Little Words Can Make a Big Difference for Text Classification

Ellen Riloff
Department of Computer Science
University of Utah
Salt Lake City, UT 84112
E-mail: riloff@cs.utah.edu

Abstract

Most information retrieval systems use stopword lists and stemming algorithms. However, we have found that recognizing singular and plural nouns, verb forms, negation, and prepositions can produce dramatically different text classification results. We present results from text classification experiments that compare *relevancy signatures*, which use local linguistic context, with corresponding indexing terms that do not. In two different domains, relevancy signatures produced better results than the simple indexing terms. These experiments suggest that stopword lists and stemming algorithms may remove or conflate many words that could be used to create more effective indexing terms.

Introduction

Most information retrieval systems use a stopword list to prevent common words from being used as indexing terms. Highly frequent words, such as determiners and prepositions, are not considered to be content words because they appear in virtually every document. Stopword lists are almost universally accepted as a necessary part of an information retrieval system. For example, consider the following quote from a recent information retrieval textbook:

> "It has been recognized since the earliest days of information retrieval (Luhn 1957) that many of the most frequently occurring words in English (like "the", "of", "and", "to", etc.) are worthless indexing terms." ([Frakes and Baeza-Yates, 1992], p. 113)

Many information retrieval systems also use a stemming algorithm to conflate morphologically related words into a single indexing term. The motivation behind stemming algorithms is to improve recall by generalizing over morphological variants. Stemming algorithms are commonly used, although experiments to determine their effectiveness have produced mixed results (e.g., see [Harman, 1991; Krovetz, 1993]).

One benefit of stopword lists and stemming algorithms is that they significantly reduce the storage requirements of inverted files. But at what price? We have found that some types of words, which would be removed by stopword lists or merged by stemming algorithms, play an important role in making certain domain discriminations. For example, similar expressions containing different prepositions and auxiliary verbs behave very differently. We have also found that singular and plural nouns produce dramatically different text classification results.

First, we will describe a text classification algorithm that uses linguistic expressions called "relevancy signatures" to classify texts. Next, we will present results from text classification experiments in two domains which show that similar signatures produce substantially different classification results. Finally, we discuss the implications of these results for information retrieval systems.

Relevancy Signatures

Relevancy signatures represent linguistic expressions that can be used to classify texts for a specific domain (i.e., topic). The linguistic expressions are extracted from texts automatically using an information extraction system called CIRCUS. The next section gives a brief introduction to information extraction and the CIRCUS sentence analyzer, and the following section describes relevancy signatures and how they are used to classify texts.

Information Extraction

CIRCUS [Lehnert, 1991] is a conceptual sentence analyzer that extracts domain-specific information from text. For example, in the domain of terrorism, CIRCUS can extract the names of perpetrators, victims, targets, weapons, dates, and locations associated with terrorist incidents. Information is

extracted using a dictionary of domain-specific structures called *concept nodes*. Each concept node recognizes a specific linguistic pattern and uses the pattern as a template for extracting information.

For example, a concept node dictionary for the terrorism domain contains a concept node called \$murder-passive-victim\$ which is triggered by the pattern "<X> was murdered" and extracts X as a murder victim. A similar concept node called \$murder-active-perpetrator\$ is triggered by the pattern "<X> murdered ..." and extracts X as the perpetrator of a murder. A concept node is activated during sentence processing when it recognizes its pattern in a text.

Figure 1 shows a sample sentence and instantiated concept nodes produced by CIRCUS. Two concept nodes are generated in response to the passive form of the verb "murdered". One concept node, \$murder-passive-victim\$, extracts the "three peasants" as murder victims, and a second concept node, \$murder-passive-perpetrator\$, extracts the "guerrillas" as perpetrators.[1]

Sentence:
Three peasants were murdered by guerrillas.

\$murder-passive-victim\$
    victim = "three peasants"

\$murder-passive-perpetrator\$
    perpetrator = "guerrillas"

Figure 1: Two instantiated concept nodes

Theoretically, concept nodes can be arbitrarily complex but, in practice, most of them recognize simple linguistic constructs. Most concept nodes represent one of the general linguistic patterns shown in Figure 2.[2]

All of the information extraction done by CIRCUS happens through concept nodes, so it is crucial to have a good concept node dictionary for a domain. Multiple concept nodes may be generated for a sentence, or no concept nodes may be generated at all. Sentences that do not activate any concept nodes are effectively ignored.

Building a concept node dictionary by hand can be extremely time-consuming and tedious. We estimate that it took approximately 1500 person-hours for two experienced system developers to build a concept node dictionary by

---

[1] In principle, a single concept node can extract more than one item. However, concept nodes produced by AutoSlog [Riloff, 1994; Riloff, 1993] extract only one item at a time. The joint venture results presented in this paper are based on a concept node dictionary produced by AutoSlog.

[2] These are the linguistic patterns used by AutoSlog to create the joint ventures dictionary (see [Riloff, 1994; Riloff, 1993] for details). The concept node dictionary for the terrorism domain was hand-crafted and contains some more complicated patterns as well.

Linguistic Pattern	Example
<**subject**> **passive-verb**	<entity> was <u>formed</u>
<**subject**> **active-verb**	<entity> <u>linked</u>
<**subject**> **verb dobj**	<entity> completed <u>acquisition</u>
<**subject**> **verb infinitive**	<entity> agreed to <u>form</u>
<**subject**> **auxiliary noun**	<entity> is <u>conglomerate</u>
active-verb <**dobj**>	<u>acquire</u> <entity>
infinitive <**dobj**>	to <u>acquire</u> <entity>
verb infinitive <**dobj**>	agreed to <u>establish</u> <entity>
gerund <**dobj**>	<u>producing</u> <product>
noun auxiliary <**dobj**>	<u>partner</u> is <entity>
noun prep <**np**>	<u>partnership</u> between <entity>
active-verb prep <**np**>	<u>buy</u> into <entity>
passive-verb prep <**np**>	was <u>signed</u> between <entity>
infinitive prep <**np**>	to <u>collaborate</u> on <product>

Figure 2: Concept node patterns and examples from the joint ventures domain

hand for the terrorism domain. However, we have since developed a system called AutoSlog [Riloff, 1994; Riloff, 1993] that creates concept node dictionaries automatically using an annotated training corpus. Given a training corpus for the terrorism domain, a dictionary created by AutoSlog achieved 98% of the performance of the hand-crafted dictionary and required only 5 person-hours to build.

Relevancy Signatures

Motivation Most information retrieval systems classify texts on the basis of multiple words and phrases. However, for some classification tasks, classifying texts on the basis of a single linguistic expression can be effective. Although single words do not usually provide enough context to be reliable indicators for a domain, slightly larger phrases can be reliable. For example, the word "dead" is not a reliable keyword for murder because people die in many ways that have nothing to do with murder. However, some expressions containing the word "dead" *are* reliable indicators of murder. Figure 3 shows several expressions involving the words "dead" and "fire", and the percentage of occurrences of each expression that appeared in relevant texts. These results are based on 1500 texts from the MUC-4 corpus.[3] The texts in the MUC-4 corpus were retrieved from a general database because they contain one or more words related to terrorism but only half of them actually describe a relevant terrorist incident.[4]

Figure 3 shows that every occurrence of the expression "was found dead" appeared in a relevant text. However, only

---

[3] MUC-4 was the Fourth Message Understanding Conference held in 1992 [MUC-4 Proceedings, 1992].

[4] The MUC-4 organizers defined terrorism according to a complicated set of guidelines but, in general, a relevant event was a specific incident that occurred in Latin America involving a terrorist perpetrator and civilian target.

Expression	Rel. %
was found dead	100%
left dead	61%
<number> dead	47%
set on fire	100%
opened fire	87%
<weapon> fire	59%

Figure 3: Strength of Associations for Related Expressions

61% of the occurrences of the expression "left dead" and 47% of the occurrences of "<number> dead" (e.g., "there were 61 dead") appeared in relevant texts. This is because the expression "was found dead" has an implicit connotation of foul play, which suggests that murder is suspected. In contrast, the expressions "left dead" and "<number> dead" often refer to military casualties that are not terrorist in nature.

Figure 3 also shows that several expressions involving the word "fire" have different correlations with relevance. The expression "set on fire" was strongly correlated with relevant texts describing arson incidents, and the expression "opened fire" was highly correlated with relevant texts describing terrorist shooting incidents. However, the expression "<weapon> fire" (e.g., "rifle fire" or "gun fire") was not highly correlated with terrorist texts because it often appears in texts describing military incidents.

These results show that similar linguistic expressions can have very different associations with relevance for a domain. Furthermore, many of these distinctions would be difficult, if not impossible, for a human to anticipate. Based on these observations, we developed a text classification algorithm that automatically identifies linguistic expressions that are strongly associated with a domain and uses them to classify new texts. Our approach uses an underlying information extraction system, CIRCUS, to recognize linguistic context.

The Relevancy Signatures Algorithm   A *signature* is defined as a pair consisting of a word and a concept node triggered by that word. Each signature represents a unique set of linguistic expressions. For example, the signature <murdered, $murder-passive-victim$> represents all expressions of the form "was murdered", "were murdered", "have been murdered", etc. Signatures are generated automatically by applying CIRCUS to a text corpus.

A relevancy signature is a signature that is highly correlated with relevant texts in a preclassified training corpus. To generate relevancy signatures for a domain, the training corpus is processed by CIRCUS, which produces a set of instantiated concept nodes for each text. Each concept node is then transformed into a signature by pairing the name of the concept node with the word that triggered it. Once a set of signatures has been acquired from the corpus, for each signature we estimate the conditional probability that a text is relevant given that it contains the signature. The formula is:

$$\Pr(\frac{text\ is\ relevant}{text\ contains\ sig_i}) = \frac{N_{sig_i \epsilon REL-TEXTS}}{N_{sig_i}}$$

where N_{sig_i} is the number of occurrences of the signature sig_i in the training corpus and $N_{sig_i \epsilon REL-TEXTS}$ is the number of occurrences of the signature sig_i in relevant texts in the training corpus. The epsilon is used loosely to denote the number of occurrences of the signature that "appeared in" relevant texts.

Finally, two thresholds are used to identify the signatures that are most highly correlated with relevant texts. A relevance threshold R selects signatures with conditional probability $\geq$ R, and a frequency threshold M selects signatures that have appeared at least M times in the training corpus. For example, R = .85 specifies that at least 85% of the occurrences of a signature in the training corpus appeared in relevant texts, and M = 3 specifies that the signature must have appeared at least 3 times in the training corpus.

To classify a new text, the text is analyzed by CIRCUS and the resulting concept nodes are transformed into signatures. Then the signatures are compared with the list of relevancy signatures for the domain. If any of the relevancy signatures are found, then the text contains an expression that is strongly associated with the domain so it is classified as relevant. If no relevancy signatures are found, then the text is classified as irrelevant. The presence of a single relevancy signature is enough to produce a relevant classification.

Experimental Results for Similar Expressions

Previous experiments demonstrated that the relevancy signatures algorithm can achieve high-precision text classification and performed better than an analogous word-based algorithm in two domains: terrorism and joint ventures (see [Riloff, 1994; Riloff and Lehnert, 1994] for details). In this paper, we focus on the effectiveness of similar linguistic expressions for classification. In many cases, similar signatures generated substantially different conditional probabilities. In particular, we found that recognizing singular and plural nouns, different verb forms, negation, and prepositions was critically important in both the terrorism and joint ventures domains. The results are based on 1500 texts from the MUC-4 terrorism corpus and 1080 texts from a joint ventures corpus.[5] In both corpora, roughly 50% of the texts were relevant to the targeted domain. Although most general-purpose corpora contain a much smaller percentage of relevant texts, our goal is to simulate a pipelined system in which a traditional information retrieval system is first applied to a general-purpose corpus to identify potentially relevant texts. This prefiltered corpus is then used by our system

---

[5]These texts were randomly selected from a corpus of 1200 texts, of which 719 came from the MUC-5 joint ventures corpus [MUC-5 Proceedings, 1993] and 481 came from the Tipster detection corpus [Tipster Proceedings, 1993; Harman, 1992] (see [Riloff, 1994] for details of how these texts were chosen).

to make more fine-grained domain discriminations.[6]

Singular and Plural Nouns

Figures 4 and 5 show signatures that represent singular and plural forms of the same noun, and their conditional probabilities in the terrorism and joint ventures corpora, respectively. Singular and plural words produced dramatically different correlations with relevant texts in both domains. For example, Figure 4 shows that 83.9% of the occurrences of the singular noun "assassination" appeared in relevant texts, but only 51.3% of the occurrences of the plural form "assassinations" appeared in relevant texts. Similarly, in the joint ventures domain, 100% of the occurrences of "venture between" appeared in relevant texts, but only 75% of the occurrences of "ventures between" appeared in relevant texts. And these were not isolated cases; Figures 4 and 5 show many more examples of this phenomenon.

Signature	Rel. %
<assassination, $murder$>	83.9%
<assassinations, $murder$>	51.3%
<car_bomb[7], $weapon-vehicle-bomb$>	100.0%
<car_bombs, $weapon-vehicle-bomb$>	75.0%
<corpse, $dead-body$>	100.0%
<corpses, $dead-body$>	50.0%
<disappearance, $disappearance$>	83.3%
<disappearances, $disappearance$>	22.2%
<grenade, $weapon-grenade$>	81.3%
<grenades, $weapon-grenade$>	34.1%
<murder, $murder$>	83.8%
<murders, $murder$>	56.7%

Figure 4: Singular/plural terrorism signatures

Signature	Rel. %
<tie-up, $entity-tie-up-with$>	100.0%
<tie-ups, $entity-tie-ups-with$>[8]	0.0%
<venture, $entity-venture-between$>	100.0%
<ventures, $entity-ventures-between$>	75.0%
<venture, $entity-venture-of$>	95.4%
<ventures, $entity-ventures-of$>	50.0%
<venture, $entity-venture-with$>	96.0%
<ventures, $entity-ventures-with$>	52.4%

Figure 5: Singular/plural joint ventures signatures

The reason revolves around the fact that singular nouns usually referred to a *specific* incident, while the plural nouns

often referred to *general* types of incidents. For example, the word "assassination" usually referred to the assassination of a specific person or group of people, such as "the assassination of John Kennedy" or "the assassination of three diplomats." In contrast, the word "assassinations" often referred to assassinations in general, such as "there were many assassinations in 1980", or "assassinations often have political ramifications." In both domains, a text was considered to be relevant only if it referred to a specific incident of the appropriate type.[9]

Verb Forms

We also observed that different verb forms (active, passive, infinitive) behaved very differently. Figures 6 and 7 show the statistics for various verb forms in both domains. In general, passive verbs were more highly correlated with relevance than active verbs in the terrorism domain. For example, 77.8% of the occurrences of "was bombed by <X>" appeared in relevant texts but only 54.1% of the occurrences of "<X> bombed ..." appeared in relevant texts. In the MUC-4 corpus, passive verbs were most frequently used to describe terrorist events, while active verbs were equally likely to describe military events. Two possible reasons are that (1) the perpetrator is often not known in terrorist events, which makes the passive form more appropriate, and (2) the passive form connotes a sense of victimization, which news reporters might have been trying to convey.

Signature	Rel. %
<blamed, $suspected-or-accused-active$>	84.6%
<blamed, $suspected-or-accused-passive$>	33.3%
<bombed, $actor-passive-bombed-by$>	77.8%
<bombed, $actor-active-bomb$>	54.1%
<broke, $damage-active$>	80.0%
<broken, $damage-passive$>	62.5%
<burned, $arson-passive$>	100.0%
<burned, $arson-active$>	76.9%
<charged, $perpetrator-passive$>	68.4%
<charged, $perpetrator-active$>	37.5%
<left, $location-passive$>	87.5%
<left, $location-active$>	20.0%

Figure 6: Terrorism signatures with different verb forms

However, the active verb form was more highly correlated with relevant texts for the words "blamed" and "broke". Terrorists were often actively "blamed" for an incident, while all kinds of people "were blamed" or "have been blamed" for other types of things. The active form "broke" was often used to describe damage to physical targets while the passive form was often used in irrelevant phrases such as "talks were broken off", or "a group was broken up".

---

[6] In fact, the MUC-4 and MUC-5 corpora were constructed by applying a keyword search to large databases of news articles.

[7] CIRCUS uses a phrasal lexicon to represent important phrases as single words. The underscore indicates that the phrase "car bomb" was treated as a single lexical item.

[8] This signature only appeared once in the corpus.

[9] The relevance criteria are based on the MUC-4 and Tipster guidelines [MUC-4 Proceedings, 1992; Tipster Proceedings, 1993].

Signature	Rel. %
<assemble, $entity-active-assemble$>	87.5%
<assemble, $prod-infinitive-to-assemble$>	68.8%
<construct, $entity-active-construct$>	100.0%
<constructed, $facility-passive-constructed$>	63.6%
<form, $entity-infinitive-to-form$>	83.1%
<form, $entity-obj-active-form$>	69.2%
<put, $entity-passive-put-up-by$>	84.2%
<put, $entity-active-put-up$>	50.0%
<manufacture, $prod-infinitive-to-manufacture$>	86.7%
<manufacture, $prod-active-manufacture$>	53.8%
<manufactured, $prod-passive-manufactured$>	52.6%
<operate, $facility-active-operate$>	85.0%
<operated, $facility-passive-operated$>	66.7%
<supplied, $entity-active-supplied$>	83.3%
<supplied, $entity-passive-supplied-by$>	65.0%

Figure 7: Joint venture signatures with different verb forms

In the joint ventures domain, Figure 7 also shows significant differences in the relevancy rates of different verb forms. In most cases, active verbs were more relevant than passive verbs because active verbs often appeared in the future tense. This makes sense when describing joint venture activities because, by definition, companies are planning events in the future. For example, many texts reported that a joint venture company "will assemble" a new product, or "will construct" a new facility. In contrast, passive verbs usually represent the past tense and don't necessarily mention the actor (e.g., the company). For example, the phrase "a facility was constructed" implies that the construction has already happened and does not indicate who was responsible for the construction. Infinitive verbs were also common in this domain because companies often intend to do things as part of joint venture agreements.

Prepositions

In the next set of experiments, we investigated the role of prepositions as part of the text representation. First, we probed the joint ventures corpus[10] with joint venture keywords and computed the recall and precision rates for these words, which appear in Figure 8. For example, we retrieved all texts containing the word "consortium" and found that 69.7% of them were relevant and 3.6% of the relevant texts were retrieved. Some of the keywords achieved high recall and precision rates. For example, 88.9% of the texts containing the words "joint" and "venture"[11] were relevant. But only 73.2% of the texts containing the hyphenated word "joint-venture" were relevant. This is because the hyphenated form "joint-venture" is often used as a modifier, as in "joint-venture law" or "joint-venture proposals", where the main concept is not a specific joint venture. Figure 8 also shows much higher precision for the singular forms "ven-

ture" and "joint venture" than for the plural forms, which is consistent with our previous results for singular and plural nouns.

Words	Recall	Precision
joint, venture	93.3%	88.9%
tie-up	2.5%	84.2%
venture	95.5%	82.8%
jointly	11.0%	78.9%
joint-venture	6.4%	73.2%
consortium	3.6%	69.7%
joint, ventures	19.3%	66.7%
partnership	7.0%	64.3%
ventures	19.8%	58.8%

Figure 8: Recall and precision scores for joint venture words

But perhaps the most surprising result was that most of the keywords did not do very well. The phrase "joint venture" achieved both high recall and precision, but even this obviously important phrase produced < 90% precision. And virtually all of the other keywords achieved modest precision; only "tie-up" and "venture" achieved greater than 80% precision.

When we add prepositions to these keywords, we produce more effective indexing terms. Figure 9 shows several signatures for the joint ventures domain that represent verbs and nouns paired with different prepositions. For example, Figure 9 shows that pairing the noun "venture" with the preposition "between" produces a signature that achieves 100% precision. Similarly, pairing the word "venture" with the prepositions "with" and "by" produces signatures that achieve over 95% precision. And pairing the word "tie-up" with the preposition "with" increases precision from 84.2% to 100%. Figure 9 also shows substantially different precision rates for the same word paired with different prepositions. For example, "project between" performs much better than "project with", and "set up with" performs much better than "set up by".

Signature	Rel. %
<project, $entity-project-between$>	100.0%
<project, $entity-project-with$>	75.0%
<set, $entity-set-up-with$>	94.7%
<set, $entity-set-up-by$>	66.7%
<tie, $entity-tie-up-with$>	100.0%
<venture, $entity-venture-between$>	100.0%
<venture, $entity-venture-with$>	95.9%
<venture, $entity-venture-of$>	95.4%
<venture, $entity-venture-by$>	90.9%

Figure 9: Joint venture signatures with different prepositions

It is important to note that the signatures are generated by CIRCUS, which is a natural language processing sys-

---

[10] These results are from the full joint ventures corpus of 1200 texts.

[11] Not necessarily in adjacent positions.

tem, so the preposition does not have to be adjacent to the noun. A prepositional phrase with the appropriate preposition only has to follow the noun and be extracted by a concept node triggered by the noun. For example, the sentence "Toyota formed a joint venture in Japan on March 12 with Nissan", would produce the signature <venture, $entity-venture-with$>, even though the preposition "with" is six words away from the word "venture".

These results show that prepositions represent distinctions that are important in the joint ventures domain. The words "venture" and "joint venture" by themselves do not necessarily represent a *specific* joint venture activity. For example, they are often used as modifiers, as in "venture capitalists" or "joint venture legislation", or are used in a general sense, as in "one strategy is to form a joint venture". The presence of the preposition "with", however, almost always implies that a specific partner is mentioned. Similarly, the preposition "between" usually indicates that more than one partner is mentioned. In some sense, the prepositions act as pointers which indicate that one or more partners exist.

As a result, keywords paired with these prepositions (e.g., "tie-up with", "venture with", "venture between", "project between") almost always refer to a specific joint venture agreement and are much more effective indexing terms than the keywords alone. Furthermore, some prepositions are better indicators of the domain than others. For example, the preposition "between" suggests that multiple partners are mentioned and is therefore more highly associated with the domain than other prepositions.[12]

Negation

We also found that negated phrases performed differently than similar phrases without negation. Figure 10 shows several examples of this phenomenon. In the terrorism domain, negative expressions were more relevant than their positive counterparts.[13] For example, the expression "no casualties" was more relevant than the word "casualties", "was not injured" was more relevant than "was injured", and "no injuries" was more relevant than "injuries". The reason for this is subtle. Reports of injuries and casualties are common in military event descriptions as well as terrorist event descriptions. However, negative reports of "no injuries" or "no casualties" are much more common in terrorist event descriptions. In most cases, these expressions implicitly suggest that there were no *civilian* casualties as a result of a terrorist attack. This is another example of a phenomenon that

---

[12] We did not conduct the preposition experiments in the terrorism domain because the concept nodes in the hand-crafted terrorism dictionary included multiple prepositions so it was difficult to compute statistics for prepositions individually. However we expect that the same phenomenon occurs in the terrorism domain as well.

[13] We did not have any concept nodes representing negation in the joint ventures domain.

would be difficult if not impossible for a human to predict. One of the main strengths of the relevancy signatures algorithm is that these associations are identified automatically using statistics generated from a training corpus.

Signature	Rel. %
<casualties, $no-injury$>	84.7%
<casualties, $injury$>	46.1%
<injured, $no-injury-passive$>	100.0%
<injured, $injury-passive$>	76.7%
<injuries, $no-injury$>	83.3%
<injuries, $injury$>	70.6%

Figure 10: Negative/positive terrorism signatures

Conclusions

To summarize, we have found that similar linguistic expressions produced dramatically different text classification results. In particular, singular nouns often represented specific incidents while plural nouns referred to general events. Different verb forms (active, passive, infinitive) distinguished different tenses (past, future) and the presence or absence of objects (e.g., passive verbs do not require a known actor). And prepositions played a major role by implicitly acting as pointers to actors and objects. Finally, negated expressions behaved differently than their non-negated counterparts, not because the negated concept was explicitly required for the domain, but because the negated expressions conveyed subtle connotations about the event.

Researchers in the information retrieval community have experimented with phrase-based indexing to create more effective indexing terms (e.g., [Croft *et al.*, 1991; Dillon, 1983; Fagan, 1989]). However, most of these systems build complex phrases from combinations of nouns, noun modifiers, and occasionally verbs. Function words, such as prepositions and auxiliary verbs, are almost always ignored. Stopword lists typically throw away function words, preventing them from being considered during the generation of indexing terms. Many information retrieval systems also lose the ability to distinguish singular and plural nouns when they use a stemming algorithm.

Stopword lists and stemming algorithms perform valuable functions for many information retrieval systems. Stopword lists substantially reduce the size of inverted files and stemming algorithms allow the system to generalize over different morphological variants. However, some common stopwords *do* contribute substantially to the meaning of phrases. We believe that certain types of common stopwords, such as prepositions and auxiliary verbs, should be available for use in building complex phrases. Similarly, stemming algorithms may be appropriate for some terms but not for others. Users would likely benefit from being able to specify whether a term should be stemmed or not. Automated

indexing systems might also produce better results if both stemmed and non-stemmed indexing terms were available to them.

As disk space becomes cheaper, space considerations are not nearly as important as they once were, and we should think twice before throwing away potentially valuable words simply for the sake of space. Although many function words do not represent complex concepts on their own, their presence provides important clues about the information surrounding them. As text corpora grow in size and scope, user queries will be more specific and information retrieval systems will need to make more subtle domain discriminations. Our results suggest that information retrieval systems would likely benefit from including these small words as part of larger phrases. We have shown that the effectiveness of slightly different linguistic expressions and word forms can vary substantially, and believe that these differences can be exploited to produce more effective indexing terms.

References

Croft, W. B.; Turtle, H. R.; and Lewis, D. D. 1991. The Use of Phrases and Structured Queries in Information Retrieval. In *Proceedings, SIGIR 1991*. 32–45.

Dillon, M. 1983. FASIT: A Fully Automatic Syntactically Based Indexing System. *Journal of the American Society for Information Science* 34(2):99–108.

Fagan, J. 1989. The Effectiveness of a Nonsyntactic Approach to Automatic Phrase Indexing for Document Retrieval. *Journal of the American Society for Information Science* 40(2):115–132.

Frakes, William B. and Baeza-Yates, Ricardo, editors 1992. *Information Retrieval: Data Structures and Algorithms*. Prentice Hall, Englewood Cliffs, NJ.

Harman, D. 1991. How Effective is Suffixing? *Journal of the American Society for Information Science* 42(1):7–15.

Harman, D. 1992. The DARPA Tipster Project. *SIGIR Forum* 26(2):26–28.

Krovetz, Robert 1993. Viewing Morphology as an Inference Process. Computer science technical report 93-36, University of Massachusetts, Amherst, MA.

Lehnert, W. 1991. Symbolic/Subsymbolic Sentence Analysis: Exploiting the Best of Two Worlds. In Barnden, J. and Pollack, J., editors 1991, *Advances in Connectionist and Neural Computation Theory, Vol. 1*. Ablex Publishers, Norwood, NJ. 135–164.

Proceedings of the Fourth Message Understanding Conference (MUC-4), San Mateo, CA. Morgan Kaufmann.

Proceedings of the Fifth Message Understanding Conference (MUC-5), San Francisco, CA. Morgan Kaufmann.

Riloff, E. and Lehnert, W. 1994. Information Extraction as a Basis for High-Precision Text Classification. *ACM Transactions on Information Systems* 12(3):296–333.

Riloff, E. 1993. Automatically Constructing a Dictionary for Information Extraction Tasks. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*. AAAI Press/The MIT Press. 811–816.

Riloff, E. 1994. *Information Extraction as a Basis for Portable Text Classification Systems*. Ph.D. Dissertation, Department of Computer Science, University of Massachusetts Amherst.

Proceedings of the TIPSTER Text Program (Phase I), San Francisco, CA. Morgan Kaufmann.