TECHNICAL RESEARCH REPORT

Improving Information Retrieval Systems using Part of Speech Tagging

by A. Chowdhury, M. McCabe

T.R. 98-48

Improving Information Retrieval Systems using Part of Speech Tagging

Abdur Chowdhury
University of Maryland
abdur@isr.umd.edu

M. Catherine McCabe
George Mason University
cmccabe@gmu.edu

Abstract

The object of Information Retrieval is to retrieve all relevant documents for a user query and only those relevant documents. Much research has focused on achieving this objective with little regard for storage overhead or performance. In the paper we evaluate the use of Part of Speech Tagging to improve, the index storage overhead and general speed of the system with only a minimal reduction to precision recall measurements. We tagged 500Mbs of the Los Angeles Times 1990 and 1989 document collection provided by TREC for parts of speech. We then experimented to find the most relevant part of speech to index. We show that 90% of precision recall is achieved with 40% of the document collections terms. We also show that this is a improvement in overhead with only a 1% reduction in precision recall.

Introduction

Information Retrieval (IR) is directed towards finding relevant "documents" from unstructured textual data, in response to user requests (commonly referred to as queries). Computerized or automatic information retrieval has been a topic of both commercial development and research for many decades. Information Retrieval has grown beyond the initial interest by academics and defense department agencies. Many commercial organizations now deploy large IR systems, i.e., Lycos, Yahoo, Excite, Inquery, and Dialog, just to name a few. Most of these systems are not as concerned with "Recall" as they are with "Precision". They are very sensitive to system constrains like response time, disk usage, and CPU usage. So any way of improving on those factors is considered a contribution.

The Vector Space Model is a popular approach to Information Retrieval system implementation. To improve speed of a VSM system a common technique is to provide a stop word list. This stop word list is used during preprocessing to eliminate common words from the system. So non-discriminate words like "the" are eliminated from processing because they do not add any value to the precision recall metrics of a system but they add considerably to the storage required because they occur in all or most documents. The exploration of what words to eliminate is a topic of research. The information retrieval system developed at Cornell called SMART uses a stop word list of 571 words and has been widely adopted by the participants at the Text Retrieval Evaluation Conference (TREC). Thirty percent of the participants of the Fifth Annual Text Retrieval Evaluation Conference (TREC5) used the SMART stop word list, while other systems used lists ranging from 0 terms to 1500 terms. These lists have generally been developed beginning with most frequent terms and then manually adding and removing terms [Fox90]. We present a method of using Part of Speech (POS) tagging to reduce the number of words indexed

by the system. We also show that this can be done with little added parsing overhead. We show that the use of this approach can reduce the number of words to be indexed by 60%, thus improving the overall performance with a reduction of less than 1% average precision recall.

The decision of when to use POS tagging as an index size reduction technique should be based on user requirements. For commercial systems whose users expect fast retrieval and high precision for lower levels of recall, (they are only willing to look at the top 20 documents and do not care if there are additional relevant ones, this technique is very important. In general, the research community has not focused on query time or even index size but has focussed on finding all relevant documents and ranking them well We show that 90% of precision recall can be achieved with 40% of the document collections terms.

In section "Vector Space Model" we discuss prior work on the vector space approach to information retrieval systems. In section "Parts of Speech Analysis" we discuss parts of speech analysis. In section "Hypothesis and Experiments" we discuss our hypothesis and our experimental design. In section "Results" our experimental results are presented. In section "Conclusions" our conclusions and future work are presented.

Vector Space Model

The vector space model defines a vector that represents each document, and a vector that represents the query [Salt75]. There is one component in each vector for every distinct term that occurs in the document collection. Once the vectors are constructed, the distance between the vectors, or the size of the angle between the vectors, is used to compute a similarity coefficient.

Consider a document collection with only two distinct terms, *a* and *ß*. All vectors contain only two components. The first component represents occurrences of *a*, and the second represents occurrences of *ß*. The simplest means of constructing a vector is to place a one in the corresponding vector component if the term appears, and a zero, if the term does not appear. Consider a document, D_1, that contains two occurrences of term *a* and zero occurrences of term *ß*. The vector, *<1,0>*, represents this document using a binary representation. This binary representation can be used to produce a similarity coefficient, but it does not take into account the frequency of a term within a document. By extending the representation to include a count of the number of occurrences of the terms in each component, these frequencies can be considered. In the example, the vector would now appear as *<2,0>*.

Early work in the field used manually assigned weights. Similarity coefficients that employed automatically assigned weights were compared to manually assigned weights [Salt69, Salt70]. Repeatedly, it was shown that automatically assigned weights would perform at least as well as manually assigned weights [Salt69, Salt70].

Unfortunately, the above approach does not include the relative weight of the term across the entire collection. The utility of including a collection-wide based weight was studied in the 1970's, and the conclusion was that relevance rankings, the ordering of documents with respect to their relevance to the user query, improved if this weight was included. Although relatively small document collections were used to conduct the experiments, the authors still determined that "in so far as anything can be called a solid result in information retrieval research, this is" [Sprj76].

To construct a vector that corresponds to each document, consider the following definitions:

n = number of distinct terms in the document collection

tf_{ij} = number of occurrences of term t_j in document D_i

df_j = number of documents which contain t_j

$idf_j = \log d / df_j$ where *d* is the total number of documents

The vector for each document is of size *n* and contains an entry for each distinct term in the entire document collection. The components in the vector are filled with weights that are computed for each term in the document collection. The terms in each document are automatically assigned weights based on how frequently they occur in the entire document collection and how often a term appears in a particular document. The weight of a term in a document increases the more often the term appears in a document and the less often it appears in all other documents.

The weights computed for each term in the document collection are non-zero only if the term appears in the document. For a large document collection consisting of numerous small documents, the document vectors are likely to contain mostly zeros. For example, a document collection with 10,000 distinct terms results in a vector of size 10,000 for each document. A given document may have only 100 distinct terms. Hence, 9,900 components of the vector contain a zero.

The calculation of the weighting factor (*w*) for a term in a document is formally defined as a combination of term frequency (*tf*), document frequency (*df*), and inverse document frequency (*idf*). To compute the value of the *j*th entry in the vector corresponding to document *i*, the following equation is used:

$D_{ij} = (tf_{ij}) (idf_j)$

Consider a document collection that contains a document, D_1, with ten occurrences of the term *green* and a document, D_2, with only five occurrences of the term *green*. If *green* is the only term found in the query, document D_1 is ranked higher than D_2.

The inverse document frequency can best be examined when term frequency is not a factor. For the query containing the terms "*the elephant*" it is assumed that "*the*" occurs substantially more frequently that the term "*elephant*". For a document collection in which document D_1 contains one occurrence of "*the*" and document D_2 contains only one occurrence of the term "*elephant*", document D_2 will be ranked higher than D_1, and "*elephant*" will have a higher inverse document frequency than "*the*".

When a document retrieval system is used to query a collection of documents with *t* terms, the system computes a vector D of size *t* for each document. The vectors are filled with term weights as described above. Similarly, a vector Q is constructed for the terms found in the query.

A simple Similarity Coefficient (SC) between a query Q and an *i*th document D_i is defined as the Euclidean distance between the two vectors where, q_j is the *j*th term in the query and d_{ij} is the *j*th term in the *i*th document.

$$SC(Q, D_i) = \sum_{j=1}^{t} q_j * d_{ij}$$

First proposed in 1975, the vector space model is still a popular means of computing a measure of similarity between a query and a document [Salt89].

In 1988, several experiments tried to improve the basic combination of *tf-idf* weights [Buck88]. Many variations were studied, and the following weight function was identified as a good performer:

$$w_{ik} = \frac{\log(tf_{ik} + 1.0) * idf_k}{\sum_{j=1}^{t}((\log(tf_{ik} + 1.0) * idf_k)^2}$$

Several different means of comparing the query vector with the document vector have been implemented. The most common of these is the cosine measure where the cosine of the angle between the query and document vector is given:

The cosine coefficient is defined as:

$$sim(Q, D_1) = \frac{\sum_{j=1}^{t} w_{qj} d_{ij}}{\sqrt{\sum_{j=1}^{t}(d_{ij})^2 \sum_{j=1}^{t}(w_{qj})^2}}$$

Note that the cosine measure "normalizes"' the result by considering the length of the document. With the inner product measure, a longer document may result in a higher score simply because it is longer, and thus, has a higher chance of containing terms that match the query -- not necessarily because it is relevant. The cosine measure levels the playing field by dividing the computation by the length of the document. We note that Singhal, et al, have recently found that the field may have been leveled too much [Sing96] as a study of recent results showed that long, relevant documents were often excluded simply because they are long. Modified normalization have been used to correct for this.

Parts of Speech Analysis

The study of Natural Language Processing including Parts of Speech has been one of interest for many years for linguistic students, but not until recently has it been used for Information Retrieval. It has been shown that NLP techniques can be used effectively for IR tasks [Tzoukermann et. al. 97] Specifically, Tzoukermann states three main categories of linguistic distinctions for indexing terms and variants:

> **syntactic:** where the same word is used with the same meaning but in a different part of speech (i.e. 'technology for developing new products' and 'new product technology')
> **morphosyntactic:** where a different form of the same word is used, sometimes the part of speech changes also (i.e. vibrating over wavelets and wavelet vibrations)
> **semantic:** where the same meaning is expressed with different words (i.e. renal failure and kidney failure)

The phrase "Parts of Speech" refers to the syntactic role of terms in written text. Examples of Parts of Speech include nouns, verbs, adverbs, prepositions, conjunctions, interjections, etc. Recently this work has focused on creating algorithms for computers to automatically identify

parts of speech in text. The algorithms of today can consistently achieve over 90% accuracy in tagging parts of speech. [MUC 97].

There are several different approaches to parts-of-speech tagging. Algorithms have been developed that are based on statistical methods, probabilistic methods, and dictionary-based approaches. There are also systems that combine several of these ideas. The topic of this paper is not to discuss the different approaches or evaluate them, but to show an implementation of it to speed up IR systems.

Many research groups have used Parts of Speech Tagging in Information Retrieval tasks [Lu et al 97], [Pederson et. al. 97], [Robertson 90]. One idea has been to add noun phrases to all terms in an effort to better represent what the document is about and thus improve precision recall. Adding noun phrases to the index of all terms has been shown to improve precision/recall [Zhai 97]. However, [Crestani97] showed that indexing noun phrases alone and using the short form of queries (three words) degrades overall recall. [Pederson et. al. 97] also indexed selectively based on parts-of-speech. He indexed nouns, verbs, adjectives, adverbs, interjections, numerals, abbreviations, and participles and left out coordinating conjunctions, subordinating conjunctions, determiners, infinite markers, prepositions, and pronouns. Pederson's work implemented a weighting scheme based on parts of speech as well, which favored noun phrases and adjective phrases to combine term weights. Their results showed that adding these phrases and weighting schemes improved precision/recall over indexing terms alone. In addition, the augmented indexing improved overall recall.

In similar work, [Strzalkowski, et. al. 97] used a stop word list to eliminate some parts-of-speech from the index. In this work, they stopped closed-class words such as determiners, prepositions, pronouns, etc. as well as certain very frequent words. Their noun phrase identification was completed using straightforward POS tagging, after stemming. Different weighting schemes were used for the various parts and tuned for the best results.

We were unable to find any prior work that used just nouns to reduce the index size. One explanation is that the research systems are concerned with precision / recall metrics, while this approach will slightly reduce the precision/recall, while improving the overhead of the system.

Hypothesis and Experiments

Hypothesis:

We feel that by using POS tagging, one can index only the most relevant terms of a document collection. By reducing the number of items tagged three results will occur:

1) The index size will be reduced.
2) The search space for queries will be reduced resulting in an overall performance increase of the system
3) The precision of the system will be reduced.

We hypothesize that certain parts of speech, i.e., nouns, verbs, adverbs are better discriminators than others. Human, heuristic analysis of a few documents indicated that nouns most closely represent what a document is *about.* A noun is any person, place or thing. By extracting and reading through only the nouns, the gist of the text was retained. This was not true for verbs or other parts of speech.. Thus, we believe that nouns are the most important discriminators for information retrieval when compared to other parts of speech. We show through a comparison of

all parts of speech for information retrieval that nouns account for the most significant retrieval portion of precision

Measuring Accuracy:

Accuracy of an Information Retrieval System is commonly measured using the metrics of precision and recall. These are defined in Equation one below. Precision is the measure of how much junk (nonrelevant) documents get returned for each relevant one. Recall is the measure of how many of the relevant documents were found no matter what else was also found. These measures assume a prior knowledge of which documents are relevant to each query. The annual Text Retrieval Evaluation Conference (TREC) generates a test set of queries and the relevant documents using the TREC corpus. We used this data in our experiments in order to allow a measure of the accuracy of using nouns-only.

$$\text{Precision} = \frac{\text{Relevant Retrieved}}{\text{Retrieved}}$$

$$\text{Recall} = \frac{\text{Relevant Retrieved}}{\text{Relevant}}$$

Equation 1 : Precision Recall Definition

Experimentation:

The corpus we will use to test this hypothesis is "The Los Angeles Times" collection. The data represents a sampling of approximately 40% of the articles published by the Los Angeles Times in the two-year period from Jan 1, 1989 -- December 31, 1990. The total collection size is 495,415,000 bytes. The collection consists of 730 files ranging from 400K to 1000K in size. The files are stored in SGML format, signifying the start and end of each document in the files.

The first portion of the experiment is to tag the collection with parts of speech tags. The tags used are:

Nouns
Verbs
Adjectives
Adverbs
Other

We used the INSO (Parts of Speech) Parser to parse the parts of speech and tag them. We used only the five categories of speech listed above because of the limitation of the INSO parser. The reason we chose the INSO parser was because it could be configured to do only tagging, where others like the Apple Pie Parser [APP Ref] did full parse trees. Although the full parse tree is better for more in depth analysis, it is very slow. The Apple Pie Parser and other POS parsers did too much work with a huge overhead cost. Our goal is to simply find the most relevant tokens as quickly as possible.

We used the GMU Information Retrieval System [GMU Ref]. We disabled all added IR techniques like relevance feedback, stemming, etc. so a comparable baseline could be obtained.

The GMU system is a Vector Space Model IR engine, which uses the inverse document frequency with normalization for document length as similarity measure.

With the INSO parser and GMU IR system, we evaluate three experiments. The goal of the experiments is to show that nouns are the most relevant portions of text while other parts of speech do not provide as much differentiation. The first experiment is a baseline experiment with all POS used as tokens and indexed into the system. The second experiment indexes nouns only, and eliminates all other POS, verbs, adjectives, adverbs, etc. The third and last experiment shows the non-relevance of other parts of speech. Below is an enumeration of the experiments:

- Nouns only
- All terms
- Verbs, Adjectives, Adverbs, and Other

The 50 queries from TREC6 (Topics 301-350) were used along with the relevant document list (qrel) provided by NIST [Voorhees96]. In addition, the TREC6 trec_eval program was used to calculate the precision recall. As in TREC, the average precision recall is used as the basis of comparison, see Equation 1 for a definition of precision and recall.

The goal of these experiments is to show that nouns are the most relevant terms. That using nouns will reduce the overhead of the system as a whole and that nouns make up less that half of the terms therefore are a good means of reducing overhead without significantly affecting precision recall.

Results

The tagging of parts of speech by the INSO Part of Speech Tagger resulted in the following breakout of terms in the corpus. (See figure 1). This reflects 500 MB Los Angeles Times Corpus described above.

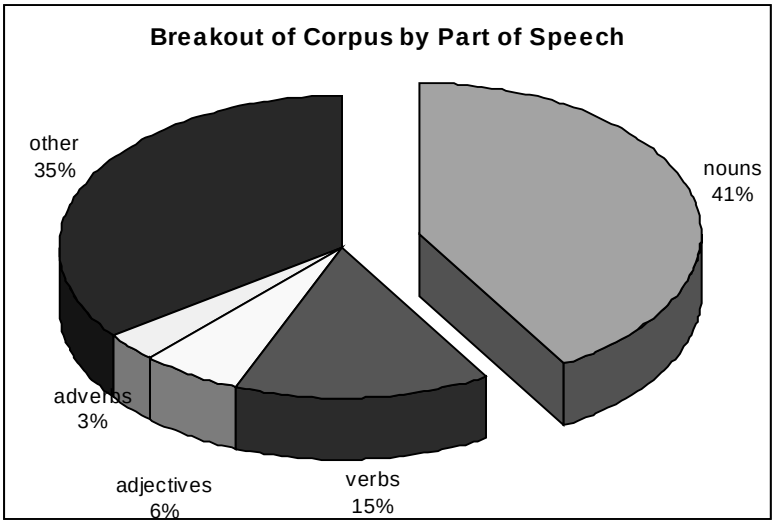

Figure 1: Breakdown of Corpus Based on POS

The First experiment indexed all tokens as a baseline experiment. The indexing and query of all terms determines a Precision / Recall metric that subsets of the POS can be evaluated against. We indexed terms only (no phrases or stemming) and used no relevance feedback or query expansion techniques. The average precision recall for all three experiments can be seen in the following figure.

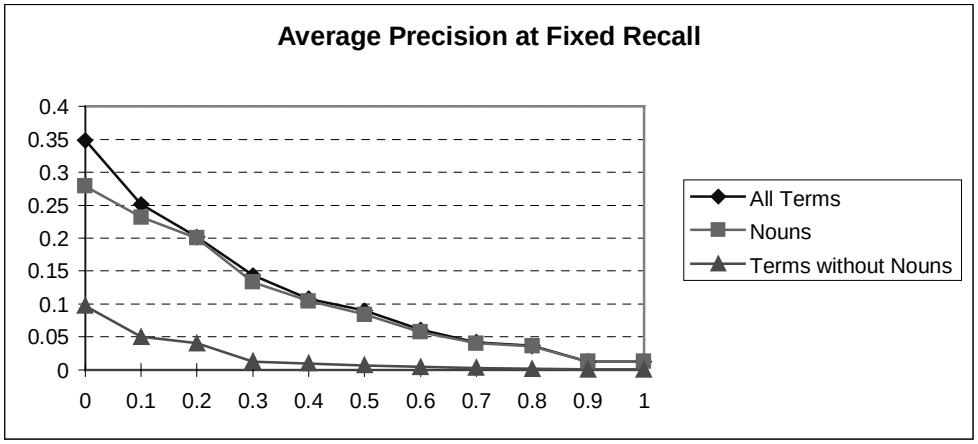

Figure 2: Average Precision at Fixed Recall

As shown in the graph the best results come from all terms. The above results do not come as a surprise, but should be noted that the goal is to reduce the load on the system with a minimal effect on precision recall. We show that nouns-only results in very close accuracy for most queries. In fact, 16 of the 50 queries actually had better average precision/recall when using nouns-only versus all terms. Therefore the most important discriminator in POS tagging is nouns for information retrieval systems.

We broke down these results by query to see the variability in results for each index on each query as shown in Figure 3.

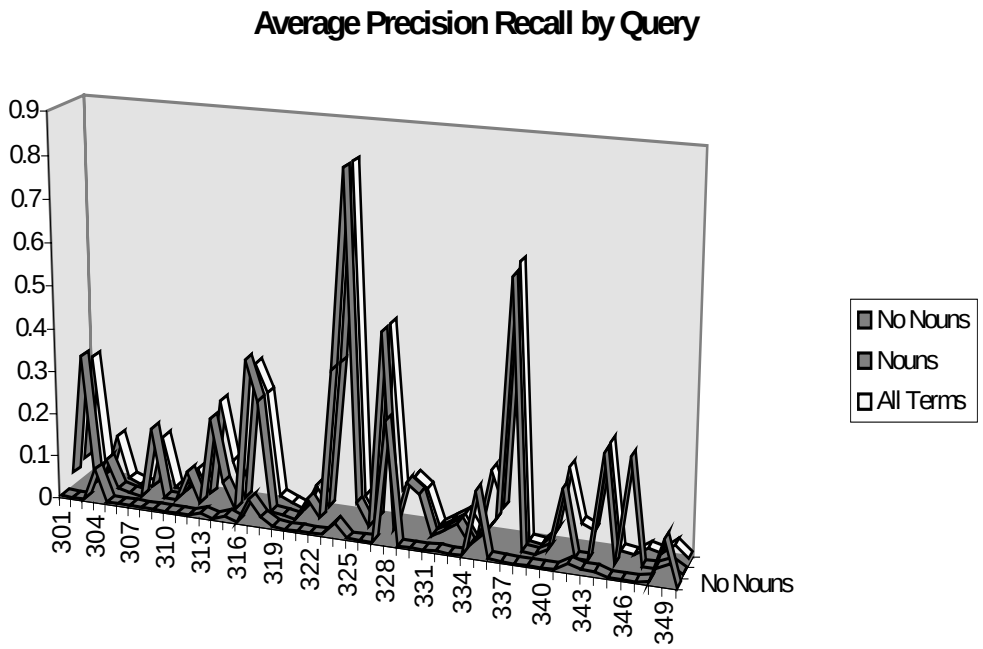

Figure 3: Precision Recall by Query

As shown in the graphs the best results come from all terms. The above results do not come as a surprise, but should be noted that the goal is to reduce the load on the system with a minimal effect on precision recall. We show that "nouns" only results in very close results for most queries. Therefor the most important discriminator in POS tagging is a noun for information retrieval systems. We found queries that performed better when nouns only or all terms when nouns were removed. Below is an example of such a query.

<top>
<num> Number: 349
<title> Metabolism
<desc> Description:
Document will discuss the chemical reactions necessary to keep living cells healthy and/or producing energy.
<narr> Narrative:
A relevant document will contain specific information on the catabolic and anabolic reactions of the metabolic process. Relevant information includes, but is not limited to, the reactions occurring in metabolism, biochemical processes (Glycolysis or Krebs cycle for production of energy), and disorders associated with the metabolic rate.

</top>

The better precision recall measurements might be explained by nouns being interpreted in different sense so the nouns bring the query closer to a different sense and by removing the nouns this a disambiguation occurs helping the P/R measurements.

When examining Figure 3 we see that nouns-only closely follow the results from all terms. In terms of actual relevant documents retrieved, the nouns-only system 565 were found by using all terms, 484 were found using only nouns. This difference of 81 documents represents 7.3% of the total number of relevant documents (1105). The average precision recall across all fifty queries was less than one percent less for nouns only than for all terms. The average precision for all terms was .1067 and for nouns-only was .0984. All terms without nouns (adverbs, adjectives,

verbs and other) were extremely low at .0159. Almost half (22) of the queries returned no relevant documents when nouns were removed from the index.

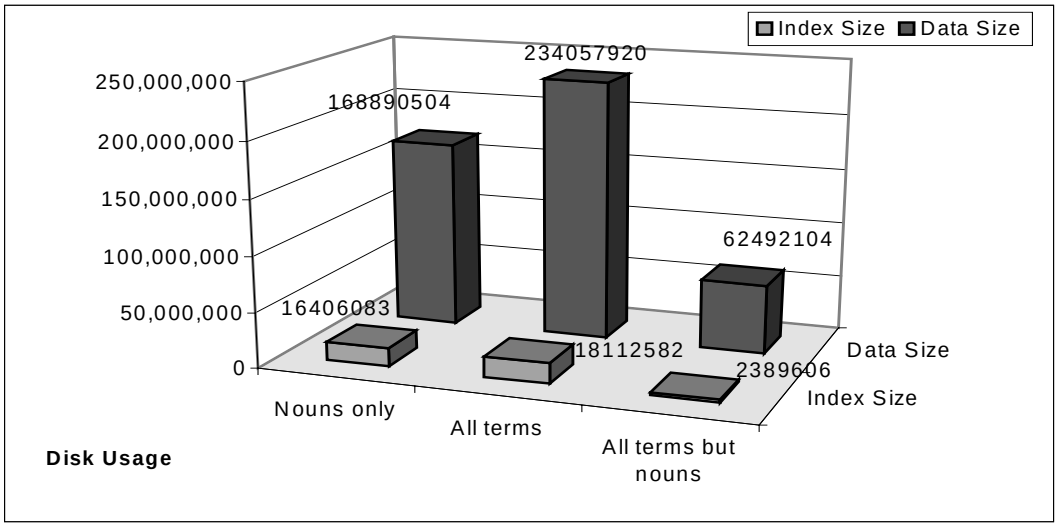

Figure 4: Disk Usage

The results of actual system usage are given in Figure 4: Disk Usage. These results are system specific, but can be used as a comparison to other systems. Figure 4 shows a reduction in the posting list of 28%, and a reduction of 9.5% for the index storage size. There is more data stored in the VSM, but data like the document index remain constant because they are the same for both implementations. The slowest parts of any IR system are disk seeks and reads. Any reduction in data size will improve the overall system speed.

Conclusions

In conclusion, only indexing nouns reduce the system's average precision recall by less than 1%. The nouns make up 40% of the collection giving us a 60% improvement in indexing overhead. The disk savings are over 28% for the posting list storage and 9.5% for the index data. The system describe here has definite advantages for commercial systems concerned with overall disk usage and speed.

In this paper, we have briefly covered related work including the Vector Space Model, and using Parts of Speech in information retrieval. We have described a method of reducing the number of tokens to index with a minimal reduction in the precision recall metrics. We have shown that this approach reduced the disk usage needed and will speed up the processing of new documents by reducing the amount of work by 60%. In our future work, we plan to index noun phrases in addition to nouns which will only add 10% to then indexing storage. We feel that this may improve the systems precision recall evaluations, better than terms only. Further future work will involve analysis of parts of speech manipulation for relevance feedback and thesauri approaches to improving precision recall. If nouns-only are effective as index entries, an approach to relevance feedback would be to only use the nouns in query expansion.

References

(Buckley95) Buckley, C., A. Singhal, M. Mitra, and G. Salton. New Retrieval Approaches Using SMART: TREC 4. Text Retrieval Conference sponsored by National Institute of Standards and Technology and Advanced Research Projects Agency, 1995.

[Blai85] Blair, D. and Maron, M. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. Communications of the ACM, 28(3):289-299.

[Buck88]

[Crestani97] Crestani, F., Sanderson, M., Theophylactou, M. Short Queries, Natural Language and Spoken Documents Retrieval: Experiments at Glasgow University. Proceedings of the Sixth Text Retrieval Conference (TREC6), sponsored by the National Institute of Standards and Technology and the Advanced Research Projects Agency, November 1997.

(Evans et al 96) Evans, D. and Zhai, C. Noun-phrase analysis in unrestricted text for information retrieval. Proceedings of the 34th annual meeting of association for computational linguistics, santa ruz, University. of CA June 24-28, 1996 p 17-24.

(Fagan 97) Fagan Joel L. Experiments in Automatic Phrase Indexing for Document Retrieval: A comparison of syntactic and nonsyntactic methods. PhD Thesis Cornell University Sept 1997.
(Fox90) Fox, Christopher. A Stop List for General Text. SIGIR Forum, (v. 24, no. 1-2) 1990 p,. 19-35.

[Fuller et. al. 97] Fuller, M., Kaszkiel, M. et. al. MDS TREC6 Report. Proceedings of Sixth Text Retrieval Conference (TREC6), sponsored by the National Institute of Standards and Technology and the Advanced Research Projects Agency, November 1997.

(Grossman95) Grossman, D., D. Holmes, O. Frieder, M. Nguyen, and C. Kingsbury. Improving Accuracy and Run-Time Performance for TREC-4. Proceedings of the Fourth Text REtrieval Conference (TREC), sponsored by the National Institute of Standards and Technology and the Advanced Research Projects Agency, November 1995.

[Lu et al 97] Lu, A. Meier, Rao, A., Miller, D., Pliske, D. Query Processing in TREC6. Proceedings of the Sixth Text Retrieval Conference (TREC6), sponsored by the National Institute of Standards and Technology and the Advanced Research Projects Agency, November 1997.

[Pederson et. al. 97] Pederson, J. Silverstein, C. and Vogt, C. Verity at TREC6: Out of the Box and Beyond. Proceedings of the Sixth Text Retrieval Conference (TREC6), sponsored by the National Institute of Standards and Technology and the Advanced Research Projects Agency, November 1997.

[Robertson 90] Robertson, S.E. On term selection for query expansion. Journal of Documentation 46. 4, 359-364

[Salt75] Salton, G., Wong, A., and C.S. Yang, "A Vector Space Model for Automatic Indexing," *Communications of the ACM*, pp. 613-620.

[Salt69] Salton, G. (1969). A comparison between manual and automatic indexing methods. Journal of American Documentation, 20(1):61-71.

[Salt70] Salton, G. (1970). Automatic text analysis. Science, 168(3929(:335-342.

[Salt89] Salton, G., Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer, Addison-Wesley Publishing Company, Inc., Reading, Massachusetts, 1989.

[Singhal96] Singhal, A., C. Buckley, and M. Mitra. Pivoted Document Length Normalization. *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Ed. Hans-Peter Frei, Donna Harman, Peter Schauble and Ross Wilkinson, SIGIR Forum, August 18-22, 1996.

[Strzalkowski,97] Strzalkowski, T. and Lin, F. Natural Language Information Retrieval TREC6 Report. Proceedings of the Sixth Text Retrieval Conference (TREC6), sponsored by the National Institute of Standards and Technology and the Advanced Research Projects Agency, November 1997.

[Tzoukermann, E. Klavans, J. Jaquemin, C. Effective Use of Natural Language Processing Techniques for Automatic Conflation of Multi-Word Terms: The Role of Derivational Morphology, Part of Speech Tagging, and Shallow Parsing. SIGIR '97. ACM 1997. P. 148 - 155

[Voorhees96] Voorhees, E. and Harmen, D. Overview of the Sixth Text Retrieval Conference (TREC-6) Proceedings of the Sixth Text REtrieval Conference (TREC), sponsored by the National Institute of Standards and Technology and the Advanced Research Projects Agency, November 1997.

[Zhai97] Zhai, C. Fast statistical parsing of noun phrases for document indexing. 5th conference on applied natural language processing. Wash. D.C. March 31 - April 3, 1997.

(Zhai95) Zhai, C. and Tong, X. Evaluation of Syntactic Phrase Indexing – CLARIT NLP Track Report. Proceedings of the Fifth Text Retrieval Conference (TREC), sponsored by the National Institute of Standards and Technology and the Advanced Research Projects Agency, November 1996.

[Sprj76]