of references, and Riloff and Jones (1999) adapted bootstrapping techniques to lexicon building targeted to information extraction.

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.

We have a different perspective than these lines of inquiry. They were specifying various semantic relationships and seeking ways to collect similar pairs. We have a less restrictive focus and are relying on surface syntactic information about clauses.

For more than a decade, a variety of statistical techniques have been developed and refined. The focus of much of this work was to develop the methods themselves. Church and Hanks (1989) explored the use of mutual information statistics in ranking co-occurrences within five-word windows. Smadja (1992) gathered co-occurrences within five-word windows to find collocations, particularly in specific domains. Hindle (1990) classified nouns on the basis of co-occurring patterns of subject-verb and verb-object pairs. Hatzivassiloglou and McKeown (1993) clustered adjectives into semantic classes, and Pereira et al. (1993) clustered nouns on their appearance in verb-object pairs. We are trying to be less restrictive in learning multiple salient relationships between words rather than seeking a particular relationship.

In a way, our idea is the mirror image of Barzilay and Elhadad (1997), who used Wordnet to identify lexical chains that would coincide with cohesive text segments. We assumed that documents are cohesive and that co-occurrence patterns can uncover word relationships.

3 Experiments

The focus of our experiment was on units of text in which the constituents must fit together in order for the discourse to be coherent. We made the assumption that the documents in our corpus were coherent and reasoned that if we had enough text, covering a broad range of topics, we could pick out domain-independent associations. For example, testimony can be about virtually anything, since anything can wind up in a court dispute. But over a large enough collection of text, the terms that directly relate to the "who," "what" and "where" of testimony per se should appear in segments with testimony more frequently than chance.

These associations do not necessarily appear in a dictionary or thesaurus. When humans explain an unfamiliar word, they often use scenarios and analogies.

We divided the experiments in two groups: one group that looks at co-occurrences within a single unit, and another that looks at a sequence of units.

In the first group of experiments, we considered paragraphs, sentences and clauses, each with and without prepositional phrases.

- Single paragraphs with/without PP
- Single sentences with/without PP
- Single clauses with/without PP

In the second group, we considered two clauses and sequences of subject noun phrases from two to six clauses. In this group, we had:

- Two clauses with/without pp
- A sequence of subject NPs from 2 clauses
- A sequence of subject NPs from 3 clauses
- A sequence of subject NPs from 4 clauses
- A sequence of subject NPs from 5 clauses
- A sequence of subject NPs from 6 clauses

The intuition for the second group is that a topic flows from one grammatical unit to another so that the salient nouns, particularly the surface subjects, in successive clauses should reveal the associations we are seeking.

To illustrate the method, consider the three-clause configuration: Say that $word_i$ appears in $clause_n$. We maintain a table of all word pairs and increment the entries for ($word_i$, $word_j$), where $word_j$ is a subject noun in $clause_n$, $clause_{n+1}$, or $clause_{n+2}$. No effort was made to resolve pronomial references, and these were skipped.

We used nouns only because preliminary tests showed that pairings between nouns seemed to stand out. We included tokens that were tagged as proper names when they also have have common meanings. For example, consider the Linguistic Data Consortium at the University of Pennsylvania. Data, Consortium and University would be on the list used to build the table of matchups with other nouns, but Pennsylvania would not. We also collected noun modifiers as well as head nouns as they can carry more information than the surface heads, such as "business group", "science class" or "crime scene."

The corpus consisted of all the general-interest articles from the New York Times newswire in 1996 in the North American News Corpus, and did not include either sports or business news. We first removed duplicate articles. The data from 1996 was too sparse for the sequence-of-subjects configurations. To balance the experiments better, we added another year's worth of newswire articles, from 1995, for the sequence-of-subject configurations so that we had more than one million matchups for each configuration (Table 1).

The process is fully automatic, requiring no supervision or training examples. The corpus was tagged with a decision-tree tagger (Schmid, 1994) and parsed with a finite-state parser (Abney, 1996) using a specially written context-free-grammar that focused on locating clause boundaries. The grammar also identified extended noun phrases in the subject position, verb phrases and other noun phrases and prepositional phrases. The nouns in the tagged, parsed corpus were reduced to their syntactic roots (removing plurals from nouns) with a lookup table created from Wordnet (Miller, 1990) and CELEX (1995). We performed this last step mainly to address the sparse data problem. There were a substantial number of pairings that occurred only once. We eliminated from consideration all such singletons, although it did not appear to have much effect on the overall outcome.

Config	*Matchups*
Para +pp	6.5 million
Sent	1.7 million
Sent +pp	4 million
1 Clause	1.1 million
1 Clause +pp	2.8 million
2 Clause	1.9 million
2 Clause +pp	5 million
Subj 2 Clause	1.1 million*
Subj 3 Clause	1.6 million*
Subj 4 Clause	2.1 million*
Subj 5 Clause	2.6 million*
Subj 6 Clause	3.1 million*

Table 1: Number of matchups found; the "*" denotes the inclusion of 1995 data

There were about 1.2 million paragraphs, 2.2 million sentences and 3.4 million clauses in the selected portions of the 1996 corpus. The total number of words was 57 million. Table 2 shows the number of distinct nouns.

	All Extracted	*Counts* > 1
No pps	74,500	44,400
W/pps	91,700	53,900
Subjs	51,000	30,800

Table 2: Distinct Nouns, 1996 Data

To score the matchups in our initial experiments, we used the Dice Coefficient, which produces values from 0 to 1, to measure the association between pairs of words and then produced an ordered association list from the co-occurrence table, ranked according to the scores of the entries.

$$score_d = \frac{2 * freq(word_i \cap word_j)}{freq(word_i) + freq(word_j)}$$

One problem was immediately apparent: The quality of the association lists varied greatly. The scoring was doing an acceptable job in ranking the words within each list, but the scores varied greatly from one list to another. Our initial strategy was to choose a cutoff, which we set at 21 for each list, and we tried several alternatives to weed out weak associations.

In one method, we filtered the association lists by cross-referencing, removing from the association list for $word_i$ any $word_j$ that failed to reciprocate and to give a high rank to $word_i$ on its association list. Another similar approach was to try to combine evidence from different experiments by taking the results from two configurations into consideration. A third strategy was to calculate the mutual information between the target word and the other words on its association list.

$$score_{mi} = p(xy) * \log\left(\frac{p(xy)}{p(x)p(y)}\right)$$

Using the mutual information computation provided an way of using a single measure that was able to compare matchups across lists. We set a threshold of $1x10^{-6}$ for all matchups. Thus these association lists vary in length, depending on the distributions for the words, allowing them to grow up to 40, while some ended up with only one or two words.

4 Evaluation

The evaluation of a system like ours is problematic. The judgments we made to determine correctness were not only highly subjective but time-consuming. We had 12 large lexicons from the different configurations. We had chosen a random sample of 10 percent of the 2,700 words that occurred at least 100 times in the corpus, and manually constructed an answer key, which ended up with almost 30,000 entries.

From the resulting 270 words, we discarded 15 of those that coincided with common names of people, such as "carter," which could refer to the former American president, Chris Carter (creator of the television show "X-Files"), among others. We thought it better to delay making decisions on how to handle such cases, especially since it would require distinguishing one Carter from another. Such words presented several difficulties. Unless the individuals involved were well-known, it was often impossible to distinguish whether the system was making errors or whether the resulting descriptive terms were informative.

Tables 3 and 4 show an example from the answer key for the word "faculty."

The overall results from the first stage of the process, before the cross-referencing filter are shown in Table 5, ranging from 73% to 80% correct. The configurations that included prepositional phrases and those that used sequences of subject noun phrases outperformed the configurations that relied on subjects and objects in a single grammatical unit. These differences were statistically significant, with $p < 0.01$ in all cases.

The overall results after cross-referencing, in Table 6, showed improvements of 5 to 10 percentage points, while the effect of the number of matchups was diminished. Here, the subject-sequence configurations showed a distinct advantage. While more noise might be expected when a large segment of text is considered, these results support the notion that the underlying coherence of a discourse can be recovered with the proper selection of linguistic features. The improvements in each configuration over the corresponding configuration in the first stage were all statistically significant, with $p < 0.01$. Likewise, the edge the sequence-of-subjects configurations had over the other configurations, was also statistically significant.

enrollment	hiring	administrator
journalism	alumnus	student
school	union	math
engineering	curriculum	trustee
group	seminar	thesis
tenure	staff	department
mathematician	educator	member
ivy	arts	college
chancellor	report	senate
activism	university	chairman
professor	teaching	law
regent	doctorate	administration
academic	committee	semester
board	campus	undergraduate
salary	council	research
president	adviser	mathematics
course	advisor	sociology
dean	study	science
teacher	cannon	provost
vote		

Table 3: Answer Key for Faculty: OK

load	trafficway	unrest
architecture	diversity	hurdle
shield	minority	revision
disburse	percent	woman
clement		

Table 4: Answer Key for Faculty: Wrong

The results from combining the evidence from different configurations, in Table 7, showed a much higher accuracy, but a sharp drop in the total number of associated words found. The most fruitful pairs of experiments were those that combined distinct approaches, for example, the five-subject configuration with either full paragraphs or with sentences with prepositional phrases. It will remain unclear until we conduct a task-based evaluation whether the smaller number of associations will be harmful.

The final experiment, computing the mutual information statistic for the matchups of a key word with co-occurring words was perhaps the most interesting because it gave us the ability to apply a

Config	OK	Wrong	Pct OK
Para +pp	3832	1054	78
Sent	3773	1270	75
Sent +pp	3973	1070	79
1 Clause	3652	1371	73
1 Clauses +pp	3935	1108	78
2 Clauses	3695	1328	74
2 Clauses +pp	3983	1018	80
Subj 2 Cl	3877	1139	77
Subj 3 Cl	3899	1117	78
Subj 4 Cl	3905	1082	78
Subj 5 Cl	3904	1076	78
Subj 6 Cl	3909	1066	79

Table 5: Results Before Cross Referencing

Config	OK	Wrong	Pct OK
Para +pp	3650	734	83
Sent	3328	742	82
Sent +pp	3751	818	82
1 Clause	3067	748	80
1 Clauses +pp	3659	826	82
2 Clauses	3048	554	85
2 Clauses +pp	3232	604	84
Subj 2 Cl	2910	450	87
Subj 3 Cl	3020	440	87
Subj 4 Cl	3050	428	88
Subj 5 Cl	3133	442	88
Subj 6 Cl	3237	449	88

Table 6: Results After Cross Referencing

single threshold across different key words, saving the effort of performing the cross-referencing calculations and providing a deeper assortment in some cases. In most of the configurations, mutual information gave us more words, and greater precision at the same time, but most of all, gave us a reasonable threshold to apply throughout the experiment. While the accuracies in most of the configurations were close to one another, those that used only single units tended to be weaker than the multi-clause units. Note that the paragraph configuration was tested with far more data than any of the others.

Our system makes no effort to account for lexical ambiguity. The uses we intend for our lexicon should provide some insulation from the effects of polysemy, since searches will be conducted on a number of terms, which should converge to one meaning. It is clear that in lists for key words with multiple senses, the dominant sense where there is one, appears much more frequently, such as "faculty ," where the meaning of "teacher" is more frequent than the meaning of "ability." Figure 1 shows the top 21 words in the sequence-of-six subjects, before the cross-referencing filter was applied. Twenty of the 21 entries were scored acceptable.

After the cross-referencing is applied, doctorate, education and revision were eliminated.

Config	OK	Wrong	Pct OK
Para	2003	183	92
Sent	1962	222	90
Sent+	2033	213	91
1 Clause	1791	218	89
1 Clause+	2004	198	91
2 Clause	2028	277	88
2 Clause+	2129	244	90

Table 7: Results of combining evidence; all configurations were combined with the sequence of six subjects

Config	OK	Wrong	Pct OK
Para +pp	4923	807	86
Sent	5193	990	84
Sent +pp	4876	775	86
1 Clause	5299	1233	81
1 Clauses +pp	5047	878	85
2 Clauses	5025	928	84
2 Clauses +pp	4668	728	87
Subj 2 Cl	5229	939	85
Subj 3 Cl	5187	860	85
Subj 4 Cl	5119	808	86
Subj 5 Cl	5003	764	87
Subj 6 Cl	4980	736	87

Table 8: Results with mutual information

The results from the single clause configuration (Figure 2) were almost as strong, with three errors, and a fair amount of overlap between the two.

The word "admiral" was more difficult for the experiment using the Dice coefficient. The list shows some of the confusion arising from our strategy on proper nouns. Admiral would be expected to occur with many proper names, including some that are spelled like common nouns, but the list for the single clause +pp configuration presented a puzzling list (Figure 3).

The sparseness of the data is also apparent, but it was the dog references that appeared quite strange at a glance: Inspection of the articles showed that they came from an article on the pets of famous people. Note that the dogs did not appear in top ranks of the sequence of subjects configuration in the Dice experiment (Figure 4), nor were they in the results from the experiments with cross-referencing, combining evidence and mutual information.

After cross-referencing, the much-shorter list for the Subj-6 configuration had "aviator", "break-up", "commander", "decoration", "equal-opportunity", "fleet", "merino", "navy", "pearl", "promotion", "rear", and "short".

The combined-evidence list contained only eight words: "navy", "short", "aviator", "merino", "dishonor", "decoration", "sub" and "break-up".

Using the mutual information scoring, the list in the Subj-6 configuration for admiral had only

faculty – trustee(51) 0.053; campus(41) 0.045; college(113) 0.034; member(369) 0.028; professor(102) 0.028; university(203) 0.027; student(206) 0.025; regent(19) 0.025; tenure(15) 0.025; chancellor(28) 0.023; administrator(34) 0.023; provost(12) 0.023; dean(27) 0.021; alumnus(13) 0.021; math(12) 0.017; **revision(8) 0.013**; salary(13) 0.013; sociology(7) 0.013; educator(11) 0.012; doctorate(6) 0.011; teaching(9) 0.011;

Figure 1: The top-ranked matchups for "faculty" from the Subj-6-Clause configuration before cross-referencing. The numbers in parentheses are the number of matchups and the real numbers following are the scores. Errors are in bold

faculty – trustee(31) 0.033; member(266) 0.025; administrator(31) 0.023; college(42) 0.012; dean(15) 0.012; tenure(8) 0.011; ivy(6) 0.011; staff(33) 0.01; semester(6) 0.01; regent(7) 0.01; salary(12) 0.01; math(7) 0.008; professor(31) 0.008; **load(6) 0.007**; curriculum(5) 0.006; **revision(4) 0.006**; **minority(11) 0.006**;

Figure 2: The top-ranked matchups for "faculty" under the single clause configuration. Errors are in bold.

nine words: "navy", "general", "commander", "vice", "promotion", "officer", "fleet", "military" and "smith."

Finally, the even-sparser mutual information list for the paragraph configuration lists only "navy" and "suicide."

5 Conclusion

Our results are encouraging. We were able to decipher a broad type of word association, and showed that our method of searching sequences of subjects outperformed the more traditional approaches in finding collocations. We believe we can use this technique to build a large-scale lexicon to help in difficult information retrieval and information extraction tasks like question answering.

The most interesting aspect of this work lies in the system's ability to look across several clauses and strengthen the connections between associated words. We are able to deal with input that contains numerous errors from the tagging and shallow parsing processes. Local context has been studied extensively in recent years with sophisticated statistical tools and the availability of enormous amounts of text in digital form. Perhaps we can expand this perspective to look at a window of perhaps several sentences by extracting the correct linguistic units in order to explore a large range of language processing problems.

admiral – navy(41) 0.027; ayalon(4) 0.024; cheating(5) 0.02; gallantry(3) 0.016; chow(4) 0.015; serviceman(4) 0.013; short(3) 0.013; wardroom(2) 0.012; american(2) 0.012; cnos(2) 0.012; self-assessment(2) 0.011; merino(2) 0.011; ocelot(2) 0.011; wolfhound(2) 0.011; igloo(2) 0.011; paprika(2) 0.011; spaniel(2) 0.01; medal(8) 0.01; awe(3) 0.01; pedigree(2) 0.009; terrier(2) 0.009;

Figure 3: Top-ranked matchups for "admiral" under the clause +pp configuration.

admiral – navy(88) 0.071; short(7) 0.03; promotion(11) 0.027; happiness(8) 0.026; fleet(11) 0.024; aviator(5) 0.022; ambition(8) 0.019; merino(3) 0.019; dishonor(3) 0.018; rear(4) 0.018; decoration(4) 0.015; sub(3) 0.013; airman(3) 0.013; graveses(2) 0.012; submariner(2) 0.012; equal-opportunity(2) 0.012; break-up(2) 0.012; commander(18) 0.012; pearl(7) 0.012; prophecy(4) 0.012; torturer(2) 0.012;

Figure 4: The list for admiral from the Subj-6 configuration.

6 Future Work

- We will have the scoring key itself evaluated by people who are not involved in the research.
- We are planning to conduct task-based evaluation in question answering.
- We are considering deploying a named entity module to provide some classification of which proper nouns should be counted and which should not.
- We plan to experiment with ways to incorporate using examining verbs and making use of surface objects in the configurations with sequences of clauses, as well as strengthen the finite state grammar.
- We will explore using the system to extract biographic information.

Acknowledgments

This material is based upon work supported by the National Science Foundation under grants Nos. IIS-96-19124 and IRI-96-18797, and work jointly supported by the National Science Foundation and the National Library of Medicine under grant No. IIS-98-17434. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Steven Abney. 1996. Partial parsing via finite-state cascades. In *Proceedings of the ESSLLI '96 Robust Parsing Workshop*.

Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for text summarization. In *Proceedings of the Intelligent Scalable Text Summarization Workshop*. ACL.

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

James P. Callan. 1994. Passage-level evidence in document retrieval. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference*, Dublin, Ireland. ACM.

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.

CELEX, 1995. *The CELEX lexical database — Dutch, English, German*. Centr for Lexical Information, Max Planck Institute for Psycholinguistics, Nijmegen.

Kenneth W. Church and Patrick Hanks. 1989. Word association norms, mutual information and lexicography. In *Proceedings of the 27th meeting of the ACL*.

Sanda M. Harabagiu and Steven J. Maiorano. 1999. Finding answers in large collections of texts: Paragraph indexing + adductive inference. In *Question Answering Systems*. AAAI, November.

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*.

Donald Hindle. 1990. Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting of the ACL*.

Yufeng Jing and W. Bruce Croft. 1994. An association thesaurus for information retrieval, tech. rep. no 94-17. Technical report, Amherst: University of Massachusetts, Center for Intelligent Information Retrieval.

G. Miller. 1990. Wordnet: An on-line lexical database. *International Journal of Lexicography*.

Thomas S. Morton. 1999. Using coreference for question answering. In *Proceedings of the Workshop on Coreference and Its Applications*, pages 85–89, College Park, Maryland, June. Association for Computational Linguistics, Association for Computation Linguistics.

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 Computation Linguistics*.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Amit Singhal. 1999. Question and answer track home page. WWW.

Frank Smadja. 1992. Retrieving collocations from text: Xtract. *Computational Linguistics*, Special Issue.

Tomek Strzalkowski and Jin Wang. 1996. A self-learning universal concept spotter. In *Proceedings of the International Conference on Computational Linguistics (Coling 1996)*.