Proceeding Number: 100/09

Examining the Impacts of Stemming Techniques on Turkish Search Results by Using Search Engine for Turkish

Erdinç UZUN
Corlu Enginering Faculty / Computer Engineering Department / Namık Kemal University
erdincuzun@nku.edu.tr

Abstract. The aim of this paper is to introduce Search Engine for Turkish (SET) Project and examine the effects of the stemming techniques on search performance of Turkish texts by using SET. SET has modules of a crawler, an indexer, a ranking, a searching and an evaluator. The crawler module downloads Milliyet news and automatically creates XML files. The indexer module prepares index to enable a rapid search. The searching module displays the search results that are sorted by the ranking module that has no-stemming, Turkish stemming and word truncation options. Moreover, it uses *tf-idf* to assist. The evaluator module helps to check and assess the search results. 110 queries are prepared by using this evaluator. As expected, analysis indicates that stemming techniques can be used for improving the search results. However, detailed analysis shows that user's suffixes are crucial in some queries so stemming techniques may have negative effects in these queries.

Keywords: *Turkish Web IR, Stemming Methods, Agglutinative Languages*

1 Introduction

The major aim of search engines is actually a question of deciding which documents in a text collection should be obtained from search algorithm to satisfy user's need for information. They use any stemming techniques and various different techniques to provide the best result first. For example, Google heavily relies on links when it comes to determine the ranking of a web site. Moreover, the ranking process varies widely from one engine to another. In addition to these, the stemming technique is used by search engines is not clear. Furthermore, different stemming techniques don't support by commercial search engines. Hence, all possible cases can't be examined in terms of the information retrieval for Turkish. Therefore, a search engine for Turkish (SET[1]) is developed . To the best of our knowledge, this is the first online project for researches on Turkish IR (Information Retrieval). With developing a custom search engine, the effects of stemming techniques can be examined for Turkish IR. In this search engine, no-stemming (NS) is selected to be the baseline; Turkish stemmer (TS) and word truncation (WT) method are used as comparator.

Another matter is the test collection used by search engine. In Turkish, only a very limited number of test collections are available for IR studies. We suffered from similar limitations in creating dataset at our previous studies [1], [2]. Therefore, a Turkish sentence collector, which obtains data over internet, was developed. For

[1]Classes of SET developed in C# programming language are open source and available via the web page http://bilgmuh.nku.edu.tr/SET/. Moreover, you can try all modules by using this web page.

this study, the Turkish sentence collector has been upgraded and developed a crawler that collects news specifically from Milliyet which is an online newspaper. Using this crawler, news of Milliyet stretching from 2004 to 2007 was able to be accessed and created.

The paper is organized as follows: Section 2 introduces properties of the SET projects and gives general information such as information models, stemming techniques, crawling techniques and evaluating techniques that is used in SET. Section 3 examines the results of obtained from an evaluator module. The paper ends with some concluding remarks.

2 Architecture of the SET

SET is a search engine project that is written in C# programming language. This search engine consists of the following fivemodules.

- A Crawler module: collects news of Milliyet between 2004 - 2007 and generates XML documents for every news.
- An index constructor module: parses XML files generated with crawler, tokenize the documents obtained from parsing, do linguistic pre-processing of these tokens and index the files that each terms occurs in.
- A ranking module: sorts the files according to a term weight is namely tf-idf (term frequency-inverse/document frequency) that is often used in information retrieval.
- A searching module: supports to use keywords and phrases in search query. Moreover, it provides the use of Boolean operators (AND, OR and NOT) and parentheses to specify the search query.
- An evaluator module: is a web-based assessment system of either relevant or irrelevant for each query-document pair.

2.1 The Crawler Module

This crawler is designed to download web pages of Milliyet between 2004 and 2007. It firstly prepares the seeds that are a list of URLs (Uniform Resource Locator) to start. For generating these seeds, it produces an appropriate date form. (For example: http://www.milliyet.com.tr/2004/01/08/, http://www.milliyet.com.tr/2005/01/01/, http://www.milliyet.com.tr/2006/03/01/.)

Then, it downloads initial URL, extracts all the hyperlinks in the downloaded page and adds them to the URL Listing. In this extraction, regular expression, which is special text string for describing a search pattern, is used. The following table indicates patterns that are used for extracting URLs.

Table 1. Link Patterns used in this Crawler

Patterns	Description
(?<=href=\").*?(?=\")	Standart HTML <a href> tag
openWindow\('(.*?)'\) popup\('(.*?)'	Links in Javascripts

Before adding an URL into the URL Listing, URL normalization process and unique web page control process are applied to URL for ensuring a duplicate record is not created. After obtaining the URLs, this crawler generates a XML file for every web page. For generating this file, it takes string between comment tags of <!-- PRINT BASLADI --> and <!-- PRINT BİTTİ --> by using regular expression because the rest is made up of banners, menus, advertisements and other unnecessary information. These comment tags didn't use in another years so news between 2004 and 2007 are chosen to use. The generated XML file begin with an information about news article (<DOCNO>, <DATE> and <CATEGORY> tags) and end with a text of a news article (<TEXT> and<P> tags), as shown in Fig. 1. The first paragraph of texts is usually the title of the news articles.

```
<DOC>
<DOCNO> 1103 </DOCNO>
<DATE>2007/08/20</DATE>
<CATEGORY>SPOR</CATEGORY>
<TEXT>
<p>
12 Dev Adam'da rota Efes Cup
</p>
<p>
A Milli Basketbol Takımımız, 3 - 16 Eylül tarihlerinde İspanya'da düzenlenecek olan 2007 Avrupa Basketbol Şampiyonası öncesi son ciddi provasını İzmir'de yapacak. Ay - yıldızlılar, 22 - 26 Ağustos'ta İzmir Halkapınar Salonu'nda gerçekleştirilecek 6. Efes Pilsen World Cup Mini Dünya Kupası'nda boy gösterecek. Yarın sabah bu turnuvaya katılmak üzere İzmir'e gidecek olan millilerimiz, bugün sabah kondisyon ve akşam da taktik olmak üzere iki antrenmanla hazırlıklarını sürdürecek. Milli Takımımız, turnuvada Letonya ve Hırvatistan ile A Grubu'nda yer alırken, B Grubu'nda ise Çin, Sırbistan ve Polonya mücadele verecek.
</p>
</TEXT>
</DOC>
```

Fig. 1. Example of a news article

Test collection that is collected in this study contains 240,364 documents and the size of this collection is around 511 MB in UTF-8. In all documents, this collection contains 94% alphabetic, 4% numeric and 2% alphanumeric characters. Each article contains 309 words (tokens) on the average without stopword elimination. The average length of a word is 6.60 characters.

2.2 The Index Constructor Module

The index constructor produces an inverted file that stores information about words and files to enable a rapid search to be made. Before storing words to the inverted file, the most widely used technique is the stemming that is the process for reducing inflected or derived words to their stem. SET uses no-stemming, stemming of Zemberek[2]and word truncation technique. No-stemming uses original words for indexing. Zemberek is morphological analyzer that can be removed both inflectional and derivational suffixes of words. In *word truncation technique*, first *n* character of each word is its stem and words with n characters are with no truncation. In Turkish IR researches, Sever and Tonta [3] also proposed the use of values 5, 6 and 7 for *n*. However, their proposition is intuitive and based on their observation that truncated and actual Turkish words display similar frequency distributions. Can et al. [4] indicate that *n=5* is appropriate for Turkish IR. Hence,

[2]Zemberek is an open source library that provides basic lexical operation for Turkic languages (http://code.google.com/p/zemberek/). Furthermore, you can try this stemmer over the web using http://bilgmuh.nku.edu.tr/set/semcoz.aspx.

*n=5*is selected in this study. While indexing the test collection with these stemming techniques, they change the size of storage. (See below Table 2.)

Table 2. The indexing information of NS and stemming techniques

Information	NS	TS	WT5
Total no. of terms	864035	364786	156328
Avg. no. of terms/doc.	214	163	148
Avg. term length	9.91	6.25	4.73
Median term length	9	6	5
Min. term length	2	2	2
Max. term length	55	55	5
Total Storage Size (MB)	1068	906	998
Size of Keywords (MB)	29	11	5
Size of Documents (MB)	49	49	49
Size of Relations (MB)	990	846	804

Stemming techniques reduce the size of storage in memory. TS and WT5 cause 15% and 20% of decrease in the storage efficiency when comparing with NS.

2.3 The Ranking Module

The *tf–idf* weight is a statistical measure used to evaluate how important a word is to a document in a text collection. This weight has two main components. First one, the term frequency (*tf*) component, should depend upon the frequency with which a query terms occurs in a given document. The other, the document frequency component, should depend upon how frequency the term occurs in all documents. In fact, inverse document frequency (*idf*), which measures the relative rarity of a term, is utilized in this study. The term weighting formula is usually given as

$$w_{t,d} = tf_{t,d} \times idf_{t,d}$$

where *t* is the weight of a term and *d* is the document vector. After obtaining term weight, the similarity vector is computed by using the formula

$$sim(q,d) = \frac{\sum_t w_{t,d} \times w_{t,q}}{\sqrt{\sum_t w_{t,d}^2} \times \sqrt{\sum_t w_{t,q}^2}}$$

where *q* is the query vector.

2.4 The Searching Module

The search module is a web interface for specifying queries. Then most widely Boolean used retrieval model was used in SET. This means you can use Boolean logic operator (AND, OR, NOT) and parenthesis in searching query. However, users prefer to search “basketball matches of Beşiktaş and Fenerbahçe”, keywords can be used “Beşiktaş OR Fenerbahçe AND Basketbol”. Moreover, synonym words can be added to query. For example, Beşiktaş is nicknamed the “Kara Kartal” (Black Eagles") and Fenerbahçe is nicknamed “Kanarya” (Canary). For

adding these words to the search term, users need parentheses to show the order in which relationships should be considered. In this case, keywords can be used "((Beşiktaş OR (Kara AND Kartal)) and (Fenerbahçe OR Kanarya)) and (Basketball OR Pota)".

2.5 The Evaluator Module

The evaluator module is a web interface for examining search results. In this evaluator, on the precision at 10 documents retrieved (P@10), precision at 20 documents retrieved (P@20), and mean uninterpolated average precision (MAP) are concentrated. P@10 and P@20 are simple and intuitive. Furthermore, these measures closely correlate with user satisfaction in tasks such as web searching and are extremely easy to interpret. However,MAP is based on a much wider set of information than P@10 and P@20. MAP is the mean of the precision scores obtained after each relevant document is retrieved. Geometrically, it is equivalent to the area underneath an uninterpolated recall-precision graph. As MAP is a more reliable measure for IR effectiveness [5], MAP value is used for comparison in this study.

3 Experiments

110 ad hoc queries are prepared to evaluate the effectiveness of stemming techniques. These query terms are divided into three sub groups by using the MAP values of NS and TS.

- 42 negative cases that stemming techniques are ineffective
- 46 positive cases that stemming techniques are effective
- 16 equal cases that the MAP values are equal

The following table gives the results of four example queries for negative cases.

Table 3. Example Query Results for Negative Cases

Query Terms	NS	TS	WT5
üniversiteye giriş sınavı	0.91	0.67	0.64
şampiyonlar ligi	1.00	0.59	0.77
işsizlik sorunu	1.00	0.21	1.00
tarihi eser kaçakçılığı	1.00	0.61	0.57

As these examples indicate, stemming techniques are statistically significantly inefficient in analyzing query terms derived with inflectional and derivational suffixes. In fourth example, the results gathered by TS and WT5 are different because the stem, "iş" (job), is used in TS however WT5 uses the word, "işsiz" (jobless). According to this selection, WT5 is more effective because the word that WT5 uses is closer to the user's original word, "işsizlik" (joblessness), in meaning than the word that TS uses. In terms of count of retrieved search results, these words (" iş", "işsiz", "işsizlik") appear in 59832, 3533 and 1809 documents, respectively. Number of relevant documents gathered by using the stem, "iş", is more than the number of relevant documents gathered by using the words, "işsiz" and "işsizlik". But the relevancy percent of the results is lesser due to the fact that there is 2058 words derived from the stem, "iş", in test collection and these 59832 documents include all possible forms of "iş". In other words, using the stem in search is most effective, if all the words derived from it have similar meanings. Now, the positive cases are examined by analyzing four example queries. (See below Table 4.)

Table 4. Example Query Results for Positive Cases

Query Term	NS	TS	WT5
film festivalleri	0.55	1.00	1.0
Nuri Bilge Ceylan	0.52	0.87	0.82
Türkiye'de futbol şikesi	0.33	0.62	0.76
Kalıtsal hastalıklar	0.15	1.0	1.00

In negative examples taking plural suffix "-lar" out has a negative effect on results, despite the fact that in positive examples taking plural suffix "-ler" has a positive effect on results. In last two examples, NS have no any effect because of query term that has no suffix. NS is efficient when user use suffixes in his query term. In four positive cases, stemming techniques are observed to be adequate to improve search engine results.

Table 5. Comparison of three different cases

Query	NS		TS		Wt5		Case
	MAP	Cnt	MAP	Cnt	MAP	Cnt	Cnt
Positive Cases	0.25	46	0.52	394	0.50	270	48
Equal Cases	0.76	561	0.76	1522	0.77	1193	16
Negative Cases	0.56	233	0.46	1519	0.53	656	46
Overall Results	**0.49**	**166**	**0.53**	**1048**	**0.56**	**493**	**110**

While evaluating the average performance on 110 query terms, stemming techniques seems to be slightly better than the other techniques. Improvements of TS, WT5 are 6.68%, 13.16% better than NS, respectively. The average count of retrieved search results show that no-stemming can be used to narrow the search results when comparing other methods. In the other hand, TS is an appropriate technique for obtaining more search results than the others.

In Turkish, Ekmekcioglu and Willett [6] compare the effectiveness of information retrieval by employing stemmed and non-stemmed query word terms using Turkish news articles of size 6,289 and 50 queries. Sever and Bitirim [7] evaluate the effectiveness of a new stemming algorithm, namely FINDSTEM, which employs inflectional and derivational stemmers. Their algorithm provides 25% retrieval precision improvement with respect to no-stemming. Pembe and Say [8] investigate the question of whether NLP techniques can improve the effectiveness of information retrieval. Can et al. [4] compare the effects of four different stemming options by using a large-scale collection. They used 72 ad hoc queries. All these studies are local. To the best our knowledge, this study is the first online project in Turkish IR.

4 Conclusion

In this study, an online Turkish search engine is described and evaluated the effects of the stemming techniques on Turkish search results obtained from this search engine. Stemming techniques usually gives satisfactory results in IR[9], [10]. Still, this study indicates that there are some cases that stemming is proved to be inadequate. Suffixes that are used in user's query term can be critical in some cases. However, stemming techniques eliminates importance of these suffixes. Regular agglutinative languages such as Turkish encode

more information with suffixes than the other languages. Because of that, not only stems but also suffixes are important in agglutinative language based IR.

Some future research possibilities are modifying this Crawler to obtain different Turkish texts. Another future work would be to try different information retrieval models such as OKAPI, language modeling and mutual information model. Moreover, synonyms of query terms are to be used in searching.

References

1. Uzun, E., Kılıçaslan, Y., H.Agun, V. and Ucar, E., (2008). "Web-based Acquisition of Subcategorization Frames for Turkish", ICAISC 2008, Editors: Rutkowski L. et.al., ISBN 978-83-60434-50-5, pp. 599-607.
2. Uzun, E., (2007). An internet-based automatic learning system supported by information retrieval [İnternet tabanlı bilgi erişimi destekli bir otomatik öğrenme sistemi], Ph.D., Department of Computer Engineering, Trakya University, Edirne, Turkey
3. Sever,H. and Tonta,Y. (2006). Truncation of Content Terms for Turkish, *CICLing 2006*, Mexico City, Mexico.
4. Can, F., Kocberber, S., Balcik, E., Kaynak, C., Ocalan, H. C., Vursavas, O. M., (2008). "Information retrieval on Turkish texts." Journal of the American Society for Information Science and Technology. Vol. 59, No. 3, pp. 407-421.
5. Sanderson, M. andZobel J. Information retrieval system evaluation: Effort, sensitivity, and reliability, *ACM SIGIR' 05*, 2005, pp. 162-169.
6. Ekmekcioglu, F.C. and Willett, P. (2000). Effectiveness of stemming for Turkish text retrieval. Program, 34(2), pp. 195-200.
7. Sever, H. and Bitirimi Y. (2003). FindStem: Analysis and evaluation of a Turkish stemming algorithm. Lecture Notes in Computer Science, 2857, pp. 238-251.
8. Pembe F. C., and Say ACC, (2004). A linguistically motivated information retrieval system for Turkish. Lecture Notes in Computer Science, 3280, pp. 741-750.
9. Krovetz,R. (1993). Viewing morphology as an inference process, *ACM SIGIR'93*, Pittsburgh: ACM, 1993, pp. 191-202.
10. Savoy, J. (2006). Light stemming approaches for the French, Portuguese, German and Hungarian languages, *ACM SAC' 06*, pp. 1031-1035.

Biography

Erdinç Uzun – received his BSc (2001), MSc (2003) and PhD (2007) degrees from Computer Engineering Department of Trakya University respectively. He was a research assistant in this university during this time. Then, he has begun to work in Computer Engineering Department of Namik Kemal University as assistant professor. He was the vice dean between 2007 and 2010 Faculty of Corlu Engineering. Since 2010, he is the vice chairman of Computer Engineering Department of this faculty. His research interests include information retrieval, machine learning, data mining, web mining and natural language processing. He is the developer of SET (Search Engine for Turkish - bilgmuh.nku.edu.tr/SET/) and Teaching (Lesson Content Management System - bilgmuh.nku.edu.tr/Teaching/)