Linguistic features to predict query difficulty

Josiane Mothe, Ludovic Tanguy

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. Three of these features are shown to have a significant impact on either recall or precision scores for previous adhoc TREC campaigns. Each of these features can be viewed as a clue to a linguistically-specific characteristic, either morphological, syntactical or semantic. These results also open the way for a more enlightened use of linguistic processing in IR systems.

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 [4] [7]. 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 [6] 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 were more or less the same ones we used for this experiment (see details below), and were obtained manually. 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 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).

Focusing on documents instead of queries, [5] 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.

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 (the runs for TREC 1,2 and 4 were not available) ; 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 consider the title part of the query as its length and format is the closest to a real user's query. We then computed the average recall and precision scores for the different runs that were submitted to the corresponding TREC task, and 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 have 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,
- Syntex [3] for shallow parsing (syntactic link detection),

In addition, we used the following resources:

- WordNet 1.6 semantic network to compute semantic ambiguity
- CELEX[2] database for derivational morphology.

According to the final objective, which is an automatic classification of queries, all the features considered are

1 *TreeTagger*, by H. Schmidt; available at www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
2 *CELEX English database* (1993). Available at www.mpi.nl/world/celex

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 a rough measure of word complexity, and is calculated in terms 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. For example, the topic 153, (TREC 3) "Term limitations for members of the U.S. congress" has a syntactic depth of 5, because of the embedded noun-phrase structure partly due to the two prepositions. 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 -> limits, limits -> for, for -> members, U.S.-> congress, etc.). This sentence therefore has an average syntactic link span of 1. From another angle, this structure does not imply that highly correlated words are distant from one another in terms of words.
The situation is different for the following topic (#171, TREC 3)
"Use of Mutual Funds in an Individual's Retirement Strategy"
Its syntactic depth is the same as the previous example, as it has roughly the same syntactic structure in terms of noun phrases and prepositions. However, this time the relations between words are more distant, specially "use -> funds" "in -> strategy", leading to an average syntactic link distance of 2.15, or more than twice that obtained for topic 153).

- *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. Correlation is a simple statistical measure, 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 is an estimation of the probability of the correlation being due to random. A significance 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. The significance level is 0.05, using Pearson's measure.

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 -CC
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 [6] 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*
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*
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.

Other experiments have been conducted using the same method, but examining each run independently, instead of using the average measures for recall and precision. 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.

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

[4] 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.

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

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

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