http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 4 (1): 32-37, 2005
ISSN 1812-5638

Improving Arabic Information Retrieval Systems Using Part of Speech Tagging

Ghassan Kanaan, Riyad al-Shalabi and Majdi Sawalha
Yarmouk University, Irbid, Jordan

Abstract: The objective 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. This study have evaluated the use of Part of Speech Tagging to improve the index storage overhead and general speed of the system with only a minimal increment in precision and recall measurements. We tagged 242 abstracts of Arabic documents using the Proceedings of the Saudi Arabian National Conferences as a source. All these abstracts involve computer science. We also built an automatic information retrieval system to handle Arabic data. We then did a series of experiments to identify the most relevant part of speech indexing method.

Key words: Arabic word, part of speech tagging, automatic indexing, information retrieval

INTRODUCTION

Computer applications, software systems and internet resources are available and easy to use in almost every field in the Arabic countries. Most of these applications use English in application interfaces and data processing since Arabic natural language processing systems are not widely available and few software development companies have shown interest in Arabic natural language processing systems. This problem has compelled many Arab computer researchers to focus on Arabic language information retrieval and natural language processing systems[1].

Arabic is the official language of twenty Middle Eastern and African countries and it is the religious language of all Muslims, regardless of their origin. It is therefore surprising that very little work has been done on Arabic corpus linguistics. Arabic differs from Indo-European languages both syntactically and morphologically. It is a Semitic language whose main characteristic feature is that most words are built up from roots by following certain fixed patterns and adding infixes, prefixes and suffixes. It is an old language and what is now known as Classical Arabic was standardized around fourteen centuries ago[1].

The modern form of Arabic is called Modern Standard Arabic (MSA) and it is the form used in all Arabic-speaking countries in publications, the media and academic institutions[1].

AUTOMATIC INDEXING

The goal of Information Retrieval (IR) is finding relevant “documents” from unstructured textual data, in response to user 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[2].

The classical model in information retrieval considers that each document is described by a set of representative key words called index terms. An index term is simply a (document) word whose semantics helps in remembering the document’s themes. Thus, index terms are used to summarize the document contents. In general, index terms are mainly nouns because nouns have meaning by themselves and thus, their semantics is easer to identify and to grasp. Adjectives, adverbs and connectives are less useful as index terms because they work mainly as complements. However, it may be interesting to consider all the distinct words in a document collection as index terms. For instance, this approach is adopted by some web search engines[2,3].

The three classic models in information retrieval are called the Boolean, the vector space and the probabilistic models. In the Boolean model, documents and queries are represented as sets of index terms. Thus, we can say that the model is set theoretic. In the vector model, documents and queries are represented as vectors in a t-dimensional space, where t is the number of index terms. Thus, we can say that the model is algebraic. In the probabilistic model, the framework for modeling document and query representation is based on probability theory. Thus, as the name indicates, we say that the model is probabilistic[3].

Corresponding Author: Ghassan Kanaan, Yarmouk University, Irbid, Jordan
Tel: 00 962 2 7211111 Fax: 00 962 2 7211128i E-mail: ghassank@yu.edu

The vector space model: The vector space model defines a vector that represents each document and a vector that represents the query. 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[3].

The vector space model recognizes that the use of binary weights is too limiting and proposes a framework in which ranking is easier. This is accomplished by assigning non-binary weights to index terms in queries and in documents. These term weights are ultimately used to compute the degree of similarity between each document stored in the system and the user query. By sorting the retrieved documents in decreasing order of this degree of similarity, the vector model takes into consideration documents that match the query terms only partially. The main resultant effect is that the ranked document answer set is a lot more precise (in the sense that it better matches the user's information needs) than the document answer set retrieved by the Boolean model[3].

Definition for the vector model, the weight $w_{i,j}$ associated with a pair (k_i,d_j) is positive and non-binary. Further, the index terms of the query are also weighted. Let $w_{i,q}$ be the weight associated with the pair $[k_i,q]$, where, $w_{i,q} >= 0$. Then, the query vector q is defined as $q = (w_{1,q}, w_{2,q}, \ldots, w_{t,q})$ where t is the total number of index terms in the system. As before, the vector for a document d_j is represented by $d_j = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$[3].

A document d_j and a user query q are represented as t-dimensional vectors. The vector space model typically evaluates the degree of similarity of the document d_j with regard to the query q as the correlation between the vectors d_j and q. This correlation can be quantified, for instance, by the cosine of the angle between these two vectors. That is,

$$\mathrm{Sim}\,(d_j,q) = (d_j .\, q)/(\,|d_j| \text{ x } |q|\,)$$

$$= (\textstyle\sum i = 1 w_{i,j} \text{ x } w_{i,q})/(\sqrt{\sum_{t=1}^{t} w_{i,j}^2} \text{ x } \sqrt{\sum_{t=1}^{t} w_{i,q}^2})$$

Where, $|d_j|$ and $|q|$ are the norms of the document and query vectors.
$\mathrm{Sim}(d_j,q)$ is the similarity between document d_j and query q.
t is the number of index terms in the system.
$w_{i,j}$ is the weight of the ith index term in the document j.
$w_{i,q}$ is the weight of the ith index term i in the query q.

After the similarity has been calculated, the documents are listed in descending order according to their similarity.

Index term weights can be calculated in many different ways. Since we use the vector space model we use the following formulas to calculate the weight of each term[3,4].

$$w_{i,j} = f_{i,j} * idf_i$$

Where, $f_{i,j}$ is called the normalized frequency and calculating according to the following formula.

$$f_{i,j} = freq_{i,j} / max_l\, freq_{l,j}$$

$freq_{i,j}$ is the frequency of the term i in the document j.
$max_l\, freq_{l,j}$ is the maximum frequency computed in the document j.
idf_i is called the inverse document frequency for the index term i, idf_i is given by :

$$idf_i = \mathrm{Log}\,(N/n_i)$$

Where:
N : is the total number of document in the system (242 in our system),
n_i: is the number of documents in which the index term i appears.

Since we use the vector model we use the following formulas to calculate the weight of each term in the queries.

$$w_{i,q} = (0.5 + ((0.5 * freq_{i,q})/max_l\, freq_{l,q}) * \log N/n_i$$

Where:
$w_{i,q}$ is the weight of index term i in the query q.
$Freq_{i,q}$ is the frequency of term i in query q.
$Max_l\, freq_{l,q}$ is the maximum frequency calculated in the query q.
N is the total number of documents in the system.
n_i is the number of documents in which the index term i appears.

THE ARABIC WORD

Most vowels in written Arabic are represented by diacritic marks. Most modern text is printed in devowelized form without these diacritic marks. While most words in a text are traditional Arabic words, some words are "Arabized" loan words from other languages (perhaps with some phonetic adjustments to facilitate pronunciation). The original Arabic words are divided in turn into two sub categories; derived Arabic words, which

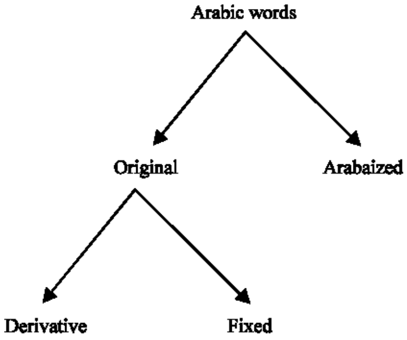

Fig. 1: The classification of Arabic words

are the verbs and nouns that are built according to the Arabic derivation rules and fixed Arabic words, which are a set of words molded by Arabs, in ancient times and do not obey the Arabic derivation rules (Fig. 1)[1].

Arabic word categories: Arabic grammarians traditionally classify words into three main categories: Nouns, verbs and particles. These categories are also divided into subcategories that collectively cover the whole of the Arabic language. These categories are:

Nouns: A noun in Arabic is a name or a word that names a person, thing, or an idea. Nouns are also divided into the following types[1]:

- An agent noun is a derived noun indicating the actor of the verb or its behavior. "كاتب" " writer".
- A patient noun is a derived noun indicating the person or thing that undergoes the action of the verb "مقتول" victim ". [or make it definite and say “the kill”]
- An instrument noun is a noun indicating the tool of an action
- An adjective is considered to be a type of noun in traditional Arabic grammar.
- An adverb is a noun that is not derived and that indicates the place or the time of the action.
- A proper noun is the name of a specific person, place, organization, thing, idea, event, date, time, or other entity.

Nouns are divided into two classes according to their origins

- Primitive noun "الأسماء الجامدة" , which are substantives having no Arabic root
- Derivative nouns “الأسماء المشتقة” , which can be substantive or adjectives "الصفات" .

Verbs: Verbs indicate an action, although the tenses and aspects are different[1]. Verbs typically appear in a variety of moods: the most common are Perfect, Imperfect and Imperative.
Verbs are categorized into three main aspects parts:

1. Perfect 2. Imperfect 3. Imperative

The definition of perfect verbs includes[1]:

- Equivalent of English past tense verbs (i.e. to describe acts completed in some past time).
- Describes acts which at the moment of speaking have already been completed and remain in a state of completion.
- Describes a past act that often took place or is still taking place (i.e. commentators are agreed (have agreed and still agree)).
- Describes an act which is just completed at the moment by the very act of speaking it.
- Describes acts which are certain to occur, that is it can be described as having already taken place (mostly used in promises, treaties and so on).

The imperfect does not in itself express any idea of time; it merely indicates a begun, incomplete, or enduring existence either in present, past or future time, while imperative verbs order or ask for something to be done in the future.

Particles: The Particle class includes: Prepositions, Adverbs, Conjunctions, Interrogative Particles, Exceptions and Interjections. The subcategories of particle are:

Prepositions	Adverbial	Conjunctions
Interjections	Exceptions	Negatives
Answers	Explanations	Subordinates

Examples of particles include:

- Prepositions "في" “in”
- Adverbial particles "سوف" “shall”
- Conjunctions "و" “and”
- Interjections "يا" “you”
- Exceptions "سوى" “Except”
- Negatives "لم" “Not”
- Answers "أجل" “yes”
- Explanations "يا" “that is”
- Subordinates "لو" “if”

PART-OF-SPEECH TAGGING

Part-of-speech tagging is the process of assigning grammatical part-of-speech tags to words based on their context. Many automated tagging systems have been developed for English and many other western languages and for some Asian languages[5-8] and have achieved accuracy rates ranging from 95 to 98%. A tagged corpus has more useful information than untagged corpus, so the tagged corpus can be used to extract grammatical and linguistic information from the corpus. This information can then be used for many applications such as creating dictionaries and grammars of a language using real language data. Tagged corpora are also useful for detailed quantitative analysis of text[5,6]. A tag is a code that represents some features or set of features and is attached to the segment in a text. Tags may carry simple or complex information[5-8].

Part-of-speech tagging is designed to assign part-of-speech tags for all index terms in the collection. We used the full automatic Arabic text tagging system implemented by Ghassan Kanaan, Riyad Al-Shalabi and Majdi Sawalha[5].

The noun tagging process: We have constructed many rules to identify nouns from analyses of the grammar of the Arabic language.

The rules for extracting nouns from documents are making use of the affixes of the word, its position in the sentence and its patterns.

Verb tagging process: This process is responsible for identifying the verbs in the document. A verb is defined as a word that indicates a meaning by itself combined with a tense or time. Verbs take words or letters as affixes such as the particles "قد" , "سوف" , the pronouns and the letters "س", "ت" , "ن" [1].

Like nouns verbs can be identified by their patterns, their affixes, their position in the sentence.

HYPOTHESIS AND EXPERIMENTS

In this study we examine the effects of using part of speech tagging in information retrieval systems applied to the Arabic language is examined. This use of part of speech tagging has the potential to reduce the index size, reduce the search space for queries resulting in an overall performance increase of the system, while it affects the precision of the system in only a minimal fashion.

Present hypothesis is that certain parts of speech (nouns, verbs, particles...) are better for use as index terms than others. 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 precision[2,3].

RESULTS

The goal of our experiments is to study the effect of using part of speech information (nouns, verbs, particles...) in choosing index terms on the performance of Arabic language information retrieval systems.

As we stated previously, we used a collection of 242 abstracts of Arabic documents from the Proceedings of the Saudi Arabian National Conferences as a source and 36 queries to test our information retrieval system.

The first step in the experiments is to tag the collection with part of speech tags.

The main tags used are:

- Nouns.
- Verbs.
- Particles.
- Other parts of speech.

We used the Full Automatic Arabic Part of Speech Tagging System. We used the word categories listed above. Figure 2 shows the resulting breakdown of terms in our corpus of 242 documents.

We designed and built an automatic information retrieval system to handle Arabic data. This automatic information retrieval system is divided into three sub parts; one handles all words as index terms, the second handles nouns as index terms and the last one handles words of other parts of speech as index terms.

The goal of our experiments was 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 was a baseline experiment with all parts of speech used as tokens and indexed in the system. The second experiment indexed nouns only and eliminated all words belonging to other parts of speech. The third and last experiment showed the non-relevance of other parts of speech. Below is an enumeration of the experiments:

- All terms
- Nouns only
- Other parts of speech.

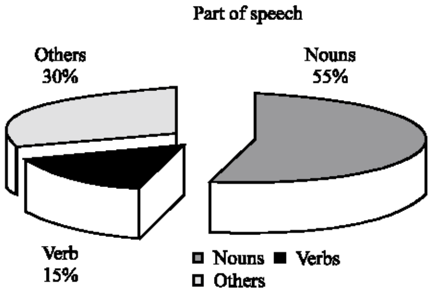

Fig. 2: Breakdown of our corpus based on part of speech

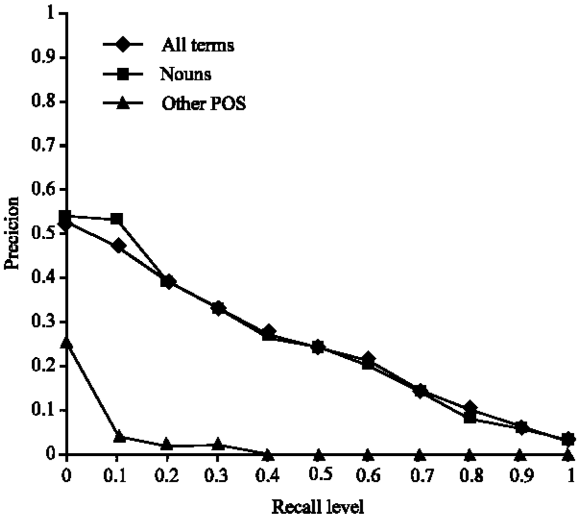

Fig. 3: Average interpolated recall and precision

We used a collection of 36 queries. Since the 242 documents in the IR System are in the computer science area, three graduate students in the computer science department were asked to write number of queries and make the relevance judgment for them. From these queries we chose 36 queries to test our system.

In addition, we built a program that evaluates the system by calculating the precision/recall and the average precision/recall, which is used as the basis of the comparisons.

Recall and precision are the common measurements used to evaluate the effectiveness of an information retrieval system.

Suppose R is the set of relevant documents for a given query.

|R| is the number of documents in the set R.
A is the retrieved documents for the query q.
|A| is the number of documents in set A.
|Ra| is the number of documents in the intersection of the sets R and A.

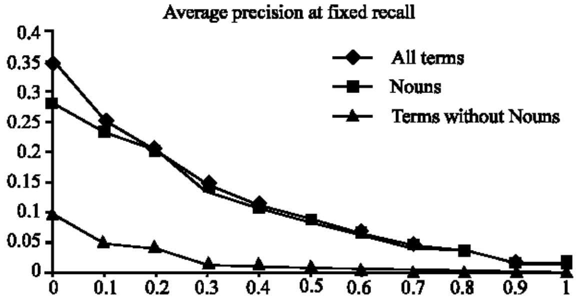

Fig. 4: Average precision at fixed recall

Recall is the fraction of the relevant documents that have been retrieved.

$$\text{Recall} = \frac{|Ra|}{|R|}$$

Precision is the fraction of the retrieved documents which are relevant.

$$\text{Precision} = \frac{|Ra|}{|A|}$$

Since the recall levels for each query may be distinct, an interpolation procedure is necessary. We used an interpolation procedure with 11 standard recall levels. Let rj, j ε {0, 1, 2,..., 10}, be a reference to j-th standard recall level (i.e., r_4 is a reference to recall level 40%). Then

$$P(r_j) = \max r_j <= r <= r_{j+1} P(r)$$

which states that the interpolated precision at the j-th standard recall level is the maximum known precision at any recall level between the j-th recall level and the (j+1)-th recall level.

Figure 3 shows the average interpolated recall and precision after applying the steps described above.

The results from using all terms and using nouns alone are almost the same (Fig. 3). The goal of this paper is to reduce the load on the system with a minimal effect on precision and recall. We show that "nouns only" results in very close results for most queries. Therefore the most important discriminator in part- of Speech tagging is a noun for information retrieval systems. We found that queries performed better using nouns only or using all terms than when nouns were removed[2].

Our experiments show results similar to those obtained in experiments done by Abdur Chowdhury and M. Catherine McCabe[2] as shown in Fig. 4., Their experiments were done for the English language.

The goal of these experiments was to show that nouns are the most relevant terms. The fact is that using nouns alone will reduce the overhead of the system as a whole, since nouns make up about half of the total terms. Therefore using nouns alone is a good means of reducing overhead without significantly affecting precision and recall[2,3].

CONCLUSIONS

In conclusion, indexing nouns only reduces the system's average precision/recall by less than 1%. The nouns make up 55% of the collection giving us a 45% improvement in indexing overhead. The research described here suggests that our approach of indexing nouns only has definite advantages for commercial systems concerned with overall disk usage and speed.

This study has 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 45%. In our next experiment, we are thinking of indexing Arabic noun phrases in addition to Arabic nouns, which will only add 10% to the indexing storage. We believe that this may improve the system precision/recall evaluations, to give a better result than single terms only. Really good morphological analysis can improve stemming and thus improve the results as well[9-11]. Further future work will involve analysis of part 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 use only the Arabic nouns in query expansion.

REFERENCES

1. Ali, N., 1988. Computers and the Arabic Language. Cairo, Egypt: Al-Khat Publishing Press, Ta'reep.
2. Abdur Chowdhury and M. Catherine McCabe, 1998. Improving information retrieval systems using part of speech tagging. ISR, Institute for Systems Research, pp: 48.
3. Ricardo Baeza-Yates and Berthier Ribeiro-Neto, 1999. Modern Information Retrieval, ACM Press, New York, Addison-Wesley.
4. Ismail Hmeidi, Ghassan Kanaan and Martha Evens, 1997. Design and implementation of automatic indexing for information retrieval with Arabic documents. J. American Society for Inform. Sci., 48: 867-881.
5. Ghassan Kanaan, Riyad Al-Shalabi and Majdi Sawalha, 2003. Full automatic Arabic text tagging system. Proceedings of the International Conference on Information Technology and Natural Sciences ICITNS, 2003, pp: 258-267.
6. Saleem Abuleil and Martha W. Evens, 2002. Extracting an Arabic lexicon from Arabic newspaper text. Computers and the Humanities, 36: 191-221.
7. Shreen Khoja, 2001.APT: Arabic part-of-speech tagger. Proceedings of the Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL2001), Carnegie Mellon University, Pittsburgh, Pennsylvania. June 2001.
8. Shreen Khoja, Porger Garside and Gerry Knowles, 2001. A tagset for the morphosynactic tagging of Arabic. Paper presented at Corpus Linguistics 2001, Lancaster University, Lancaster, UK, March 2001.
9. Mohamed Attia Mohamed Elaraby Ahmed, 2000. A Large-scale computational processor of the Arabic morphology and applications M.Sc. Thesis, Faculty of Engineering, Cairo University, Cairo, Egypt.
10. Saleem Abuleil, Khalid Alsamara and Martha Evens, 2002. Acquisition system for Arabic noun morphology. Proceedings of the Computational Approaches to Semitic LanguagesWorkshop, University of Pennsylvania, 11th July 2002, pp: 19-26.
11. Riyad Al-Shalabi and Martha Evens, 1998. Computational morphology system for Arabic. Workshop on Semitic Language Processing. COLING-ACL98, University of Montreal, Montreal, PQ, Canada, pp: 66-72.