SEMI-AUTOMATIC DOCUMENT PROCESSING IN AN EXPERIMENTAL MODEL OF INFORMATION RETRIEVAL SYSTEM

V. Pachovski[1], M. Kon-Popovska[2]

[1)] Faculty of Civil Engineering, Sts. Cyril and Methodius
Partizanski odredi bb, 1000 Skopje, Macedonia
[2)] Faculty of Natural Sciences and Mathematics, Sts. Cyril and Methodius
Arhimedova b.b., PO Box 162, 1000 Skopje, Macedonia
matematika@gf.ukim.edu.mk, margita@pmf.ukim.edu.mk

Abstract: The paper presents an approach to document processing in an information retrieval system. Namely, when documents are entered in the Information retrieval system, some kind of language processing has to be applied. Considering the diversity of grammar rules of different natural languages, this processing can be performed by partial expert intervention while entering the documents. In the paper, certain characteristics of documents written in Macedonian language (grammar) and Cyrillic alphabet are discussed and a kind of semi-automatic system approach to document processing and related database structures in order to store parts of the expert knowledge is proposed. The aim is to enable the system to perform automatic language processing with some degree of confidence and use expert interventions, to avoid automatic context analysis. The approach is continually tested on a test-collection containing over 4.000 rather short documents, common news, growing by 10-15 on daily bases. The work represents a part of research in Information retrieval systems, encountering specifics of documents in native language and alphabet, aiming to improve the quality of retrieval in national information retrieval systems.

Keywords: retrieval, database, processing, information retrieval, database systems, document processing

1 Introductory remarks

The paper represents a part of our continuous research in Information retrieval of documents in Macedonian language and Cyrillic alphabet. As a starting point we presume previous results. The database of documents is a well-defined relational database structure, with removed stop-words [4]. Quotients of relevance are available [5],[6] as elements of weighted retrieval, based on a Bayes formula, measuring the strength of a document-word relation as well as word-subject relation. We discuss language aspects of our developed information retrieval system for archiving documents in "small scale" organizations [7]. It has been tested with acceptable response time on a rather small collection of documents (under 10^4, see its visual representation

below), and has yet to be tested on a larger collection of 10^6 documents or above. A test-collection contains rather short documents common news growing on the rate up to 30 documents per day, with a growing dictionary of over 40.000 words, almost 15000 non-identified words and a table of 380.000 records of words within the documents.

2 The outline of the retrieving proccess

Generally the process of retrieving is done in two phases. In the first phase, the system generates **primary answer table.** This table of retrieved words contains all the different forms of the words in request that could be found in the database, the number of occurrences and weights (relative relevancy) to their documents as well as document identifiers. The second phase is to analyze that table (sequentially or otherwise) and to generate the **secondary answer table** while simultaneously calculating the relevance. Finally, the **partial result table** is created, containing all the documents with a calculated relevance greater than a user defined threshold, ordered by their relevance in a descending order [1].

So, the user has a set of retrieved documents, and can judge their relevance. In order to help the user, we provide some metadata about the documents, as well as some visual representation of the retrieved set.

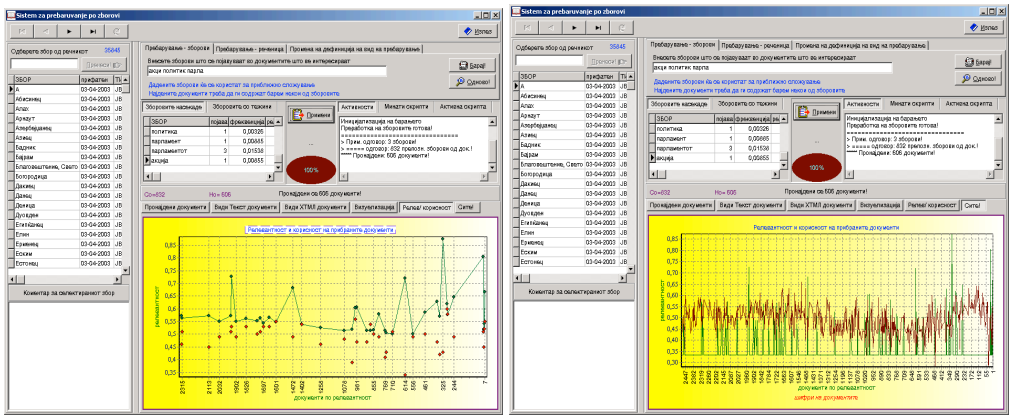

Fig. 1: The retrieving process - a visualisation

The quality of answer and the response time of the system can be affected by several factors. If the table of words within the documents is too big and the request is general (which happens quite often when the user enters his request for the first time), the generation of the primary answer table can take a long time. To address this, the table of words within the documents must be kept as small as possible – the language aspect.

For the analyses of the primary answer table to be as short as possible, we must address the size of the table and the time that is needed for computation of relevancy.

Again, one possibility is to reduce the word forms in the database of identified words, as much as possible.

3 The language aspect – general considerations

The language aspect of the system is very important because it has direct impact on the size of the primary answer table. Namely, if we do not apply any filtering or optimization while accepting the documents into the system, we will end up with a big table with all kinds of words, which will include even misspelled words.

In order to get familiar with the problem, we will point out some language specifics, which can be encountered while processing documents in Macedonian language. For that purpose, we are going to use the word cat and its variations.

English	Macedonian	Experimental separation	Suffix	Comment (potential loss of information)
a cat	machka			
the cat	machkata	machka + ta	ta	definite article
this cat	machkava	machka + va	va	Spatial determination
that cat	machkana	machka + na	na	spatial determination
the cats	machkite *(or)* tie machki	machki + te	te	definite article
these cats	machkive *(or)* ovie machki*	machki + ve	ve (ovie)	spatial determination
those cats	machkine *(or)* onie machki*	machki + ne	ne (onie)	spatial determination

* - same meaning with two words, as a word for word translation of English words

Table 1: An example of word analyses (determining the suffixes)

It can be noted that the definite article in English is a separate word, unlike in Macedonian, where it is an integral part of the word, just like the suffixes for spatial determination of the noun. So, English language uses two words, (in which case getting the basic meaning is easier, because the words are separate), but in Macedonian it takes one word, so the word analysis is needed, meaning that the word should be analyzed and part of it separated. As a result, that word form's occurrence will increase; the information content will be somewhat preserved, on the account of losing definite article or spatial determination (see columns 4 and 5 in Table 1).

The conclusion would be that we can delete some types of word extensions to get the basic meaning, but not always. For example, the following table shows an example of incorrect separation

English	Macedonian	experimental separation	suffix to be deleted	comment
the door	vratata	vrata + ta	ta	correct!
door	vrata	vra + ta	ta	incorrect

Table 2: An example of word analyses (correct and incorrect separation of the possible article)

What about the meaning? For example, the word **machka** means (besides *cat, an animal*), also **to spread** (like in *spreading butter*) or **to paint**, **to colour** (having negative connotation, *painting badly*).

English	Macedonian (basic word form)	various word forms	suffix & change	comment
a cat	machka *(as animal)*	machkata machkava machkana machki machkive machkine machkite	*ta* *va* *na* *i -> a* *ve & i -> a* *ne& i -> a* *te& i -> a*	
to spread the paint	machka *(verb)*	machkan* machkana* machkano* machkani* machkanje* machkanja*	*N* *na* *no* *ni* *nje* *nja*	adjectives - // - - // - - // - a verbal noun - // -
that cat *or* painted on		machkana	*na*	double meaning

These word (*) forms can also have an article or a prefix.

Table 3: An example of different word forms, some may have mixed meaning

Although the situation is much more complicated with prefixes and our system doesn't deal with them, we will give some examples just to make a point.

These examples show that in Macedonian, if you remove the prefixes, the varieties of the word meaning or action defining are lost, and, in some cases, even the tense of the action is changed.

English	Macedonian	experimental separation	prefix	comment
to grease; to bribe	podmachka	pod + machka	pod	all these prefixes are action defining
to spread over	namachka	na + machka	na	
to paint over; to deceive	zamachka	za + machka	za	
to bloat	razmachka	raz + machka	raz	
to smear	izmachka	iz + machka	iz	
to paint over	premachka	pre + machka	pre	
to	premachkuva	pre + machkuva	pre	

Table 4: An example of word analyses (the prefixes in variations of the word machka used as a verb)

The plural also creates problems. In English, the plural is formed by adding 's' or 'es' to the word, or by changing 'y' with 'ies'. The exceptions are groups of nouns whose plural is known as foreign plural. In Macedonian, last letter is changed or a letter 'i' is added, but not always (see rows 4, 5 and 6 in Table 4).

English			Macedonian			comment
singular	plural	change	singular	plural	change	
door	doors	-> s	vrata	vrati	a -> i	regular
cat	cats	-> s	machka	machki	a -> i	regular
lady	ladies	y -> ies	dama	dami	a -> i	regular
city	cities	y -> ies	grad	gradovi	_ -> ovi	special case
stomach	stomachs	-> s	stomak	stomaci	k -> ci	- // -
citizen	citizens	-> s	gragjanin	gragjani	n -> _	- // -

Table 5: An example of creating plural

So, we conclude that if the word contains some specific extensions, while identifying and removing them, we can also discover the meaning. As you can see in Table 3, we have 7 forms with similar meaning and 6 forms with another, and only one word form appears in both groups. So, it looks that the difference in meaning can be established with rather high degree of confidence. But, we don't have a definite answer to this yet and that is where the expert knowledge may be applied.

4 Document processing

Considering various document formats, existing (doc, rtf, txt, html, php, asp, etc) or yet to come, we presume that the document has been prepared for processing (pic-

tures, graphs, and other non-text items removed) and entered into a unified database structure.

When the documents are entered into the system, the processing goes through two phases. First, the document is dissolved into set of words and the system does the basic analyses, which consists of attempting to identify and classify the word, and then counting the occurrences. At this time, the words are classified into four categories: 'r' – dictionary, 's' – stop-word, '?' – un-identified and 'b' – number.

The dictionary category means that the word is identified; the system 'knows' about it and has a strategy of simplifying. The non-identified category means that the system cannot recognize the word. This can be the result of misspelling, or the word can be from different language (very rare, but possible) or simply that the system has not encountered that particular word before.

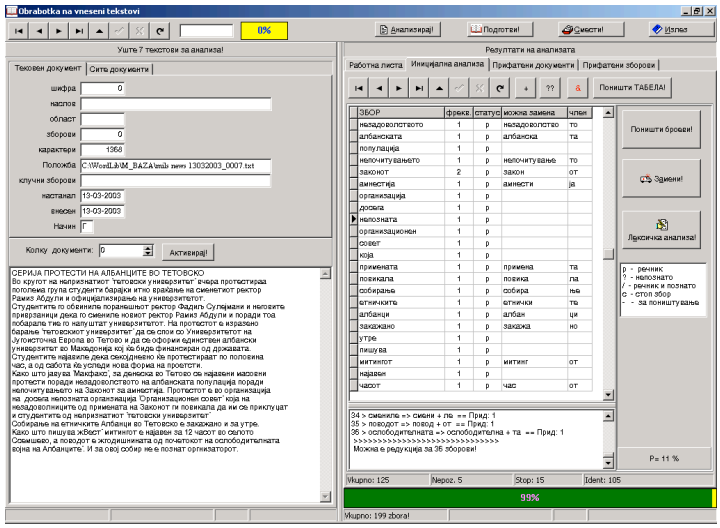

Fig. 2: A phase in document processing

Now, we come to the second phase that is more language dependent (see Figure 3). The system tries to determine which words are predefined as basic word forms and which words can be simplified (even suggesting the possible changes, like removing the suffixes, or changing some letters, as shown in examples before). If the part of the word that remains after being reduced according to certain rules (removing the possible suffixes, for example) is the basic word form and/or a word that system recognizes, than that transformation will be automatically processed. But, if that basic word form is not in the dictionary, the system doesn't know how to treat it.

At this point, the expert intervention is needed, and if the status is changed from "?" to '/', the system will enter the word into its dictionary. If the expert is not present or chooses not to act, the word will be classified as non-identified, and it will await further analyses.The expert, who is overseeing the document processing, decides which changes are allowed, and to what degree. After this, we consider the document pro-

cessed and it is finally entered in the database, i.e. its internal representation is created.

At the beginning, we have used expert knowledge in the first phase of document processing extensively when deciding which words or word forms should enter the internal dictionary, because that dictionary was empty. The dictionary now contains more than 40.000 words, so current experiments have shown almost 90% accuracy of the system suggestions in separating some of the word extensions (ta, va, na, te, ve, ne, to, vo, no) (see Table 1).

5 Conclusion

Some aspects of Macedonian language grammar were discussed and, in order to store parts of the expert knowledge, a kind of semi-automatic system approach to document processing was proposed. The language aspect was addressed through specific examples from the point of expert interaction while processing the documents. It is considered essential to record the expert's decisions and the research in that direction will continue. The aim is enabling the system to perform automatic processing with some degree of confidence without doing the actual context analysis.

The further research will attempt to improve the system considering the language specifics as well as the interaction with the expert (to record the knowledge) and the user (the relevance feedback) as a way of improving the quality of retrieval.

6 References

1. Baeza-Yates, R, Ribeiro-Neto, B, (1999), *Modern information retrieval,* ACM press, New York, Addison-Wesley
2. Korfage, R.R. (1997), *Information Storage and Retrieval*, Wiley Computer Publishing, NY, USA
3. Mizzaro, S. (1997) Relevance: The Whole History, *JASIS, Vol. 48(9), 810–832*
4. Pachovski, V. (2000), *Organizacija na podatoci i relevantnost vo sistemi za pribiranje informacii*, [Data Organization and Relevancy in Information Retrieval Systems], masters thesis, Institute for Informatics, Faculty of Natural Sciences, "St. Cyril and Methodius" University, Skopje
5. Pachovski, V. (2002) A proposition for a method of calculating relevancy in an Information Retrieval System, *Proceedings of VI Balkan Conference on Operational Research,* Thessaloniki, 22-25 May
6. Pachovski, V., Kon-Popovska M. (2002) A proposal for a matching function for an Information Retrieval System, *III International conference on Informatics and Information technologies,* Bitola 13-15 December
7. Pachovski, V., Kon-Popovska M. (2003) A model for an information retrieval system for "small scale" organizations, Proceedings of 1st International Conference On Mathematics in Industry MII 2003*,* Thessaloniki, 14-16 April, 245-253