Failed Queries: a Morpho-Syntactic Analysis Based on Transaction Log Files

Anna Mastora[1], Maria Monopoli[2] and Sarantos Kapidakis[1]
[1] Laboratory on Digital Libraries and Electronic Publishing, Department of Archives and Library Sciences, Ionian University, 72, Ioannou Theotoki str., GR-49100, Corfu, Greece.
[2] Library Section, Economic Research Department, Bank of Greece, 21, El. Venizelos Ave., Athens, GR-10250
{mastora, sarantos}@ionio.gr, mmonopoli@bankofgreece.gr

Abstract. The aim of the study is to elaborate on the procedure needed in order to analyze morpho-syntactically the typing-error queries submitted in Greek during the search process. In the context of our analysis a *failed query* is a query which returned no hits. The analysis showed that failed queries represent 36% of the submitted queries. More specifically, 19.6% of failed queries occurred due to typing errors. We discovered that for analyzing morpho-syntactically a Greek text corpus the PoS tools need to be rich in tags in order to work adequately. Open Xerox tokenizer performed well but with significant pre-processing of the queries and the analyzer seems to require additional tools to improve its performance. MS Word which was used for spelling corrections seems to perform satisfactorily. All tools were challenged in terms of named entities recognition.

Keywords: Failed queries, Morpho-syntactic analysis, PoS tagging, Typing errors

1 Introduction

Information retrieval techniques do not work effectively at all times. *Not working effectively* includes both not retrieving relevant documents, i.e. low recall, and retrieving non relevant documents, i.e. low precision. Part of studying what is not retrieved during an information search process is the analysis of *failed queries* or *failure analysis*. This is also the motivation of our study with respect to Natural Language Processing (NLP) techniques.

In this study we explore the failed queries caused due to typing errors. The grouped queries are analyzed morpho-syntactically in order to develop a clear image of the required process before stepping to the next phases of the data analysis in the future.

2 Aims and Objectives

The aim of the study is to elaborate on the procedure needed in order to analyze morpho-syntactically the typing-error queries submitted in Greek during the search process.

The objectives of the study are twofold. First, we explore the extent and types of failed queries due to typing errors. Second, we explore the procedure and feasibility of their morpho-syntactic analysis.

3 Related Research

The discussion concerning what constitutes a *failed query* is extensive [1, 2, 3] Different perspectives of search failures are presented. Some researchers consider *failure* in terms of precision and recall applying retrieval effectiveness measures. Others examine *failure* in terms of user satisfaction applying users' criteria to measure whether a query failed or not. Others use transaction log files and treat input terms either as "bag of words" or apply relevance feedback and assign more interpretations to the result set. Finally, there are techniques which study the human behavior by observation.

Significant interest has been expressed on failed queries as the outcome of subject searching [4, 5]. This strategy has been identified as the most common for delivering failed queries due to various reasons but mostly because of the inherent difficulty of matching the index terms to the users' queries. This identified difficulty and the documented analysis [6] which supports that for information needs related to environmental issues users tend to perform subject searching explain the focus of our study on subject searching.

A considerable aspect of the research on failed queries is the techniques used for Natural Language Processing. These techniques are essential especially in highly inflectional languages [7] such as the Greek language. While the main goal at all times is to assign the proper semantic information to each query, this cannot be accomplished without prior identification of the morho-syntactic information of the terms used. The techniques applied for this purpose are the Part of Speech (PoS) tagging which is accompanied by more detailed morpho-syntactic information (see Fig.2 for an example).

4 Definitions and Methodology

In this section we provide the definitions of the terminology used in our study as well as the analysis on the methodology used.

4.1 Definitions

Through the study of related research, as presented in the previous section, what becomes obvious is that failed queries constitute a disputable area concerning the very definition of what actually should be considered as a *failed query*.

In the context of our analysis a *failed query* is a query which returned no hits. We took into consideration the objections on the issue yet we support this decision by the fact that the analysis of the data was based on terms extracted from transaction log files without any relevance feedback from the users' perspective. This is also why we proceeded with a morpho-syntactic analysis leaving for later phases the processes related to word-sense disambiguation. An additional factor which strengthens our decision is that both the content of the database and the information needs belonged to the same domain and it was expected that most queries would return hits.

The *morpho-syntactic analysis* of the data is a cognitive process that constitutes an intermediate layer between morphological and syntactic analysis and aims to assign unambiguous morpho-syntactic information to words of texts [8].

The *morpho-syntactic information* consists of the morphological origin and the morpho-syntactic properties of a word. For example, the word *ανθρώπου* is the genitive singular form of the masculine noun *άνθρωπος* [8].

Inflectional languages are the languages with a high morpheme-per-word ratio whereas the *morpheme* is the smallest meaningful linguistic unit. The Greek language is considered a highly inflectional language.

More definitions on terminology used across this paper can be found in the corresponding sections.

4.2 Methodology

The data analyzed in this paper was gathered from an in vitro experiment with the participation of 27 undergraduate students at the Department of Archives and Library Sciences at the Ionian University in Corfu. They were given 13 information needs related to environmental issues and asked to submit appropriate queries in order to retrieve relevant documents. The database they were searching in contained material mainly from the environmental domain.

For the purpose of this experiment we selected and customized approximately 14,400 bibliographic records of the Evonymos Ecological Library[1]. The queries were submitted in Greek as well as the records contained information only in Greek. This is a significant factor when analyzing data in the context of Natural Language Processing because it eliminates the possibilities of arbitrarily assigning characteristics to words due to the intervening stage of their translation.

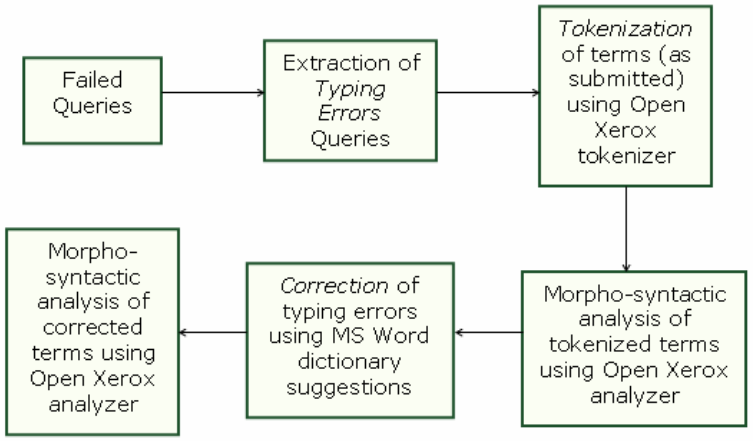

Fig. 1. Synopsis of the procedures' workflow during the processing of the data.

The participants could search only in the *Subject* field. According to Jones et al. [2] users rarely change default settings. This observation suggests that the customization of the interface did not record either an unrealistic or biased users' behavior. The transaction log files kept in a Zclient consist of one xml document per user per session. All participants logged in the system using their matriculation numbers thus making it easier to potentially relocate them for providing feedback at a later stage of the research.

Concerning the processing of the data, the first step involved the selection of *failed queries* and, more specifically, the selection of *typing error queries*. The next step involved the tokenization of the selected corpus of queries and then their morpho-syntactic analysis. Following was the processing of correcting the spelling errors of the tokens and running from scratch the analyzer. Figure 1 above visualizes the workflow of the data processing while Figure 2 below gives an example of the processed data.

[1] Full database available at http://www.evonymos.org/index.html (last accessed 17 April 2011).

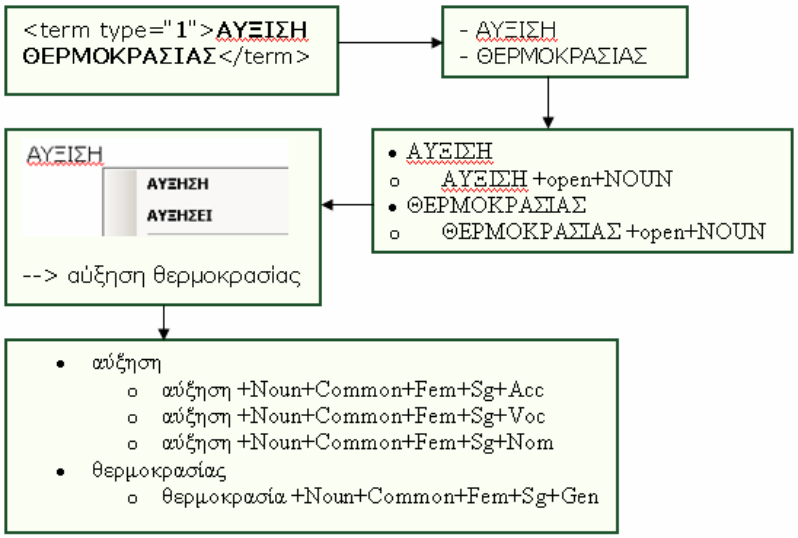

Fig. 2. Example of the processing of the data.

5 Results

This section presents the findings of our study. There were 1,284 queries submitted overall, while 459 of them were failed queries, i.e. 36%, meaning that they returned no hits. Consistent to our previous work [9], we further categorized those failed queries to four subcategories, namely *Valid terms with no hits*, *Typing errors*, *Inseparable terms* and *Undefined terms.*

The failed queries subcategory named *Valid terms with no hits* is the most populated one with a ratio of 75.8% and includes terms which were valid both morphologically and syntactically yet they did not deliver any hits. The second subcategory, i.e. *Typing errors*, comes next in delivering failed queries with a ratio of 19.6%. Third appears the subcategory containing the words which were not separated during typing. They represent a ratio of 2.4%. And, finally, the last subcategory includes some undefined terms, meaning words that do not appear in official dictionaries, in 2.2% of the failed queries overall.

Since the focus of this study is on *Typing error* queries, we analyzed them further by dividing them, based on previous work [9], to five new subcategories, namely *Substitutions*, *Transpositions*, *Omissions*, *Insertions* and *Divisions*. *Substitutions* include the changing of a letter with another letter, like in the case of typing ***φ***εωθερμική instead of the correct ***γ***εωθερμική. *Transpositions* include the cases where one or more characters within a word do not appear in the right order, for example *αν****ιτ***ρρήσεις instead of the correct *αντ****ι***ρρήσεις. *Omissions* include the cases where one or more characters within a word are missing, for example *οπω****φ***όρα δέντρα instead of the correct *οπω****ρο***φόρα δέντρα. *Insertions* include the cases where one or more characters are added within a word, as in the case of *μεσσ*όγειος instead of the correct *με****σ***όγειος. Finally, the last subcategory of typing error queries is *Divisions*, including splitting terms which should appear as one. Table 1 right below shows the distribution percentage of each subcategory.

Table 1. Categorization of failed queries due to typing errors (percentage, %).

Substitutions	Transpositions	Omissions	Insertions	Divisions
36.7	4.4	28.9	28.9	1.1

At this point we remind that the total number of failed queries was 459 out of 1,284 submitted queries. Ninety (90) of the failed queries were due to typing errors. In order to proceed to the morpho-syntactic analysis we had to identify the tokens to analyze. For this purpose we uploaded the terms to the Open Xerox Tokenizer[2]. The outcome of this process was 156 tokens.

The following step was to explore whether the Open Xerox analyzer[3] would directly identify the misspelled tokens during the morpho-syntactic analysis. As shown in Table 2, the tool did not manage to recognize the misspelled tokens, thus, performing poorly since it only managed to identify 20 out of 156 tokens.

Table 2. Categorization of identified tokens when analyzed as submitted (exact numbers).

Regular words	Punctuation	Pronouns	Prepositions	Others
10	5	3	1	1

In order to overcome the barrier of this poor performance we proceeded with the correction of the identified tokens using the spelling suggestions of the MS Word's default dictionary. During this stage, since the data was processed manually, we interfered with the results by assigning the semantically correct suggestion to each token. Table 3 below shows the performance of the MS Word dictionary.

Table 3. Categorization of MS Word correction suggestions.

Action	**Percentage (%)**	**Actual number**
No suggestion required	30.1	47
No suggestion provided	12.8	20
Irrelevant suggestion	3.2	5
MS Word's 1st suggestion=correct	45.5	71
MS Word's 2nd suggestion=correct	7.1	11
MS Word's 3rd suggestion=correct	1.3	2
Total	100	156

As shown in Table 3, for approximately 30% of the cases no suggestion was required. This includes the tokens which did not contain any typing error. Their assignment to typing error queries was due to the fact that they belonged to multi-word terms in which at least one typing error was identified. After the tokenization stage, these tokens were isolated from the original term and when processed during the next stage, that is the stage of typing errors' correction, no intervention was required. Punctuation was also included in this category.

After having corrected the originally identified tokens, we proceeded with the morpho-syntactic analyzer anew. This time it performed significantly better identifying 139 out of 156 tokens. Table 4 below shows a categorization of the missed identifications. We need to mention at this point that in the documentation for the Part of Speech tag set for Greek it is mentioned that the analyzer identifies words in other languages and tags them as *+FM*, i.e. Foreign Words[4]. We observed an inconsistency concerning this feature since words in English included in our corpus were not identified as expected. Instead they were rather arbitrarily assigned a general tag, like *noun*.

[2] Available at http://open.xerox.com/Services/fst-nlp-tools/Consume/175 (last accessed 17 April 2011).

[3] Available at http://open.xerox.com/Services/fst-nlp-tools/Consume/176 (last accessed 17 April 2011).

[4] The full Part of Speech (PoS) tag set for Greek is available here http://open.xerox.com/Services/fst-nlp-tools/Pages/Greek%20Part-of-Speech%20Tagset (last accessed 17 April 2011).

Table 4. Categorization of corrected tokens not recognized during the morpho-syntactic analysis.

Category of the token	Percentage (%)	Actual number
Named entities	17.6	3
Regular words	35.3	6
Truncated words	29.4	5
English words	11.8	2
Punctuation	5.9	1
Total	100	17

6 Conclusions

The analysis of *failed queries* shows that they represent 36% of the submitted queries overall in our experiment. More specifically, 19.6% of failed queries are due to typing errors. During Natural Language Processing the queries which contain typing errors require more steps and extra mechanisms involved in order to achieve a trustworthy and effective morpho-syntactic analysis. This is both a practical and a substantial problem to solve considering their proportion within the overall submitted queries.

In the process of data analysis we discovered that the tools for morpho-syntactic analysis for the Greek language need to be rich in tags in order to work adequately. Since the Greek language is a highly inflectional language it requires the combination of more mechanisms, such as dictionaries, discovering synsets etc., for proper analysis. This practice affects the complexity of the tools used but it seems inevitable. Such tools should aim at making the less possible suggestions for each segment and that the suggestion is as close as it gets to the *true* sense of the segment, where by *true* is meant the sense which the user intended.

Transaction log files serve as good starting points for processing the data quantitatively but more measures need to be applied in order to extract adequate qualitative information for the terms used in submitted queries.

Concerning the tools we used for the analysis of our data we observed important deficiencies which complicated the process. First, we observed that in order for the Open Xerox tokenizer to work properly all input words should be lower case and stress marked. This caused extra load of work because we had to convert the words submitted in capitalized form and stress them. Additionally, we had to implement this step to all the words that were originally in lower case but had no stress mark as well.

Another challenge of the tools used was that they did not recognize *named entities*. This covers a whole separate field of research but within our dataset the use of named entities was not extensive and did not severely affect the outcome. In other cases, however, this could play a significant role.

7 Future Work

Future planning concerning this work includes research on *named entities recognition, language identification* and *word-sense disambiguation* in order to achieve higher rates of morpho-syntactic analysis. All three aforementioned areas are important in terms of analyzing the input of the user and delivering better results.

Another aspect of future research on this area is the exploration of how and to what extent could we incorporate Knowledge Organization Systems (KOS) to query expansion techniques in terms of improving the retrieved result set in cases of prior failed queries.

Acknowledgement: This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

References

1. Tonta, Y., 1992. Analysis of Search Failures in Document Retrieval Systems: A Review. *Public-Access Computer Systems Review*, 3(1), pp. 4-53.
2. Jones, S. et al., 2000. A Transaction Log Analysis of a Digital Library. *International Journal on Digital Libraries*, 3, pp. 152-169.
3. Pu, H.-T., 2008. An analysis of failed queries for web image retrieval. *Journal of Information Science*, 34(3), p.275–289.
4. Lau, E.P. & Goh, D.H.-L., 2006. In search of query patterns: a case study of a university OPAC. *Information Processing and Management: an International Journal*, 42(5), pp. 1316–1329.
5. Villén-Rueda, L. et al. 2007. The Use of OPAC in a Large Academic Library: A Transactional Log Analysis Study of Subject Searching. *The Journal of Academic Librarianship*, 33(3), pp. 327-337.
6. Nicholas, D. et al., 2008. User diversity: as demonstrated by deep log analysis. *The Electronic Library*, 26(1), pp. 21-38.
7. Acedański, S., 2010. A morphosyntactic Brill Tagger for inflectional languages. In *Proceedings of the 7th international conference on Advances in natural language processing*. IceTAL'10. Berlin, Heidelberg: Springer-Verlag, pp. 3–14.
8. Orphanos, G., 2000. *Computational morphosyntactic analysis of modern Greek*. Unpublished PhD thesis. Patras: University of Patras. School of engineering. Department of computer engineering and Informatics.
9. Mastora, A. et al., 2007. Exploring users' online search behaviour: a preliminary study in a library collection, 2nd DELOS Conference on Digital Libraries, Pisa, Italy, December 5-7.