Joining Statistics with NLP for Text Categorization

Paul S. Jacobs
Artificial Intelligence Laboratory
GE Research and Development Center
Schenectady, NY 12301 USA
psjacobs@crd.ge.com

Automatic news categorization systems have produced high accuracy, consistency, and flexibility using some natural language processing techniques. These knowledge-based categorization methods are more powerful and accurate than statistical techniques. However, the phrasal pre-processing and pattern matching methods that seem to work for categorization have the disadvantage of requiring a fair amount of knowledge-encoding by human beings. In addition, they work much better at certain tasks, such as identifying major events in texts, than at others, such as determining what sort of business or product is involved in a news event.

Statistical methods for categorization, on the other hand, are easy to implement and require little or no human customization. But they don't offer any of the benefits of natural language processing, such as the ability to identify relationships and enforce linguistic constraints.

Our approach has been to use statistics in the knowledge acquisition component of a linguistic pattern-based categorization system, using statistical methods, for example, to associate words with industries and identify phrases that information about businesses or products. Instead of replacing knowledge-based methods with statistics, statistical training replaces knowledge engineering. This has resulted in high accuracy, shorter customization time, and good prospects for the application of the statistical methods to problems in lexical acquisition.

1 Introduction

Text categorization is an excellent application domain for natural language processing systems. First, it is a task in which NLP techniques have born fruit, producing high accuracy along with other benefits [Hayes and Weinstein, 1990; Kuhns, 1990; Tong *et al.*, 1986]. Second, it provides an easy way of measuring success, by comparing system responses with "expert" category assignments. Third, it is a ripe domain for exploring statistical methods for automated knowledge acquisition. Published work on text categorization has focused on the first item above, arguing convincingly for knowledge-based techniques and their accuracy, but has not yet opened the way for the investigation of category assignment as a way of testing NLP methods or on the prospects for acquisition. This work focuses on combining statistics and NLP in a knowledge-based categorization system, using statistics as way of augmenting hand-coded knowledge.

The context of this research is a commercially-developed system [Rau and Jacobs, 1991] that automatically assigns categories to news stories for "custom clipping" and other markets. Like Construe/TIS [Hayes and Weinstein, 1990], the work derives from, and coordinates with, NLP efforts, but the system primarily uses a lexico-semantic pattern matcher for categorization [Jacobs *et al.*, 1991]. Categorization tasks vary greatly in difficulty, but the recall and precision results produced in our tests are similar to those reported by other systems, with coverage of over 90% on topic assignment and performance better than human indexers on most aspects of the task.

Figure 1 shows a typical example of a news story, with associated human-assigned categories. Retrieval is performed by matching a desired set of categories (termed a *query* or *profile*) against those assigned in the text database. Our system, known as NLDB, mimics these category assignments, extracting company names [Rau, 1991], topics or subject indicators, industries, and others (including, for example, stock exchanges and geographic regions). The program also incorporates portions of the SCISOR system [Jacobs and Rau, 1990], which can fill certain other fields, such as the target and suitor of a takeover.

This sort of system has a simple appeal: the "answers" (the set of category assignments) are usually clear-cut, yet they clearly require some detailed content analysis. On the other hand, the technologies that could contribute to this analysis are bafflingly complex, from discourse methods that distinguish topics from background events to word sense techniques that help to distinguish, for example, **COMMUNICATIONS** from **BROADCASTING** and **HEALTH CARE** from **PHARMACEUTICALS**.

Figure 2 shows the complete list of industry and topic assignments currently in use to categorize texts in the NLDB system.

The development of this system has advanced the state of the art in practical NLP by proving the utility of statistical training methods on a knowledge-based NLP task. Feeding in large volumes of texts with human answers has found new ways around old problems in knowledge acquisition. This paper explains the relationship between problems in NLP and performance in categorization and describes a statistical method for automatically creating lexico-semantic patterns for categorization.

Companies	Industries	Topics	Other
ARGOSYSTEMS	AVIATION	BUSINESS	CORPORATE
ARGOSYSTEMS INC	DEFENSE CONTRACTING	CONTRACT	NEWSGRID
BOEING	ELECTRONICS		NYASE
BOEING CO(BA)			OTC
UTL			USA
UTL CORP(UTLC)			

BOEING'S ARGOSYSTEMS SUBSIDIARY TO MAKE TENDER OFFER FOR ALL UTL CORP. SHARES

SEATTLE (JULY 31) PR NEWSWIRE - The Boeing Co. (NYSE: BA) has announced its agreement to cause its wholly-owned subsidiary ARGOSystems of Sunnyvale, Calif., to make a cash tender offer at $4.75 per share for all shares of UTL Corp. (NASDAQ: UTLC) of Dallas.

The transaction is valued at approximately $20 million. The boards of ARGOSystems and UTL have approved the transaction.

The tender offer will commence no later than Aug. 6. Upon completion of the tender offer, the agreement calls for a merger in which the remaining UTL stockholders will also receive $4.75 in cash per share. The tender offer is subject to certain conditions including the tender of at least a majority of the outstanding UTL shares.

UTL Corp. designs, develops, manufactures and markets electronic warfare systems used for reconnaissance and surveillance. The systems provide information on the location and identification of radar and communications emitters....

Figure 1: Input Text and Assigned Categories

2 Overall Results

Figure 3 summarizes the overall results of this experiment, including the results of assigning topic categories (the task generally reported in this sort of work) and industry categories with statistics only, natural language, and the combination; and the overall effect of the combined approach.

Recall here has essentially the same meaning as in information retrieval; i.e. the percentage of human-assigned categories that the system also produced. *Precision* is the percentage of system-assigned categories that also appeared in the human indices. These statistics make the dubious (and often incorrect) assumption that the human-assigned categories are always correct. In Figure 1, for example, NLDB included AEROSPACE on the industry list—This hurts precision because it is not included on the human list.

The major achievement here is that the combination of statistical analysis and natural language based categorization is considerably better than either alone. The system uses statistical methods where they do better (i.e. industries) and NLP where it does better (i.e. topics), and shows that combined NLP and statistics can be better than either technique alone within a particular task.

The results tend to understate the real impact of this combination, in part because of the large differences in difficulty among sets of categories, and in part because of the portion of the human-assigned categories that are incorrect. In analyzing sample texts where the human categories differ from the automatically-assigned categories, we have found that the system tends to be correct about as often as the human indexer, with many cases so difficult to judge that multiple independent assessments differ. Since this means that the system recall against the human is higher than human recall against the system, the results indicate that the system's overall performance is better than human performance. However, for the purpose of this experiment, we use these results to compare different system configurations, and not to illustrate an absolute measure of accuracy.

3 Categorization & Natural Language

While it is easy to attain a certain level of accuracy in text categorization using a single layer of techniques and to combine all texts and all categories in evaluating the results, this aggregation of results obscures many of the real problems where NLP and categorization come together. Each type of category can highlight a different set of NLP issues, and different texts reveal different processing problems. For example, the problem of mistaking a totally irrelevant text for a text of a particular category is very different from the subtle task of distinguishing texts about one category from another. Similarly, NLP results in processing texts *within* a category are quite different from results in assigning texts to categories, for reasons that will be explained below.

The best way to evaluate NLP techniques, therefore, is within the context of a more precise task than categorization in general, and as a complement to statistical methods. Even in component tasks where pure statistical methods tend to outperform pure NLP methods, NLP can play an important role in improving results, and statistics can play a role in improving NLP.

Industry Segments

advertising
aerospace
agriculture
autos
aviation
banking
beverages
biotechnology
broadcasting
building + material
business + services
chemicals
computers
construction
consumer + products
defense + contracting
educational + services
electronic + publishing
electronics
entertainment
environmental + services
financial + services
food
forestry + products
freight
health + care
industrial + products
insurance
machinery
metals
mining
nuclear + energy
office + equipment
personal + care + products
petroleum + products
pharmaceuticals
photography
plastics
precious + metals
publishing
railroads
real + estate
restaurants
retail
rubber
ship building
telecommunications
textiles
tobacco
toys
travel services
trucks
utilities

Subject Indicators

air + force
antitrust
appointment
bankruptcy
boycott
budget
business
cabinet
capitol
career
chg-naq
commodity
congress
contract
corporate
coup
crime
debt
deficit
democracy
depression
divestiture
dividend
earnings
economy
election
executive change
expansion
export
government
import
inflation
insider + trading
joint venture
labor
lawsuit
layoff
legislation
market
merger
military
money
nasd halt
nasd resume
navy
new product
news
newsbrief
prime + rate
public offering
recession
refinancing
resignation
restructuring
socialism
space
strike
taxes
trade
unemployment

Figure 2: Keywords for Industry and Topic Segments

Categorization Task	*Recall*	*Precision*
Topic assignment (NL)	.94	.61
Topic assignment (Stats)	.73	.79
Topic assignment (Stats + NL)	.95	.65
Industry assignment (NL)	.34	.18
Industry assignment (Stats)	.64	.49
Industry assignment (Stats + NL)	.67	.50
All categories (NL)	.74	.46
All categories (Stats + NL)	.79	.64

Figure 3: Overall Results

3.1 Word Weights for Industry Assignment

For example, in the news categorization task described earlier, natural language delivered the weakest performance relative to statistics on the assignment of industry categories. Performance on topic assignment was generally much higher using natural language, and company name extraction and variation was handled using a separate mechanism [Rau, 1991]. The two most obvious differences between topic assignment and industry assignment are:

1. It is generally easy to determine where the topic of a story is expressed, either in the first sentence or by spotting certain words and phrases that are good indicators, while the industries involved can appear almost anywhere in a story,
2. The breadth of language that expresses topic is much narrower than the breadth required to handle the different industries; for example, it is quite easy to identify texts dealing with bankruptcies, lawsuits, and mergers using a few key words and phrases (vocabularies of no more than 20 or 30 words per topic), but a single industry can be indicated by any number of words or expressions, including names of specific customers, products, or devices.

For these reasons, statistics have the upper hand in the identification of industries, and the first pass at the NLDB system used linguistic methods for topic assignment and statistical methods for industries. This was unsatisfying, because we quickly came across errors in the results that might have been prevented using simple NLP methods, as well as places where the results could have helped to augment or correct linguistic knowledge

The statistical methods, which will be described later involve weighting individual words and phrases according to their value in distinguishing industries. The most

obvious errors resulting from these methods were finding a good indicator in the wrong place (either in irrelevant text or in background text), and finding a good indicator used in a different way. These errors pervade the results of statistical categorization, with the most obvious problems coming when the results clearly derived from the misinterpretation of individual words and phrases.

For example, the word *brewing* occurred 14 times in a sample of about 11,000 news stories. In 9 out of those 14 cases, or 64%, the story was correctly categorized under the industry **BEVERAGES**, which includes less than 1% of the stories. By most any statistical metric, *brewing* is a strong indicator for **BEVERAGES** (better, in fact, than *beer* and *beverages*, although not quite as good as *Pepsi*). However, the statistical categorization method failed, for example, on the following text, incorrectly assigning the text to **BEVERAGES**:

> The issue first surfaced Monday when Dawes complained there is a "black hole" of information about how Richards deposited state money while the S&L crisis was *brewing* [1].

The word *gas* is not quite as good an indicator as *brewing*—in 55% of occurrences it indicates **PETROLEUM PRODUCTS**, and 11% of the time **UTILITIES**, with scattered other interpretations. But *gas* is much more frequent than *brewing*, occurring 835 times in the same training sample where *brewing* occurred 14 times. So, in terms of overall performance, knowing when *gas* is a good indicator of an industry can make more of a difference. The problem is with texts such as the following:

> Of 55 check-ups of the 17 patients, mild diarrhea was reported during 2 percent of check-ups, nausea or vomiting in about 3 percent and a moderate increase in intestinal *gas* in about 10 percent.

While the problem with *brewing* above can easily be solved by using any simple method of filtering out irrelevant text (the sentence appears in the middle of a story about a political campaign), this is not the case with *gas*. The *gas* example, like many similar errors, appears in relevant text describing the health effects of bran cereal, which could be correctly categorized as **HEALTH CARE** and **FOOD**.

Note also that it is difficult to compensate for errors in individual word weights by using combined statistical weights for categorization. This point will be discussed more in Section 5, but the main problem is that content words like *patients*, *nausea* and *check-ups* simply don't have enough information content to act as good discriminators compared to *gas*.

3.2 Recall and Precision

While it is easy to spot places where statistics tend to introduce erroneous categories, thus lowering *precision* in examples such as *brewing* and *gas* above, it is harder to understand why statistical methods also fail to produce enough information to assign a category, thus producing low *recall*. Since the task of assigning industries depends more on what businesses companies are in than what an individual story is about, the information about the industry is often localized, perhaps even in a single mention of a company or product. For example, the following is a typical, though difficult, story about an executive change:

> ....James W. Nelson has been named vice president-manufacturing and distribution for the household products group at Lehn & Fink Products, maker of such well-known brands as Lysol, Love My Carpet, Resolve, Chubs, Mop & Glo, Ogilvie, Minwax and Thompson's.
> ....
> Lehn & Fink Products, headquartered in Montvale, is a leading international marketer of household and do-it-yourself products.

The human categorizer assigned the story to the categories of **CONSUMER PRODUCTS** and **PERSONAL CARE PRODUCTS**, although the latter is probably an error. While **CONSUMER PRODUCTS** is the strongest category indicated statistically, from words such as *household* and *brands*, it is still weakly indicated; in fact, it would be difficult to get a statistical measure to admit **CONSUMER PRODUCTS** without also including **RETAIL**, **BEVERAGES**, **BUILDING MATERIAL**, and even **TEXTILES**, which are loosely coupled with terms such as *brands*, *do-it-yourself* and *carpet*. It is also quite difficult, because of the high independent frequencies of the words, to identify *household products* as a collocation or combination that should be considered.

The key to getting good recall and precision on texts such as these is to consider the weights of the individual words and phrases to determine *what* industries could be involved, but to use the structure of the texts to help determine *where* the industry information might be. Phrases like *X is a leading marketer of Y* or *X is the maker of Y* appear throughout news stories, and are sure indicators of industry information, even though they do not point to any particular industry. Linguistic approaches probably won't help to guess that *Love My Carpet* is a consumer product, but they *can* help to determine that the industry discriminators lie in the text following patterns such as *X is the maker of Y*. Statistics can then guess the industries associated with *Y*.

The NLP method used in NLDB associates categories with linguistic patterns. We will next describe the pattern language, then explain how statistical methods can automatically add simple patterns.

4 Lexico-Semantic Patterns

In SCISOR [Jacobs and Rau, 1990], MUC [Jacobs *et al.*, 1991; Krupka *et al.*, 1991], and other applications, we have found that lexically-driven pre-processing serves as a complement to parsing and semantic interpretation, both in identifying portions of relevant text and in marking the input text to make it easier to process. Our lexico-semantic pattern rules are quite similar to those in CONSTRUE/TIS [Hayes and Weinstein, 1990], associating each pattern with an action rule that can manipulate text or activate or de-activate a category. This

[1] italics added

type of knowledge structure has proven effective for topic identification as well as other forms of pre-processing.

Because the pattern matcher is designed as an efficient "trigger" mechanism and an aid in parsing, the patterns are mostly simple combinations of lexical categories. The patterns largely adopt the language of regular expressions, including the following terms and operators:

- Lexical features that can be tested in a pattern
- Logical combination of lexical feature tests—OR, AND , and NOT
- Wild cards
- Variable assignment (*?X =*)
- Grouping operators— <> for grouping, [] for disjunctive grouping
- Repetition— * for 0 or more, + for 1 or more
- Range— *N for 0 to N, +N for 1 to N
- Optional Constituents— {} for optional

While this pattern language provides a tool for recognizing linguistic constructs, most patterns for categorization are simple lexical items, semantic categories, or combinations. For example, the word root *dividend* and the phrase *holders of record* are good indicators of a **DIVIDEND** story.

Most topics have rules that include such sure-fire single words and phrases, along with some more complex patterns. For example, the following two rules help to recognize stories about mergers and acquisitions:

```
(or tender merger) offer => C-TAKEOVER ;

;;  C1 ... announced ... acquisition of ... C2
?C1={cname} * $announce-verb * $merger-verb
*5 [of with] $ ?C2={cname}
   => (C-TAKEOVER (r-agent ?C1) (r-target ?C2)) ;
```

In addition to helping to catch a broader range of constructs that indicate takeovers, the more complex patterns like the second one above can make preliminary assignments of roles, which can greatly speed and aid parsing in systems that perform both parsing and categorization. The ability to construct and add these more sophisticated patterns by hand is a major advantage, which accounts for much of the benefit of knowledge-based methods over statistical means. However, the more simple patterns are required, the more labor-intensive this process can be, and the more manual tuning must be done in order to get accurate results.

Statistical methods have the advantage of building rules automatically from a training set. But, in order to get the benefit of statistics, the methods used must add knowledge in the same form as the knowledge-based rules, and must produce a clear result that is accessible for knowledge engineering. In other words, the statistical methods themselves must be an aid rather than a replacement for knowledge acquisition. The next section describes how this is accomplished.

5 Statistics and Acquisition

The strategies for statistical training described here all use a "training" set of 11,500 news stories, including about 3,000,000 words, with human-assigned categories assigned to each story. The "test" set used was a new sample of one day's news, or 700 stories including 200,000 words.

Many statistical methods in information retrieval use probabilistic weights of individual *terms* [Salton and McGill, 1983], where a term can be a single word, root, or combination of words. The techniques we explored include various weighting schemes, with the end goal being to use heavily-weighted terms as the building blocks for patterns. A term can be weighted with respect to its overall relevance, or with respect to its ability to determine a particular category. The "pure" statistical results reported earlier summed the weights of all terms in a text with respect to each industry category.

Automating the process of acquiring lexico-semantic patterns poses a number of distinct problems. First, there has to be some means of distinguishing *where* industry information might appear in a text. Second, single words should be distinguished from phrases in certain cases where the individual words might be misleading. Third, the statistical methods must produce individual actions, not weights that must be combined to derive the final answer or answers. Fourth, there has to be some good way of determining when the statistical results were bound to introduce errors.

All of these requirements come together to assure that the statistical methods can be used to improve existing sets of pattern-action rules, that manual methods will not counteract the results of acquisition, and that the statistical and NLP methods can interact gracefully.

5.1 Identifying Relevant Text

Many of the errors with statistical methods, especially when single words and phrases are used for categorization, come from unusual occurrences in background or irrelevant texts, such as the *brewing* example earlier. A similar case is the word *Yankee* which is an excellent indicator of **NUCLEAR ENERGY** (because of the *New Hampshire Yankee Power Plant*), except in an isolated cluster of articles about George Steinbrenner, the owner of the New York Yankees.

We tried two methods of correcting for such problems. First, we noted that most industry information is contained in the headline, first, and last paragraphs of texts, and tested using only these paragraphs for categorization. Second, we tried a simple filter to score how much each paragraph was like a first or last paragraph, using the following calculation for the relevance weight of a term:

$$1000^d \log_2 b$$

Where b is number of occurrences of each term in the first and last paragraphs of text, and d is the ratio of occurrences in these texts to all occurrences, minus a constant.

The following are some of the noteworthy results of these tests:

- A separate relevance filter produced some improvement over simple term weighting, about 3 points in precision, and a larger improvement, about 5 points, when only "in-or-out" patterns were used. This was above and beyond a comparable gain from considering only headline, first, and last paragraphs.
- Almost all combinations of looking at first, last, and additional relevant paragraphs ended up with about the same results. In other words, looking at only paragraphs with high relevance scores that were *also* at the beginning or end of a story produced about the same results, combining recall and precision, as looking at highly relevant paragraphs *in addition* to the first and last paragraphs. We settled on using only first and last paragraphs with relevant scores because this simply minimizes the amount of text that must be processed.
- Using only first and last paragraphs for training, as well as categorizing, produced no improvement in results. We think that this is because the improvement from having a more accurate training sample is neutralized by having *less* text to train on. This suggests using a still larger training sample.

These tests showed that a simple relevance filter produced a consistent, small, advantage in accuracy over using the whole text of each story. However, surprisingly, it was hard to see any gain from considering the middle parts of stories where those parts had high relevance scores. This suggests that further work could produce better indicators of relevance that get industry information out of the middle parts without introducing more extraneous categories.

5.2 Collocations and Names

Statistical methods looking for sure-fire indicators must, by their nature, consider overwhelming statistical evidence even when these indicators are relatively infrequent. Unfortunately, even when a word or phrase correlates with an industry 100% of the time in a sample, this does not mean that it will be a sure indicator in a new sample. This is especially a problem with proper names and names of locations, but is also an issue with words that occur frequently in collocations.

Treating words as individual indicators when they really are part of a name or collocation can hurt precision by increasing the chance that the single word will appear independently. It can also hurt recall, because the combined evidence derived from a name or collocation can be strong even when the individual words contribute little.

As evidence of the problem with proper names, the words *Flint* (a city in Michigan where General Motors produces cars and trucks), *Donahue* (the name of a popular daytime TV show), and *Warner* (a communications conglomerate) are all good statistical indicators. In fact, in the training sample of about a month's worth of news, *Donahue* was an indicator of **BROADCASTING** 18 out of 18 times, making it a better indicator even than *Pepsi*. But in the one-day test sample, *Donahue* occurred only once, in a story about the president of Nike (the athletic shoe company). *Flint* occurred in a story about an art display in the Michigan town, and *Warner* as a name unrelated to the communications company.

While accurate recognition of proper names is necessary for good precision, it is also a requirement for recall. A company may be involved in industries that are difficult to get from either the parts of the name or the surrounding context. For example, *household* is a weak indicator for **CONSUMER PRODUCTS**, but the company *Household International* is in the **INDUSTRIAL PRODUCTS** category. Similarly, *Digital Equipment* is in a different industry from *Digital Communications*. Since these names are so important for industry assignment, we found that the best results came from handling all proper names separately and being much more lenient about when to admit an industry name based on the name of a company than for individual words and phrases.

With other compounds, there were again problems with both recall and precision. Recall problems came from common words that have a special meaning when conjoined, such as *real estate*—both *real* and *estate* frequently occur, and have no significance with respect to industry, but *real estate* is a good indicator of the **REAL ESTATE** category. *Television* is a weak indicator of several industry categories, but *television viewers* and *television networks* are strong indicators.

We took a simple approach to handling such collocations, by computing mutual information statistics [Church *et al.*, 1989] for bigrams (two-word sequences) in the training corpus, and treating combinations with a high degree of mutual information as if they were single terms. So *real estate* would be treated as an atom, the individual words not being considered for categorization. This yielded fair results, but probably is not as good a method as looking for discriminators specifically when the individual words are indicators of multiple categories.

The following are some of the key results from separate processing of names and collocations:

- Company name extraction was the best contributor to accuracy, with a 10% improvement in recall and no significant loss of precision over treating company names as any other word. This is an especially compelling result because the individual components of company names are often themselves good indicators, and because the size of the training set is not nearly large enough to cover many of the companies that occur in a given test (hence the training data inherently miss a fair number of companies).
- The use of bigrams yielded about a 6% gain in precision with no real effect on recall. We expect that the number of bigrams was not adequate to have a major positive impact on recall, while the method of ignoring the individual component words can neutralize some of the positive effect.

5.3 Setting Word Thresholds

Knowledge-based methods aim at high-accuracy individual patterns. It is hard to balance these against weighted terms, so it is best to tune the statistics to identify simple indicators rather than weights to be combined with

other weights. This way, a rule can combine hand-coded knowledge with automatically-acquired data by looking for industry information in a particular place and getting the industry from a single indicator in that place, as in *the company manufactures satellites.*

Because work in information retrieval [Salton and McGill, 1983] has suggested that combinations of weighted terms could be more accurate than single in-or-out assignments, we compared a number of different weighting methods with a number of different methods for discriminating key indicators. We found that, in general, the combination of weighted terms produced better results than simply taking the union of the industries activated by the "best" terms. However, the results for the best discrete assignment of industries to individual terms were very close to those of the weighted terms, and the benefits of this approach—including the ability to integrate statistics with knowledge-based methods, the identification of important ambiguities in word meaning, and speed and simplicity—suggest that statistical thresholds for individual terms, without any combining of weights, is a good approach.

We had to devise a statistical means of distinguishing only those terms that were very good indicators *by themselves* of a particular category, without having to compute a score for each paragraph. This would not have worked without the pre-processing of relevant text, name and collocation extraction. In fact, the performance of categorization using single in-or-out terms was more than 10% lower in precision than the combination of weights without the pre-processing, but about the same with the combined method. The apparent explanation for this is that most of the error introduced in using single terms as discriminators comes from the confusion of terms either in special combinations or in irrelevant text.

The following is the formula for weighting terms that are individual indicators of a particular category:

$$(200^d)(\log_2 b)(\log_2 r)$$

where b is number of times a term appears is a story about a particular category, d is the difference between that number and the overall percentage of words in texts of that category, and r is the ratio of combined probabilities to the product of independent probabilities, the mutual information statistic.

With a threshold of 100, this weight identifies terms that are good independent indicators of each topic. These terms can then be used automatically, either by themselves or in combination with other terms, to create patterns in the knowledge base.

5.4 Combining Information

The statistical pre-processing methods and calculations of relevance weights and weights for category indicators lay the groundwork for automatically constructing linguistic patterns for categorization. Because these individual discriminators can be combined with hand-coded knowledge, the statistical recognition of these terms is sufficient to augment a knowledge base automatically, and the combination of the hand-coded rules with the statistical patterns is better than either alone (although it could always be argued that, with a little more work, the same rules could have been hand-coded).

These results are not completely satisfying. The statistical acquisition method uses only a fraction of the power of the pattern language, and the error rate of the system could still be reduced. We tried three different types of methods–finding *exceptions*, *co-indicators*, and *meta-indicators*–to try to improve results using combinations rather than single term rules. These three methods are described as follows:

Exceptions: Exceptions spot secondary terms that would override a category indicated by a "good" term, for example, if *oil* appears in a text about the automobile industry rather than petroleum, it might appear near *motor* or *engine.*

Co-indicators: Co-indicators spot combinations of words, where at least one was a "good" term, where the combination was a much better indicator that the single term.

Meta-indicators: Meta-indicators spot terms such as *produces* and *manufactures*, which are not themselves indicators of a particular industry, but often appear near terms that indicate an industry.

For this process, we compiled a set of tables covering, for each good discriminator, the words that appeared as neighbors of that term along with the number of times those words appeared in texts about the "right" category vs. texts not about the category. This was a computation-intensive process for 3 million words, so much so that we had to reduce the size of the table by considering only terms with moderate frequency—the highest frequency terms are "stop" words, and the lowest frequency terms are not good discriminators.

The following are the major results of this analysis:

- Using exception lists to try to correct for precision problems turned out to be of surprisingly little value (about 2% precision). While it is possible that different methods would yield better results, it seems that the data on exceptions are just too sparse—it is much easier to get good data on positive examples from the training set than negative examples.
- Using co-indicators, or second order relations, appeared to be much more promising than exceptions. Like the use of bigrams, this produced only a small effect (about 2% in both recall and precision), but any technique that improves recall without a loss of precision is worth exploring.
- The use of meta-indicators, while also not producing a major effect on results, looks like the most promising method. Like co-indicators, these meta-indicators rely mainly on positive examples from the text, but they have the additional advantage of being able to use much larger volumes of data.

6 Fertile Areas for Future Research

While the improvements in overall system performance on this task came as a result of many months of engineering and experiments, the most promising aspect

of this evaluation is the prospect for new areas where statistics can help natural language and *vice versa*. We have identified three critical research areas that are likely to improve both system performance and general NLP performance.

The first major area of research is in discourse, or text structure analysis. This experiment showed that the first and last paragraphs of a news story give much more accurate information than others, and that a general assessment of relevance of a paragraph serves as a good filter for the extraction of information from that paragraph. However, this simple filter falls far short of really identifying where the information in a story lies. For example, in many stories the "last" paragraph really comes in the middle of a text, with some additional material coming at the end because of incidental information or strange editing. Similarly, in more complicated texts, there can be more than one introductory paragraph, multiple concluding paragraphs, and other critical information in the middle. Statistical techniques are a very weak means of guessing this type of structural information. We expect that we can improve the results slightly with some more sophisticated discourse analysis; and, perhaps more importantly, this type of evaluation can measure how well a structural theory of text can perform.

The second area to explore is developing generalized patterns from detailed statistical analysis. The first-order logarithmic measures used for acquisition here are overly simplistic, and assume that relationships between words and categories are basically independent. This assumption is false, because the words are a manifestation of concepts and linguistic relations in the text that the statistics are ignoring. We are investigating a variety of more complex means, including multivariate discriminant analysis, that help to determine when, for example, the effect of combining *satellite* with *weapon* is really an effect of combining *satellite* with *any* military concept.

The third, more linguistic, area is in identifying thematic roles. In meta-indicators described earlier, we are really identifying potential function words in the text, such as *produces* or *manufactures*. Since the use of these terms in category assignment really assumes that any company or industry appearing around them is involved in the function or operation, it should be possible to use more detailed parsing to check whether this assumption fits linguistically, and to use the statistical analysis to acquire functional or thematic relations that can help in more detailed analysis.

7 Summary and Conclusion

This paper has addressed the area where statistical and linguistic analysis come together with an application focus — the assignment of categories to news stories, particularly the identification of topics and industries. The experiments reported here show that using statistical methods to acquire simple lexical patterns helps knowledge-based processing and leads to a substantial improvement in overall system performance. In addition, this method promises to ease the burden of hand-coding knowledge for each application, by automatically identifying the significant terms and combinations of terms to use knowledge-intensive NLP applications.

References

[Church *et al.*, 1989] K. Church, W. Gale, P. Hanks, and D. Hindle. Parsing, word associations, and predicate-argument relations. In *Proceedings of the International Workshop on Parsing Technologies*, Carnegie Mellon University, 1989.

[Hayes and Weinstein, 1990] Philip J. Hayes and Steven P. Weinstein. CONSTRUE/TIS: A system for content-based indexing of a database of news stories. In *Proceedings of the Second Annual Conference on Innovative Applications of Artificial Intelligence*, May 1990.

[Jacobs and Rau, 1990] Paul Jacobs and Lisa Rau. SCISOR: Extracting information from on-line news. *Communications of the Association for Computing Machinery*, 33(11):88–97, November 1990.

[Jacobs *et al.*, 1991] Paul S. Jacobs, George R. Krupka, and Lisa F. Rau. Lexico-semantic pattern matching as a companion to parsing in text understanding. In *Fourth DARPA Speech and Natural Language Workshop*, pages 337–342, San Mateo, CA, February 1991. Morgan-Kaufmann.

[Krupka *et al.*, 1991] George R. Krupka, Paul S. Jacobs, Lisa F. Rau, and Lucja Iwańska. Description of the GE NLToolset system as used for MUC-3. In *Proceedings of the Third Message Understanding Conference (MUC-3)*, San Mateo, CA, May 1991. Morgan Kaufmann Publishers.

[Kuhns, 1990] Robert Kuhns. News analysis: A natural language application to text processing. In Paul S. Jacobs, editor, *Text-Based Intelligent Systems: Current Research in Text Analysis, Information Extraction, and Retrieval*, pages 147–150. September 1990. GE Research and Development Center Report CRD90/198.

[Rau and Jacobs, 1991] Lisa F. Rau and Paul S. Jacobs. Creating segmented databases from free text for text retrieval. In *Proceedings of the 14th International Conference on Research and Development in Information Retrieval*, pages 337–346, October 1991.

[Rau, 1991] Lisa F. Rau. Extracting company names from text. In Tim Finin, editor, *Sixth IEEE Conference on Artificial Intelligence Applications*. IEEE Computer Society Press, Miami Beach, Florida, February 1991.

[Salton and McGill, 1983] G. Salton and M. McGill. *An Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

[Tong *et al.*, 1986] Richard M. Tong, L. A. Appelbaum, V. N. Askman, and J. F. Cunningham. RUBRIC III: An object-oriented expert system for information retrieval. In *Proceedings of the 2nd Annual IEEE Symposium on Expert Systems in Government*, pages 106–115, Washington, DC., October 1986. IEEE Computer Society Press.