A News Story Categorization System

Philip J. Hayes, Laura E. Knecht, and Monica J. Cellio

Carnegie Group Inc
650 Commerce Court at Station Square
Pittsburgh, PA 15219

Abstract

This paper describes a pilot version of a commercial application of natural language processing techniques to the problem of categorizing news stories into broad topic categories. The system does not perform a complete semantic or syntactic analyses of the input stories. Its categorizations are dependent on fragmentary recognition using pattern-matching techniques. The fragments it looks for are determined by a set of knowledge-based rules. The accuracy of the system is only slightly lower than that of human categorizers.

1. Introduction

The large economic potential of automatic text processing is leading to an increasing interest in its commercial applications. This paper describes a pilot version of a commercial application of natural language processing techniques to the problem of categorizing news stories into broad topic categories.

The conventional way to process natural language texts is to have people read them and perform some action based on what they have read. People, for instance, currently categorize news stories for routing purposes and extract information from banking payment telexes so that transactions can be executed.

Unfortunately, using people tends to be:

- **slow** - people read text slowly;
- **expensive** - if the volume of text is high, processing it requires the efforts of many people;
- **inconsistent** - it is very hard to get a group of people to make consistent decisions about text.

In many cases, the proper processing of text is central to a company's revenue stream, so that improvements in the processing can provide major leverage and justify major contract system expenditures.

Automatic text processing offers the possibility of such improvements in all three areas. A single text processing machine can potentially do the job of several people faster, cheaper, and more consistently.

This paper describes an implementation of a system to do text categorization. The texts it operates on are news stories, but similar techniques could be employed on electronic mail messages, telex traffic, technical abstracts, etc.. Once categorization has been accomplished, the results can be used to route the texts involved to interested parties or to facilitate later retrieval of the texts from an archival database.

The system described here uses the well-established natural language processing technique of pattern-matching [1, 5]. Since the input to the system is an arbitrary news story on any topic whatsoever, no attempt is made to perform a complete syntactic or semantic analysis. Instead, categorization is based on the presence of particular words and phrases in particular lexical contexts. As the more detailed description in Section 3 will make clear, however, the approach used goes well beyond the keyword approaches used in information retrieval (e.g. [6]). In particular, the words and phrases the system looks for and the context in which it looks for them are specified through a modified version of the powerful pattern matching language used in Carnegie Group's Language Craft™ product[1] [3]. Moreover, the system determines which words and phrases to search for in a given story and how to interpret the presence of these words and phrases according to knowledge-based rules.

As simple as these techniques are by current natural language processing standards, the accuracy of the system is high. As described in more detail in Section 4, the system had an average accuracy of

[1]Language Craft also uses caseframe parsing techniques for complete linguistic analyses.

93%[2] on a sample of 500 random stories that had not been previously processed by the system or seen by its developers. Moreover, this accuracy was obtained without sacrificing computational efficiency. The average processing time was 15 seconds per story[3] on a Symbolics 3640, a figure which we believe could be considerably improved through a detailed performance optimization which we have not performed.

The remaining sections of the paper describe in more detail: the problem tackled by the system, the approach used, and the results obtained.

2. The Problem

The primary goal in developing the system described in this paper was to demonstrate the feasibility of categorizing news stories by computer in small amounts of time (a few seconds) using natural language processing techniques. The specific task chosen to do this was emulation of the performance of a group of human categorizers. Our raw material was a data base containing many thousands of news stories that had been hand-categorized[4] for any of 72 categories. Our system was required to assign 6 of the 72 categories: acquisitions/mergers, metals, shipping, bonds, war, and disorders. A story could be assigned one or more of these codes, or no code at all if none of the chosen six was appropriate. The restriction to six codes was imposed to keep the effort required to build the system within certain budgetary limits. As Section 3 will show, the approach taken is equally applicable to the larger set of categories.

Modelling the categorizations produced by human beings presented some difficulties. To summarize:

- The text processing techniques used in the system were oriented to identifying concepts explicitly mentioned in a story. They were not well suited to identifying the class of people that a story might be of interest to. The human categorizers of the stories in our data base used both these kinds of considerations when they assigned topic codes to stories.
- Some topic codes had relatively vague, subjective definitions.
- The human categorizers were not always consistent in the way they made their topic assignments.

The news stories themselves posed another challenge. Though the set of topics to be assigned by the system was narrowed from 72 to 6, there was no parallel narrowing of the stream of stories that would serve as input to the system. The full range of story types found in a newspaper occurred in the data base of news stories. As a consequence, our task was not the relatively simple one of, for instance, distinguishing a story about war from one about bonds. War stories also had to be distinguished from military, disaster, crime, diplomacy, politics, and sports stories, to name just a few.

It was often the case that we could characterize the kind of stories that might mislead the system. We were prepared for sports stories that looked like metals stories (*"...captured the gold medal at the summer Olympics..."*) or like war/disorders stories (*"...the battle on center court at Wimbledon..."*). A more difficult challenge was posed by words and phrases that were good predictors of a particular topic but occurred randomly across all story types, sometimes with the same meaning, sometimes not. For instance, the noun *note*, in the sense of financial instrument, was useful for finding stories about bonds; however, numerous, random stories used that word in a different sense. Metaphorical language was also a problem -- not use of fixed phrases (we had no trouble failing to assign the category **metals** to a story that contained the phrase *like a lead balloon*) -- but rather creative metaphorical language. So, a story about a series of battles in the continuing disposable diaper war between Proctor and Gamble and its competitors was assigned to the disorders category.

---

[2]More precisely, its average recall was 93% (i.e. it made 93% of the topic assignments it should have made) and its average precision was also 93% (i.e. 93% of the topics it did assign were correct).

[3]The average story length was 250 words; stories varied from about a 100 to about 3000 words.

[4]The hand-categorizations were done by a group of people who had no involvement with or knowledge of the system we developed.

3. Approach

3.1. Overview

The system tackles story categorization in two distinct phases:

- **hypothesization**: an attempt to pick out all categories into which the story might fall on the basis of the words and phrases it contains; if particular words and phrases suggest more than one category, they will contribute to the hypothesization of each of these categories;
- **confirmation**: an attempt to find additional evidence in support of a hypothesized topic or to determine whether or not the language that led to the topic's being hypothesized was used in a way that misled the system; it is this phase, for instance, that would detect that conflict vocabulary was being used in the context of a sports story and disconfirm the **war** and **disorders** categories for that story. This phase thus has an expert system flavor to it.

Both phases use the same basic kind of processing: a contextually limited search for words and phrases using pattern-matching techniques. They are also both organized (conceptually) as a collection of knowledge-based rules. The phases differ only in the directedness with which the search is conducted. Hypothesization always looks for the same words and phrases. Confirmation looks for different words and phrases using specific knowledge-based rules associated with each of the topics that have been hypothesized.

The search for words and phrases in both phases is organized around *patternsets*. A patternset represents a collection of words and phrases that are associated with a given concept, such as *conflict*. The concepts associated with patternsets sometimes correspond to the topics we are trying to categorize stories into, but they may also be more specific or may span several topics.

The basic operation on a patternset is to determine how many of the words and phrases it represents appear in a story. System actions are taken when the number of matches crosses a threshold, at which point we say that the patternset has matched. The thresholds are empirically determined and differ from patternset to patternset and even from use to use of the patternset.

Hypothesization is typically performed on the basis of matches of single patternsets. Confirmation rules typically involve branching conditions depending on the results of multiple patternset matches. Individual patternsets may be involved in both hypothesization and confirmation phases.

The remainder of this section describes the operation of the system in greater technical detail.

3.2. Patterns and Patternsets

Patternsets are collections of *patterns*. A pattern is an expression in a pattern-matching language that corresponds to one or more words and phrases that might appear in a story. A pattern is said to match the story if any of the words or phrases that it specifies appear in the story. Each pattern has a weight, either *probable* or *possible*, with matches of *probables* counting more than matches of *possibles*, according to a scheme explained below. Patterns also have names.

The following pattern, called "titanium", will match the word *titanium* and assign the match a weight of "probable".

```
(titanium) -> probable
= titanium
```

Eight operators are available to allow individual patterns to specify several words and phrases. They are:

- **?**: specifies an optional subpattern;
- **!** and **!!**: specify alternatives (i.e. they both mean "or");[5]
- **~** and **&not**: specify a subpattern that should not be matched;[6]
- **&skip**: specifies the maximum number of words to skip over;
- **+N**: specifies that a word is a noun and can therefore be pluralized;
- **+V**: specifies that a word is a verb and can therefore occur with the full range of verbal inflections.

The following examples illustrate how these operators are used.

---

[5]The operator !! is more efficient than !, but there are some situations where it cannot be used.

[6]The operator &not filters out a subpattern to the left of the subpattern to be matched; ~ filters out a subpattern to the right.

- (par (pricing !! ?issue price))
  -> probable
  = parprice
  [This rule matches the phrases *par pricing*, *par issue price*, and *par price*.]

- ((&not ratings) war +N) -> possible
  = war
  [This rule matches *war* or *wars* preceded by anything except the word *ratings*.]

- (sell +V (&skip 6 (company !! business !! unit))) -> possible
  = sell-co
  [This rule matches any form of the verb *sell* followed by *company*, *business*, or *unit*, with as many as 6 words intervening.]

The pattern operator **&skip** deserves special comment. It allows us to find key expressions even when it is impossible to predict exactly what extraneous words they will contain. Consider, for example, the phrases *sell the business* and *sell the unit*; these phrases must be matched if the system is to detect stories about acquisitions. The problem is that expressions like *sell the business* are rare. Examples of the sorts of phrases that we actually find in acquisitions stories are given below:

sell the Seven-Up business
sell the ailing Seven-Up unit
sell its Seven-Up soft drink business
sell 51 pct of the common stock of its unit
sell the worldwide franchise beverage business
sell about 5 mln dlrs worth of shares in the company

With **&skip**, we can look for the verb *sell* followed by *company*, *unit*, or *business* without having to specify what the intervening words might be.

In addition to pattern operators, a set of **wildcards** is also available to rule-writers for matching words that cannot be specified in advance. **$** is the general wildcard: it matches any single word or other symbol. **$d** matches any determiner (*a*, *the*, *this*, etc.); **$n** matches any number; **$q** matches any quantifier (*much*, *many*, *some*, etc.); and **$p** matches any punctuation mark.

3.3. Hypothesization and Confirmation

After a story has been read in, the system begins the process of topic determination by applying its hypothesization rules. A hypothesization rule tells the system to hypothesize one or more specified topics if a given patternset matches the story with a strength greater than a given threshold.

For example, one of the system's hypothesization rules specifies that the topics **war** and **disorders** should be hypothesized if the score for matches in the "conflict" patternset is 4 or greater; another rule specifies that the **metals** topic be hypothesized if the "metals" patternset matches with a score greater than 2. The thresholds for each rule are determined empirically based on the rule developer's observation of the performance of the system when different thresholds are used. Note also that there is not necessarily a direct correlation between topics and patternsets; some patternsets could provide evidence for more than one topic, and some topics could make use of more than one patternset.

The scores for patternset matches are calculated according to the formula:

```
(2 x probables) + possibles
```

i.e. a match with a "probable" pattern has a weight of 2 while a match with a "possible" pattern has a weight of 1. In the course of establishing this weighting system, we experimented with several more complex and finely-grained schemes, but found that they provided no significant advantage in practice.

After the hypothesization phase comes confirmation. This involves more detailed topic-specific processing to determine whether or not the vocabulary used in hypothesizing the topic was used in a misleading way. The confirmation phase uses topic-specific knowledge-based rules which may try to match additional patterns or patternsets.

The most complex confirmation rules in the system are those for the **war** and **disorders** topics. These topics were difficult to tell apart, so considerable additional processing was involved. The rules use additional specialized patternsets: one patternset looked specifically for words (including proper names) that occur in war but not disorders stories and another looked for vocabulary that occurs in stories that are both war and disorders stories. There are also patternsets for sports, crime, and disaster vocabulary. The confirmation rules associated with **war** and **disorders** attempt to match these rules according to a branching set of conditions.

Consider the following story, for example. The

words and phrases in boldface match patterns in the "conflict" patternset; the total value of matches is great enough to get the story hypothesized as **war** and **disorders**. In the confirmation phase, additional patternsets are run against the story. As soon as *Iran* and *Iraq* are matched, the topic **war** is confirmed and the topic **disorders** is disconfirmed.

IRAN ANNOUNCES END OF MAJOR **OFFENSIVE** IN GULF **WAR**

LONDON, Feb 26 - Iran announced tonight that its major **offensive against** Iraq in the Gulf **war** had ended after dealing savage blows against the Baghdad government.

[...]

The statement by the Iranian High Command appeared to herald the close of an **assault** on the port city of Basra in southern Iraq.

[...]

It said 81 Iraqi **brigades** and **battalions** were totally destroyed, along with 700 **tanks** and 1,500 other vehicles. The victory list also included 80 **warplanes downed**, 250 **anti-aircraft guns** and 400 pieces of military hardware destroyed and the seizure of 220 **tanks** and **armored personnel carriers.**

For the story that follows, the topics **war** and **disorders** are also originally hypothesized. In the confirmation phase, two things are discovered: the story mentions no wars by name nor contains any references to countries or organizations involved in conflicts that are classified as wars; and there is nothing in the story that suggests that the topic **disorders** should be disconfirmed. Hence **war** is disconfirmed and **disorders** is confirmed.

RIOT REPORTED IN SOUTH KOREAN PRISON

Seoul, July 5 - Twelve South Korean women **detainees** refused food for the fifth consecutive day today after a **riot** against their maltreatment in a Seoul prison was put down, dissident sources said.

The 12, **detained** for **anti-government protests** and awaiting trial, pushed away prison officials, smashed windows and occupied a prison building on Tuesday as a **protest** against what they called "suppression of prisoners' human rights".

After two hours, about 40 **riot** police, **firing tear gas**, stormed the building and overpowered the **protesters**, the sources said. Some **protesters** were injured, they added.

For the story below, both **war** and **disorders** are hypothesized and then disconfirmed because *tennis* is matched during the disconfirmation phase.

LENDL **DEMONSTRATES** GRASS COURT MATURITY

LONDON, July 2 - Czechoslovak top seed Ivan Lendl served warning that he may finally have come of age on grass when he emerged victorious from a pitched **battle** with one of the finest exponents of the fast court game at Wimbledon today.

The U.S. and French Open tennis champion has never won a title on grass but he outlasted American 10th seed Tim Mayotte 6-4 4-6 6-4 3-6 9-7 over three and a half hours to join Boris Becker, Henri Leconte and Slobodan Zivojinovic in Friday's semifinals.

The titanic **struggle** on court one upstaged the centre court **clash** between seventh seed Leconte and the remarkable Australian Pat Cash, which had been billed as the day's main attraction [...]

The story below is the rare sports story which is also a disorders story. Even though the name of a sporting event, *Asian Games*, occurs in the text, the topic **disorders** is not disconfirmed. The reason is that the confirmation patternsets match words and phrases in the story (e.g. *radicals* and *violent protests*) that very strongly suggest that real disorders are being described.

POLICE SEEK 160 SOUTH KOREAN **RADICALS**

SEOUL, July 2 - Police said today they wanted to **detain** 160 South Koreans to stop **sabotage** attempts during September's Asian Games in Seoul.

The 160, mostly students and workers, masterminded various **violent protests** against the government and the United States in the past months but managed to escape arrest, police said.

They had been tipped that the **radicals** were trying to organise big **demonstrations** against the government during the Asiad, which is to run from September 20 to October 5.

"It is highly probable that they will form **radical** underground groups to step up their **anti-government** and anti-U.S. **protests** and may disrupt the Asian Games in an attempt to defame the government," a senior police officer told reporters.

[...]

3.4. Flow of Control

Rather than being expressed in a formal rule language, topic hypothesization and confirmation rules are specified through a lisp program. Having a

program allows for fine-grained control by the rule developer. Rather than having a set of hypothesization and confirmation rules which are processed in a fixed order, we allow the rule developer to specify the order and manner of processing in a topic-dependent manner. The major kinds of activities available to rule developers for incorporation into the control code are the following: running one or more patternsets, applying evaluation functions to the resulting matches, and confirming or disconfirming topics.

In developing the system, we observed many regularities in the lisp code which controls the flow of processing and we believe it would be possible and profitable to provide rule developers with a more restricted control language which embodies many of these regularities in its primitives.

3.5. Rulebase Development

The process of formulating the *rulebase* of the system, i.e. the collection of patterns, patternsets, and hypothesization and confirmation rules it uses, is an empirical one. It requires human rule developers to examine many stories, create rulebase components according to their intuitions, run stories through the system, observe the results, and modify the system to avoid any miscategorizations that have occurred without introducing new miscategorizations. This task is time-consuming and sometimes tedious. Nevertheless, our experience with the system suggests that it does tend to converge without undue oscillation at an accuracy level that while far from perfect is adequate for many tasks of practical importance (see Section 4). The rule development effort on this system took approximately six person months.

An important factor in the success of the rulebase development effort was the separation of the vocabulary the system looks for into a collection of abstract concepts represented by patternsets. The patternsets provide rule developers with a way of thinking about the themes they are looking for in a story when they write the hypothesization and confirmation rules without becoming mired in questions about which specific words and phrases indicate those themes.

In designing the system, we also considered a different approach in which the selection of words and phrases to look for would be determined automatically by a statistical method. Since we did not adopt this approach, we have no direct evidence that it would not have worked as well as the labor-intensive method chosen. However, our choice was influenced by a belief that a statistical method would not provide us with a choice of words and phrases that could be used to make distinctions as precisely as the patterns of the kind described above that were chosen by humans.

As shown in [2], accuracy is particularly problematic with a traditional keyword approach regardless of whether the keywords are selected by humans or statistically. And if we had adopted a statistical approach, it would have been computationally expensive to vary the length of the phrases chosen as much as human rule developers do. It would also have been difficult to establish the contextual restrictions that human rule developers establish (e.g. this word, so long as it is not followed by one of these four others). Rules of the complexity of the confirmation rule for **war** and **disorders** described in Section 3.3 are of course essentially impossible to establish by statistical means.

Some interesting possibilities for a statistical approach to defining keywords have appeared recently in conjunction with semantic information about potential keywords [7] and in conjunction with very powerful parallel hardware devices [4]. However, given the current state of the art, we continue to believe that our decision to use rules formulated and refined by human developers was a sound one from the point of view of the accuracy of the resulting system.

4. Performance

4.1. Measuring Performance

The accuracy of the system for topic assignments was measured through two percentages for each of the six topics:

- **recall**: the percentage of stories assigned the topic code by human categorizers that were also assigned that code by the system;
- **precision**: the percentage of stories assigned the topic code by the system that actually carried the topic code assigned by the human categorizers.

The recall rate serves as a measure of the number of stories for which the system misses an appropriate topic code; a high recall percentage will therefore mean few such false negatives. The precision rate, on the other hand, measures the number of stories for which the system chooses an incorrect topic. A high precision percentage means few such false positives. We emphasized high recall over high precision.

4.2. Results

The results obtained from the system were very promising. After certain necessary adjustments (described below) to the raw results, the system had an average recall rate of 93% (i.e. it made 93% of the topic assignments it should have made and missed only 7%) and an average precision rate also of 93% (i.e. 93% of the topics it did assign were correct). Another way of expressing this is that it had on average only 7% false negatives and 7% false positives in its topic assignments. This level of accuracy was achieved in an average of around 15 seconds per story on a Symbolics 3640 in Common Lisp. Little effort was spent to optimize the execution time and we believe that a substantial improvement in speed is possible.

Adjustments to the raw recall and precision figures produced by the system were necessary because, as described in Section 2, we discovered three problematic features of the hand-categorizations against which the system was being evaluated: they were not always content-based; they were not always consistent; and some topic definitions were vague. Given this, it was clear that raw performance scores would not give a meaningful picture of how well the system worked, so we devised a score-adjustment procedure to provide results that would reflect system performance more accurately. The remainder of this section describes that procedure and presents the raw and adjusted results we obtained.

We used an adjustment procedure that was based on the assumption that there are three explanations for disagreements between the system and the human categorizers about the assignment of a topic to a story:

- The human categorizer is clearly wrong.
- The system is clearly wrong.
- The topic assignment is debatable. This case can typically be attributed to one of the three sources of difficulty described above.

A set of 500 stories was run through the system. These stories had never before been processed, and no hypothesization or confirmation rules had ever been based on them. A Carnegie Group employee who was not involved with the system produced score adjustments for each topic disagreement between the system and the human categorizers. The employee was presented with a story and told that there was a disgreement on a specific named topic; she was not told which choice the system or the human categorizers had made. The employee was asked to decide whether the topic was appropriate for the story, inappropriate, or debatable. Debatable cases counted in favor of the system.

The results of this experiment before and after adjustment of the system's scores were as follows (where **acq** is acquisitions/mergers, **mtl** is metals, **shp** is shipping, **bnd** is bonds, and **dis** is disorders).

	Raw Rec.	Raw Prec.	Adj. Rec.	Adj. Prec.
acq	85%	82%	92%	92%
bnd	91%	89%	97%	100%
dis	90%	58%	93%	84%
mtl	80%	70%	95%	90%
shp	72%	49%	88%	92%
war	88%	82%	92%	100%

Recall is 92% or higher, except in the case of the **shipping** code. This is not surprising because it turned out that **shipping** was a strongly interest-based category, as far as the human categorizers were concerned. So, stories about rough weather in the St. Lawrence seaway (but not the Rhine) and the devaluation of the rupee (but not the Turkish lira) were classified as shipping stories because human categorizers possessed the expert knowledge that shippers are interested in that particular waterway and that particular currency.

The precision scores are actually higher than the corresponding recall scores in the case of **war** and **bonds**. Since we have found that precision can be traded off against recall by appropriate manipulation of thresholds associated with our rulebase, this suggests that the recall rate for those two topics could be further improved while still maintaining an acceptable precision rate.

The adjustment procedure also allowed us to

measure the performance of the hand-categorizers. While adjusted precision scores were perfect for all six topics, adjusted recall scores ranged from 81% to 100%, with an average of 94%.

	Adj. Rec.	Adj. Prec.
acq	100%	100%
bnd	97%	100%
dis	81%	100%
mtl	95%	100%
shp	100%	100%
war	90%	100%

While human performance on precision is clearly superior to that of the system, the average recall rates of human categorizers and of the system are very similar (94% v. 93%). Closer examination of the results, however, shows that the kind of errors made are quite different. Human errors stem mainly from inconsistent application of categories, especially the categories with the vaguest definitions, and from failing to specify all the categories when several should have been assigned to a story. System errors on the other hand stem largely from misinterpretation of the way in which language is being used. This sometimes results in ridiculous categorizations of a kind that humans never produce.

Out of 500 stories, the system produced a total of 28 "lemons" (stories that were clearly assigned the wrong categories). We analyzed these stories and discovered six sources of errors:

- The system did not match useful words or phrases, or the disconfirmation rules were too powerful.
- The topic vocabulary was not much used in the story.
- The system used the story background to derive the topic.
- The topic vocabulary came too late in the story.
- The topic vocabulary was used with different meanings.
- The topic vocabulary was used with the same meaning, but different focus.

Examples and further discussion follow.

4.2.1. Topic Vocabulary Not Much Used In Story

Some stories did not use the topic vocabulary more than one or two times. Setting thresholds very low would catch these stories, but generate many false positives as well. Most stories that had this problem were also very short, so we added length-dependent thresholds to address the problem. This technique worked for **metals** stories, where the vocabulary is somewhat distinctive, but would not work for **acquisitions** stories, where the vocabulary consists of very common words like *buy* and *sell*.

4.2.2. Story Background Used To Derive Topic

News stories sometimes have background information included which does not have much to do with the main point of the story. For example, the following story, about the Pope's visit to Colombia, was miscategorized as a **metals** story because of the background information about the country. Solving this problem requires a deep understanding of the structure of the story.

> SECURITY FOR POPE TIGHTENS
>
> Chiquinquira, Colombia, July 3 - Security precautions for Pope John Paul II were tightened today, with hundreds of troops making thorough body searches of visitors to this colonial town high in the Andes mountains.
>
> [...]
>
> Chiquinquira has been spared the guerilla warfare which has torn much of Colombia over the past three decades. But the nearby Muzo emerald **mines**, the country's biggest, have attracted adventurers who often feud violently in the town.
>
> Some Muzo **miners** have moved on to the more lucrative drug traffic [...].

4.2.3. Topic Vocabulary Used With Different Meaning

Sometimes stories are miscategorized because of the metaphorical language they use. For example, in one story the word *revolution* appeared numerous times: the British government was calling for a revolution in broadcasting. Another contained the phrases *ready to go to war*, *make peace*, *make war*, *target*, and *heavy losses*; the subject of the story was labor negotiations in the automobile industry. Since the system does not really understand the texts it processes, it is inevitable that it will be fooled from time to time by such usage.

4.2.4. Topic Vocabulary Used with Same Meaning But Different Focus

The following story illustrates another problem for which there is no obvious solution. The word *army* occurs four times (not all shown), and the sense of the word in this **military** story is exactly the sense it might have in an actual **war** or **disorders** story.

> CHINESE **ARMY** TO HAVE NCOS FOR FIRST TIME
>
> Peking, July 4 - The Chinese **army** will allow non-commissioned officer ranks for the first time as part of its reform program, the New China News Agency said today.
>
> It said soldiers who have been in the **army** for one year and had a good record would, after training at two special schools, serve as NCOS.
>
> [...]

5. Conclusion

This paper has shown that high accuracy automatic text categorization is feasible for texts covering a diverse set of topics, using the well-established natural language processing technique of pattern matching applied according to knowledge-based rules, but without requiring a complete syntactic or semantic analysis of the texts. Automatic text processing of this kind has many potential applications with high economic paybacks in the routing and archiving of news stories, electronic messages, or other forms of on-line text. We expect that many such systems will be in commercial use within the next few years.

Acknowledgements

Peter Neuss and Scott Safier contributed substantially to the design and implementation of the system described in this paper.

References

1. Carbonell, J. G., Boggs, W. M., Mauldin, M. L., and Anick, P. G. The XCALIBUR Project: A Natural Language Interface to Expert Systems. Proc. Eighth Int. Jt. Conf. on Artificial Intelligence, Karlsruhe, August, 1983.

2. Furnas, G. W., Landauer, T. K., Dumais, S. T., and Gomez, L. M. "Statistical semantics: Analysis of the potential performance of keyword information systems". *Bell System Technical Journal 62*, 6 (1983), 1753-1806.

3. Hayes, P. J., Andersen, P., Safier, S. Semantic Case Frame Parsing and Syntactic Generality. Proc. of 23rd Annual Meeting of the Assoc. for Comput. Ling., Chicago, June, 1985.

4. Hillis, W. D.. *The Connection Machine*. MIT Press, Cambridge, Mass., 1985.

5. Parkison, R. C., Colby, K. M., and Faught, W. S. "Conversational Language Comprehension Using Integrated Pattern-Matching and Parsing". *Artificial Intelligence 9* (1977), 111-134.

6. Salton, G. and McGill, M. J.. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

7. Walker, D. E. and Amsler, R. A. The Use of Machine-Readable Dictionaries in Sublanguage Analysis. In *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, R. Grishman and R. Kittredge, Ed., Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1986, pp. 69-83.