Linguistic features to predict query difficulty - a case study on previous TREC campaigns

Josiane Mothe
IRIT – University of Toulouse 3 / CNRS
mothe@irit.fr

Ludovic Tanguy
ERSS – University of Toulouse 2 / CNRS
tanguy@univ-tlse2.fr

Categories and Subject Descriptors
H.3 [**Information storage and retrieval**]

Keywords
Difficult query, linguistic features, TREC

ABSTRACT

Query difficulty can be linked to a number of causes. Some of these causes can be related to the query expression itself, and can therefore be detected through a linguistic analysis of the query text. Using 16 different linguistic features, automatically computed on TREC queries, we looked for significant correlations between these features and the average recall and precision scores obtained by systems. Each of these features can be viewed as a clue to a linguistically-specific characteristic, either morphological, syntactical or semantic. Two of these features (syntactic links span and polysemy value) are shown to have a significant impact on either recall or precision scores for previous adhoc TREC campaigns. Although the correlation values are not very high, they indicate a promising link between some linguistic characteristics and query difficulty.

1. CONTEXT

This study has been conducted in the context of the ARIEL research project, in which we investigate the impact of linguistic processing in IR systems. The ultimate objective is to build an adaptive IR system, in which several natural language processing (NLP) techniques are available, but are selectively used for a given query, depending on the predicted efficiency of each technique.

2. OBJECTIVE

Although linguistics and NLP have been viewed as natural solutions for IR, the overall efficiency of the techniques used in IR systems is doubtful at best. From fine-grained morphological analysis to query expansion based on semantic word classes, the use of linguistically-sound techniques and resources has often been proven to be as efficient as other cruder techniques [5] [8]. In this paper, we consider linguistics as a way to predict query difficulty rather than a means to model IR.

3. RELATED WORK

A closely-related approach is the analysis performed by [7] on the CLEF topics. Their intent was to discover if some query features could be correlated to system performance and thus indicate a kind of bias in this evaluation campaign, and further to build a fusion-based IR engine. The linguistic features they used to describe each topic mostly concerned syntactic and word forms aspects, and were calculated by hand. They used a correlation measure between these features and the average precision, but the only significant result was a correlation of 0.4 between the number of proper nouns and average precision. Further studies led the authors to named entities as a useful feature, and they were able to propose a fusion-based model that improved overall precision after a classification of topics according to the number of named entities. The precision increase using this feature varied from 0 to 10%, across several tasks (mono- and multi-lingual). Our study deals with more linguistic features, especially in order to deal with syntactic complexity. In addition, we only used automatic analysis methods with NLP techniques.
Focusing on documents instead of queries, [6] also used linguistic features in order to characterize documents in IR collections. His main point was to study the notion of relevance, and test whether it could be related to stylistic features, and if the genre of a document could be useful for relevant document selection.

[3] also used documents in order to predict query difficulty using a clarity score that depends on both the query and target collection. Both the previous studies therefore need to have exhaustive information on the collection; while we decided to focus on queries only, in order to deal with a wider range of IR situations.

In [2] several classes of topic failures were drawn manually, but no elements were given on how to assign automatically a topic to a category.

4. METHOD

We selected the following data: TREC 3, 5, 6 and 7 results for the adhoc task; that corresponds to a total of 200 queries (50 per year). Each query in these collections was automatically analysed and described with 16 variables, each corresponding to a specific linguistic feature. We considered the title part of the query as its length and format is the closest to a real user's query. Because TREC web site makes participants' runs available (i.e. lists of retrieved documents for each query), it was possible to compute the average recall and precision scores for each run and each query (using the trec-eval utility). We then computed the average recall and precision values over runs for each query. Finally, we computed the correlation between these scores and the linguistic features variables. These correlation values were tested for statistical significance.

As a first result, if simple features dealing with the number or size of words in a query or the presence of certain parts of speech do not have clear consequences on a query's difficulty, more sophisticated variables led to interesting results. Globally, the syntactic complexity of a query has a negative impact on the precision scores, and the semantic ambiguity of the query words has a negative impact on the recall scores. A little less significantly, the morphological complexity of words also has a negative effect on recall.

4.1. Linguistic Features

The use of linguistic features in order to study a document is a well-known technique. It has been thoroughly used in several NLP tasks, ranging from classification to genre analysis. The principles are quite simple: the text (i.e. query in our case) is first analysed using some generic parsing techniques (e.g. part of speech tagging, chunking, and parsing). Based on the tagged text data, simple programs compute the corresponding information. We used:

- Tree Tagger[1] for part-of-speech tagging and lemmatisation: this tool attributes a single

[1] *TreeTagger*, by H. Schmidt; available at www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

morphosyntactic category to each word in the input text, based on a general lexicon and a language model;

- Syntex [4] for shallow parsing (syntactic link detection): this analyser identifies syntactic relation between words in a sentence, based on grammatical rules;

In addition, we used the following resources:

- WordNet 1.6 semantic network to compute semantic ambiguity: this database provides, among other information, the possible meanings for a given word;
- CELEX[2] database for derivational morphology: this resource gives the morphological decomposition of a given word.

According to the final objective, which is an automatic classification of queries, all the features considered are computed without any human intervention, and are as such prone to processing errors.

The 16 linguistic features we computed are in Table 1, categorized in three different classes according to their level of linguistic analysis:

Table 1: List of linguistic features

Morphological features :	
- number (#) of words	NBWORDS
- average word length	LENGTH
- average # of morphemes per word	MORPH
- average # of suffixed tokens word	SUFFIX
- average # of proper nouns	PN
- average # of acronyms	ACRO
- average # of numeral values (dates, quantities, etc.)	NUM
- average # of unknown tokens	UNKNOWN
Syntactical features :	
- average # of conjunctions	CONJ
- average # of prepositions	PREP
- average # of personal pronouns	PP
- average syntactic depth	SYNTDEPTH
- average syntactic links span	SYNTDIST
Semantic feature :	
- average polysemy value	SYNSETS

- *Word length* is the average length of terms in the query, measured in numbers of characters.

- The *number of morphemes* per word is obtained using the CELEX morphological database, which describes, for around 40,000 lemmas, their morphological construction. For example, ”additionally” is a 4-morpheme word (“add+ition+al+ly”). Heavily constructed words are known to be more difficult to match with morphologically similar words, thus requiring specific rules, often more complicated than the Porter algorithm. The limit of this method is of course the database coverage, which leaves rare, new, or misspelled words as mono-morphemic.

- The *number of suffixed* tokens is a more general method, which can lead to consistent results with any word form. We used a bootstrapping method in order to extract the most frequent suffixes from the CELEX database, and then tested for each lemma in the topic if it was eligible for a suffix from this list.

- The *number of proper nouns* was obtained through the POS tagger's analysis, and with a more robust method based on upper-case word forms.

- *Acronyms* and *numerals* are detected using a simple pattern-matching technique.

- *Unknown words* are those marked up as such by the POS tagger (i.e. that are absent from its reference wordlist), excluding proper nouns, acronyms and badly-segmented forms. Most unknown words are constructed words such as “mainstreaming”, “postmenopausal” or “multilingualism”.

- *Conjunctions, prepositions and pronouns* were detected through POS tagging only.

- *Syntactic depth and syntactic links span* are computed from the results of the syntactic analyzer. Syntactic depth is a straightforward measure of syntactic complexity in terms of hierarchy. It simply corresponds to the maximum number of nested syntactic constituents in the query. Figure 1 displays the syntactic tree for Topic 158 (TREC3) : “Term limitations for members of the US congress”, which has a maximum depth of 5.

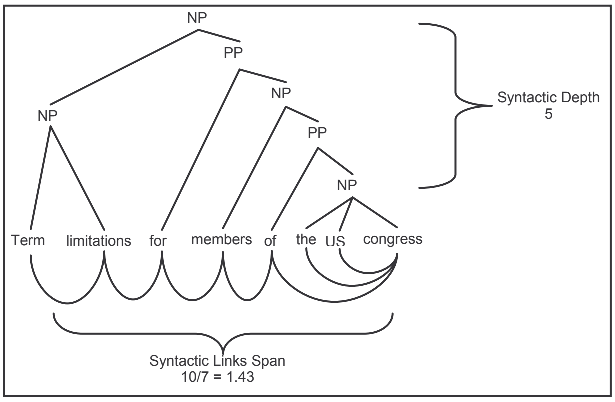
However, the “horizontal” analysis of this structure is quite straightforward, as each of the five nouns is linked to its immediate neighbour (e.g. term -> limitations, limitations -> for, for -> members, U.S.-> congress, etc.).

The measure that we used to take this into account is the Syntactic Links Span: taking each individual syntactic link identified in the sentence, we simply compute its distance in terms of number of words (1 for adjacent words, etc.). We then average this value over all syntactic links.

Figure 1 presents the results of a syntactic analysis for topic #158. The upper half of this figure shows the syntactic tree associated to the title: it identifies the nested Noun Phrases (NP) and Prepositional Phrases (PP), and therefore gives the syntactic depth. The lower half shows the syntactic dependency links between pairs of words. For example, in the phrase ‘the US congress’, the determiner ‘the’ is related to ‘congress’ and not ‘US’. Therefore, this particular link covers a distance of 2 words. By adding all distances, and then dividing by the number of identified links, we get the syntactic links span value for this sentence (10 / 7 = 1.43).

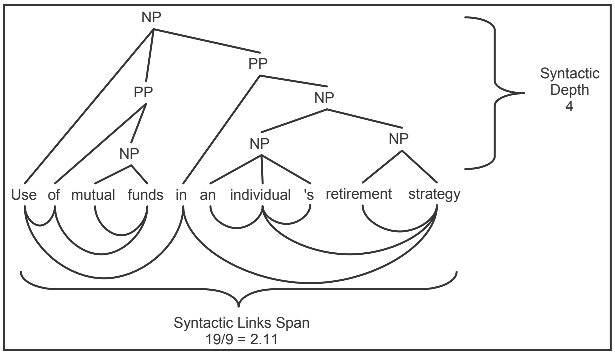
Figure 1: Syntactic Depth and Syntactic Links Span for Topic 158

[2] *CELEX English database* (1993). Available at www.mpi.nl/world/celex



The situation is different for the following topic (#171, TREC 3) "Use of Mutual Funds in an Individual's Retirement Strategy", as seen in figure 2.

Figure 2: Syntactic Depth and Syntactic Links Span for Topic 171



This topic has a syntactic depth of 4, i.e. less than topic #153. However, its syntactic links span is much higher, thus presenting a different kind of syntactic complexity.

Syntactic links span is a better representation of words being separated from each other, even for sentences with a similar tree structure.

- *The average polysemy value* corresponds to the number of synsets in the WordNet database each word belongs to. This value is directly available in WordNet, and roughly corresponds to the different meanings a given word can have. Once again, the database coverage is a limit to this method, but it is a safe assumption to say that rare or new words are monosemous, so the default value of 1 used for words absent from WordNet is a good approximation.

5. ANALYSIS

As mentioned above, we computed correlation scores between these features and the average recall and precision scores for these queries, separately for each TREC campaign. We used Pearson's product moment correlation, which corresponds to the covariance of the two considered variables divided by their standard deviation, thus ranging from -1 to +1. Higher absolute values indicate a stronger correlation; positive values indicate a positive correlation, i.e. that the higher the value for the variable, the higher the recall/precision score. Negative value indicates a relation in the other direction. Significance in the correlations value is expressed by the associated p-value. P-value is an estimation of the probability that results as extreme or more extreme occur by chance. A p-value of 0 indicates a high confidence in the correlation, while a high value indicates a high chance for independence between the variables.

The following table gives, for each TREC campaign, the significantly correlated variables for both average recall and precision scores; a minus sign in front of the variable indicates a negative correlation. Significant correlations are those for which p-value is less than 0.05. These measures were obtained using the SPSS software.

Table 2 : Significant correlations between linguistic features and recall / precision

TREC Campaign	*Significant variables for Recall*	*Significant variables for Precision*
TREC 3	- PREP - SYNTDEPTH **- SYNSETS**	- SUFFIX - NBWORDS - CONJ
TREC 5		**- SYNTDIST** - SYNTDEPTH
TREC 6	**- SYNSETS** + PN	
TREC 7	**- SYNSETS**	+ PN - LENGTH **- SYNTDIST**

As can be seen in the above table:

- the only positively correlated feature is the number of proper nouns. The same conclusion was obtained by [7] on CLEF topics;

- many variables do not have significant impact on any evaluation measure. Only the more "sophisticated" features appear more than once;

- the only two variables appearing more than once with the same sign in the same column are SYNTDIST for precision and SYNSETS for recall. The following tables give the detailed results.

Table 3 : Correlation and significance values between SYNTDIST and Precision

TREC Campaign	*Correlation (Pearson)*	*Significance (p-value)*
TREC 3	-0.224	0.117
TREC 5	-0.396	0.000
TREC 6	+0.091	0.528
TREC 7	-0.234	0.047

Table 4 : Correlation and significance values between SYNSETS and Recall

TREC Campaign	*Correlation (Pearson)*	*Significance (p-value)*
TREC 3	-0.302	0.033
TREC 5	-0.053	0.714
TREC 6	-0.354	0.012
TREC 7	-0.284	0.045

As can be seen in these figures, correlations are significantly negative for 3 out of 4 TREC campaigns. The non-significant cases, however, are very close to independence (high score for significance).

The main result of this study is therefore that semantic ambiguity and “horizontal” syntactic complexity are good indicators of query difficulty.

Possible explanations vary depending on the techniques used by IR systems. A high SYNTDIST is an obstacle to the identification of significant collocates (thus lowering precision), while a high SYNSETS indicates polysemous words that can lead to unrelated documents, thus lowering recall.

Other experiments have been conducted using the same method, but examining each run independently, instead of using the average measures for recall and precision over all the systems. It appeared that, for both selected features, correlations were very close from one system to another. For other features, however, sensitivity to linguistic phenomena differs widely. Most notably morphological features (especially SUFFIX) can lead to varying level of correlation, supposedly due to the difference in terms of morphological processing (stemming methods), while having an overall negative impact.

Another interesting track is the possibility to automatically reformulate queries diagnosed as complex using these features. If semantic disambiguation has already been investigated in IR, reduction of syntactic complexity has yet to be studied.

6. CONCLUSION

This study presents a closer look at the correlation between a query difficulty (as shown by the average scores obtained by IR systems in TREC campaigns) and some linguistic features of the query itself. We have shown that the most significant features are syntactic complexity (in terms of distance between syntactically linked words) and word polysemy (in terms of number of semantic classes a given word belongs to). The results we obtained are promising clues towards an adaptive IR system, as well as towards new specific techniques. An example work in progress following these results is to use different word stemming techniques depending on the number of suffixed words, to add a semantic disambiguation module when dealing with highly polysemous words, or to change the word order of syntactically complex sentences, while doing simpler (and less error-prone) processing for "simple" queries.

7. REFERENCES

[1] Biber, D. (1988). Variation *across speech and writing*. Cambridge: Cambridge University Press.

[2] Buckley, C and Harman, D. (2004) *Reliable Information Access* Final Workshop Report. http://nrrc.mitre.org/NRRC/Docs_Data/RIA_2003/ria_final.pdf

[3] Cronen-Townsend S., Zhou Y. and Croft, W.B. (2002) *Predicting Query Performance*, in proceedings of the 25th annual international ACM-SIGIR conference on research and development in information retrieval, pp. 299-306, Tampere..

[4] Fabre, C. and Bourigault D. (2001). *Linguistic clues for corpus-based acquisition of lexical dependencies*, in proceeding of Corpus Linguistics, Lancaster.

[5] Harman, D. (2000). *What we have learned and have not learned from TREC*. Paper presented at the British Computer Society Information Retrieval Special Group 22nd Annual Colloquium on IR Research, Cambridge.

[6] Karlgren, J. (1999). *Stylistic Experiments in Information Retrieval*, in Natural Language Information Retrieval, Kluwer.

[7] Mandl, T. Womser-Hacker, C. (2002) *Linguistic and Statistical Analysis of the CLEF Topics*, CLEF Workshop.

[8] Sparck Jones, K., Galliers, J.R., (1996). *Evaluating natural language processing systems*. Berlin: Springer-Verlag, Lecture Notes in Artificial Intelligence 1083.