Linguistic Analysis of Users' Queries: towards an adaptive Information Retrieval System

Josiane Mothe(1)(3), Ludovic Tanguy(2)
(1) Institut de Recherche en Informatique de Toulouse, UMR 5505, Université de Toulouse, 118 Route de Narbonne, 31062 Toulouse Cedex 04, France
(2) Laboratoire Cognition, Langues, Langage, Ergonomie - Équipe de Recherche en Syntaxe et Sémantique, UMR 5263, CNRS et Université de Toulouse, 5, allées Machado F-31058 Toulouse CEDEX 9, France
(3) Institut Universitaire de Formation des Maîtres, 56 av de l'URSS, 31079 Toulouse, France
josiane.mothe@irit.fr, ludovic.tanguy@univ-tlse2.frl

Abstract

Most of Information Retrieval Systems transform natural language users' queries into bags of words that are matched to documents, also represented as bags of words. Through such process, the richness of the query is lost. In this paper we show that linguistic features of a query are good indicators to predict systems failure to answer it. The experiments described here are based on 42 systems or system variants and 50 TREC topics that consist of a descriptive part expressed in natural language.

1. Introduction

Research in Information Retrieval (IR) aims at proposing models and methods in order to build systems that answer a user's need as completely and as precisely as possible: retrieving the relevant information while avoiding non-relevant information.

Different IR models have been proposed in the literature. In the Vector Space Model (VSM) [1] a vector represents the document in the indexing term space. A query is represented in the same way and possibly relevant documents are selected according to the similarity of the query and document vectors. Another commonly used model is the probabilistic retrieval model which calculates the probability of a document being relevant to a query [2],[3] whereas Language Modeling [6] is based on the probability of the language model of the document to generate the query. The Latent Semantic Indexing (LSI) [4] improves the VSM in the fact that it reduces the document dimension. In that model the document dimension is not the number of index terms but a smaller dimension obtained using the single value decomposition theory [5].

Whatever the underlying model, in most systems, texts (documents or queries) are first parsed in order to remove stop words and the remaining terms are stemmed in order to represent the different surface variations of a term by a unique root or word. However, some terms are considered more important than others, depending on their discriminatory power. The importance of a term is directly linked to its frequency both in the parsed text and in the entire collection. In the vector space model for example, the document coordinates are given by $d_{ij}=tf_{ij} \cdot idf_i$ where tf_{ij} is the term frequency of the term *i* in the document *j* and idf_i is the inverse document frequency, basically $\frac{1}{N_i}$ where N_i is the number of documents where *i* occurs.

More advanced text parsing techniques have also been used. Considering phrases rather than single terms has been studied in different contexts. In [7], phrases were used in an ad-hoc retrieval task. Two different ways of phrase extraction were used: statistical and syntactical. No significant difference was found. More importantly, the use of phrases instead of simple words did not significantly affect the overall results. In [8], phrases were used for passage retrieval: once again, no significant difference was found compared to the use of single terms. Other similar studies led to comparable results, including works on morphological analysis, use of semantic information, etc.

Intuitively a richer and more linguistically-aware processing of texts should lead to better retrieval results and therefore, Natural Language Processing (NLP) has been used in IR in different ways. However, despite those efforts, improvements on IR efficiency have not been proved on a large scale evaluation [9]. It has to be noted that this conclusion

is drawn from standard evaluation programs, in which system results are computed averaging recall and precision measures over fifty queries. Doing so, variability is hidden. We argue that detailed analysis of retrieved results should help us achieve a better understanding of the mechanisms and their influence on the results, as well as predict when systems will succeed or fail.

Some recent works aim at studying result variability. [10] distinguishes three types of factors that can contribute to variability on system performance: topic statement, relationship between topics and documents and system features. The RIA workshop focused on query expansion issue and analysed both system and topic variability factors on TREC collections. [11] report a work on CLEF topics, studying correlation between system performance and query features. They found a correlation of 0.4 between the number of proper nouns and average precision. [12] analyses TREC topics according to linguistics features and shows that the *average polysemy value* of query terms is correlated to recall. [13] show that topic difficulty depends on the distances between three topic components: topic description, the set of relevant documents, and the entire document collection.

The work presented here has similar objectives: can we identify some characteristics in users' queries that can explain the variations between systems, and lead us to both better understanding of IR mechanisms and weaknesses, and some guidelines towards more efficient techniques. Therefore, we carry out a deep analysis of some results obtained in the TREC[1] environment. We show that it is possible to cluster topics according to linguistic features and that these clusters can be correlated with systems when considering recall.

In section 2 we present the framework of the study: the IR task, the data used, and the query features. Section 3 reports the analysis of the system results. Section 4 discusses the results and present future works.

2. Experiment framework

International experimental environments such as TREC accumulate retrieval results with a large variety in terms of systems, tasks and test collections. Because it was impossible to analyze all the results that came out from international evaluation programs, we made the decision to focus on TREC Novelty Track. Whatever the track, an evaluation collection consists in the following:

- a number of pre-defined documents (e.g. newspaper articles),
- a set of topics. Each topic consists in a user's query (see below) in natural language,
- and the list of relevant information items corresponding to each query.

Both queries and relevant sets are manually defined. The relevance judgments are used to measure the system performance.

TREC Novelty track has been introduced in TREC 2002.

2.1. Information retrieval task: TREC Novelty track

The TREC novelty track has been leading the development of new research in passage retrieval within non structured documents at the sentence level [14]. The TREC novelty track is composed of two different goals namely (1) retrieving relevant sentences from relevant documents and (2) selecting the sentences that bring new information (information not seen before in the document or in a previous document). These two goals are declined in different contexts, each one leading to a TREC sub-task:

- task A: given a set of relevant documents, for each topic, NIST selected relevant documents with a maximum of 25 documents per topic. These documents are given to participants, sentences being marked-up. Goals (1) and (2), as explained in the previous paragraph, are proceeded.
- task B: given the set of relevant sentences, for each topic, NIST indicates the relevant sentences. Participants have to proceed goal (2).

2.2. Collection description

In TREC 2002, 49 topics were used from the TREC collection. As said previously, for each topic NIST selected 25 relevant documents from previous TREC tasks.

For each topic, after runs were submitted, NIST evaluators decided which ones among them were new.

Figure 1 corresponds to an example of a TREC topic. It is composed of three textual parts: a title that is supposed to correspond to a typical user's query. It is composed of just a few words. It is written under the form of keywords and not necessarily in real natural language. The two other parts are written in natural language. The descriptive part explains the title whereas the narrative part describes what will be a relevant sentence and a non-relevant sentence.

```
Topic: 310
Title: Radio Waves and Brain Cancer
Description: Evidence that radio waves
   from radio towers or car phones affect
   brain cancer occurrence.
```

[1] TREC : Text REtrieval Conference : trec.nist.gov

Narrative: Persons living near radio towers and more recently persons using car phones have been diagnosed with brain cancer. The argument rages regarding the direct association

Figure 1: Sample topic (TREC 2002)

Table 1 reports some features of the TREC collection.

	NIST2002
Number of topics	49
Number of documents per topic (avg over topics)	22.3
Number of sentences per topic (avg)	1321
Relevant sentences per topic (avg)	27.9
% of relevant sentences (avg)	2.1
New sentences per topic (avg)	25.3
% of new sentences (avg)	90.9

Table 1: TREC 2002 Novelty track collection

2.3. Evaluation

Each participant submits runs to NIST that are evaluated against human judgments. The evaluation measures proposed in TREC Novelty track are based on commonly used measures of recall and precision. In the general framework of document retrieval, recall and precision are defined in terms of number of documents. When considering the sentence level, these measures become[14]:

$$Rs = \frac{Number\ of\ relevant\ retrieved\ sentences}{Number\ of\ relevant\ sentences}$$

$$Ps = \frac{Number\ of\ relevant\ retrieved\ sentences}{Number\ of\ retrieved\ sentences}$$

$$Fs = \frac{2 \cdot Ps \cdot Rs}{Ps + Rs}$$

The *Fs* measure is also used. It is defined above in terms of Precision and Recall.

These measures are computed for each query and then averaged over all topics. Similar measures are used to evaluate novelty detection.

2.4. Runs

Each run a participant submits is available on the TREC server for active participants. Also available are the measures obtained for each query by each run. Table 2 provides some examples of run results (average results over the set of queries).

Run	Recall	Precision	R*P
Dubrun	0.49	0.15	0.19
Thunv1	0.34	0.23	0.235
Thunv3	0.41	0.20	0.235
Pircs2N01	0.49	0.16	0.209
Nttcslabnvr2	0.60	0.10	0.166

Table 2. Average recall and precision for some runs.

There are 42 systems or system variants for the Novelty 2002 task 1 that we consider in this paper.

Runs and evaluation of these runs are the inputs of the analysis we report in section 4.

2.5. Topic features

The use of linguistic features to characterize a text is a commonly-used technique in text classification. It has been used for the identification of text genre characteristics [15] and even stylistic studies on IR documents [16]. The purpose of these features is to describe some of the linguistic characteristics of a given text, and to study their correlations with themselves and other phenomena. [11] used such techniques, and manually identified some features on CLEF topics.

We calculated a number of such features for each topic, taking their title and description parts into account (thus ignoring the longer narrative parts, as most IR systems do). As these parts only contains between one and three sentences, some of the more statistically-oriented features could not be computed, or led to too many sparse values. We also restricted our study to features that can easily be obtained automatically, as relevant features could thus be used in an adaptive system. We also focused on features that could be matched with known NLP techniques, and as such are clues to specific difficulties in the processing. The three categories are morphology, syntax and semantics. Morphology deals with the variation of words across documents and queries, and is processed through well-known normalization techniques such as stemming and lemmatization. Syntax deals with the functional relations between words, and its area covers the notions of phrase identification. Semantics deals with word senses, and is the area that covers query expansion techniques (i.e. automatic adjunction of words in a query).

In the end, we selected the following features:

a) Morphological features :
- average word (token) length — LENGTH
- average number of morphemes per word — MORPH
- number of suffixed words — SUFFIX

b) Syntactical features :
- number of conjunctions — CONJ
- number of prepositions — PREP
- number of verbs — VERBS
- average syntactic tree depth — SYNT DEPTH
- average syntactic distance — SYNT DIST

c) Semantic features :
- average polysemy value — POLYSEM

Each topic was first processed by a POS tagger and lemmatizer (we used Schmid's TreeTagger[2]) and a syntactic analyzer named SYNTEX [17].

Morphological features are used to reflect the morphological complexity of words used in a query. The most crude measure is the word length (measured in numbers of characters), which does not need any specific linguistic resource. The average number of morphemes per word is a more sophisticated measure, relying on the CELEX[3] morphological database, in which 40,000 base word forms are described. For example, we find in this database that "*additionally*" is a 4-morpheme word ("add+ition+al+ly"). Heavily constructed words are known to be more difficult to be matched with morphologically similar words, thus requiring specific processing. The limit of this method is of course the database coverage, which declares rare, new, or misspelled words as mono-morphemic. To provide a more robust analysis method, we developed a third measure, which focused on the more common morphological operation: suffixation. We computed a list of the most common suffixes for English, and used it to detect whether a word is constructed by suffixation or not. As an example, this measure is able to detect that a rare undescribed word such as "*postmenopausal*" is suffixed (*-al*).

Syntactic features focus on sentence complexity. We used two different techniques to characterize such complexity: the first one is to look for specific word classes such as pronouns and conjunctions, as simple clues to complex structures and phenomena (coordination, anaphora, etc.). The second one is to take advantage of an automated syntactic analysis. The SYNTEX parser is a dependency grammar analyzer that gives for each word in a sentence, the ones to which it is syntactically linked (e.g. it identifies relations between a verb and its subject, object, between an adjective and a noun, etc.). This information can then be used to build a more classic syntactic tree. Given these two possibilities, we computed two different measures of syntactic complexity. Syntactic depth is the degree of hierarchical complexity for each sentence. Syntactic distance, on the other hand, measures the average span of a syntactic link on the syntagmatic axis. For example, a subject noun that is separated from the verb because of complex noun phrases, or subordinates, will lead to an increase in this latter measure, but not necessarily in terms of syntactic depth.

[2]http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

[3]http://www.ru.nl/celex/

Another important set of characteristics is related to semantics. The main semantic problem encountered in IR is polysemy, as the terms used in a query can have different meaning in different contexts. We focused on a simple measure, using the WordNet[4] lexical database. This polysemy value is directly available in WordNet (in terms of the number of different synsets the words belongs to), 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 monosemic, so the default value of one used for words absent from WordNet is supposed to be a good approximation.

Features relying on occurrences are expressed as percentage values. For example, a PREP value of 0.12 indicates that 12% of the words in a query are prepositions. Other measures are averaged over every words or sentences in a query.

For every NLP technique used, a certain amount of error is expected, depending on its complexity. However, we tried to use the most reliable clues for each phenomenon, and manually checked each feature detection technique.

3. Analysing runs: clustering queries and systems

The study presented in this section aims at discovering correlations between topic features and system performances. This has been done through two different steps. The first step (3.1) consists in a cross-analysis of topics and linguistic features, leading to the definition of topics clusters based on linguistics features, without taking systems performance into account. In the second step (3.2) we cross-analyze topics and systems, and then project the classes resulting from the first step.

3.1. Classifying topics using linguistic features

As explained in section 2, we automatically parsed every topic/query in order to obtain numerical feature values. Table 3 reports these values for the first 10 topics.

We then used an agglomerative hierarchical clustering in order to build topics classes. A Hierarchical Agglomerative Clustering (HAC) produces a set of partitions of the initial objects, P_n, P_{n-1}, ……. , P_1. At one extreme, P_n consists of n single objects, at the other extreme, P_1, consists of a single group that consists of all n objects. In such a clustering, at each particular stage the two clusters

[4] www.cogsci.princeton.edu/~wn/

which are closest together are joined to form a new cluster. At the first stage, each cluster has a single object. We used the Euclidean distance to compute cluster similarity. The input vectors are not normalized and each feature is considered as equally important as the other ones. The fact that features are not equivalent considering the scale of their value is not taken into account in this study; but will be in future works. The resulting dendrogram is shown figure 2.

FEATURES	MORPHOLOGICAL			SYNTACTICAL					SEMANTIC
TOPIC	LENGTH	MORPH	SUFFIX	CONJ	PREP	VERBS	SYNT DEPTH	SYNT DIST	POLYSEM
TOP303	5.41	1.19	0.13	0.03	0.09	0.16	5.60	1.81	3.63
TOP305	4.79	1.05	0.09	0.02	0.12	0.16	4.17	1.61	4.92
TOP310	4.64	1.11	0.08	0.03	0.13	0.14	4.00	1.73	4.71
TOP312	5.19	1.14	0.17	0.06	0.13	0.12	2.60	1.40	4.00
TOP314	4.82	1.23	0.19	0.05	0.10	0.10	3.40	1.54	3.44
TOP315	4.99	1.13	0.09	0.00	0.05	0.17	3.60	1.59	2.86
TOP316	4.45	1.11	0.10	0.04	0.12	0.09	5.30	1.51	4.58
TOP317	4.83	1.15	0.10	0.05	0.05	0.17	3.29	1.52	4.09

Table 3. Linguistic topic features.

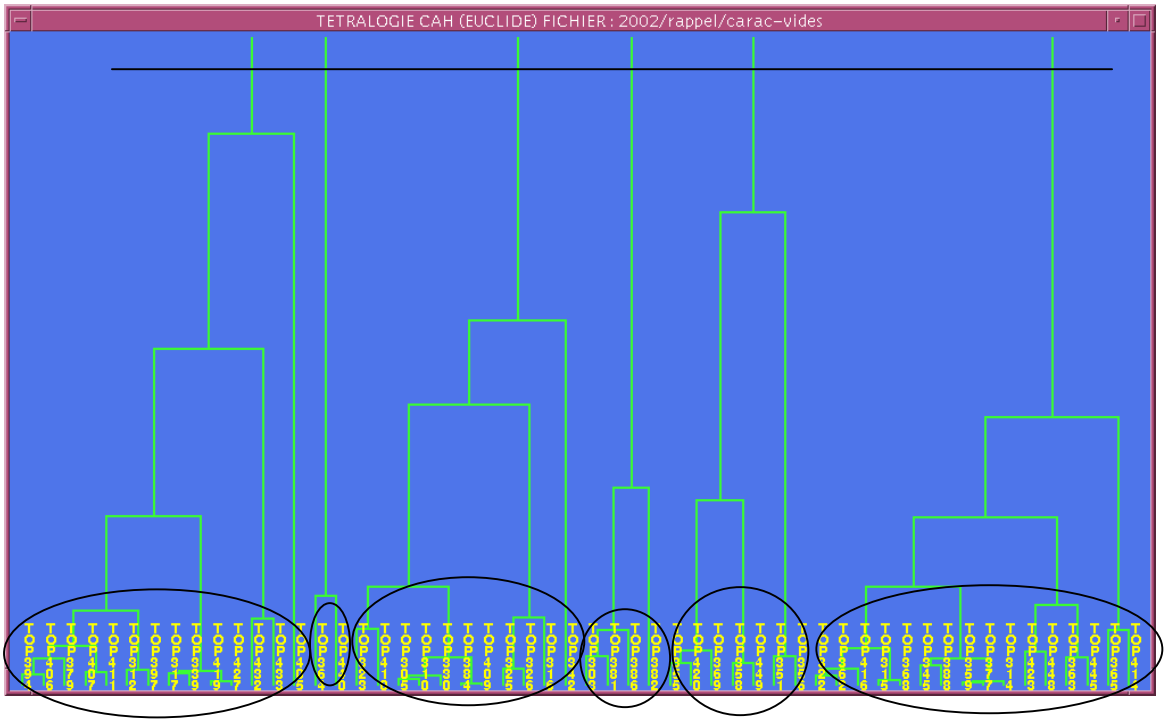


Figure 2. Dendrogram resulting from topic clustering.

In the following, we will consider only one of the possible partitions presented in the dendrogram. The number of classes we chose as a partition takes into account the fact that the partition should lead to a small number of clusters in which objects are similar enough. The inter-cluster distance leads us to consider a 6-class partition. Although this clustering method led to 6 clearly identifiable topics clusters, no simple correlation was found between linguistic features (taken individually) and these classes. This has already been pointed out in similar studies [11]. However, as shown in the next section, this does not prevent a high significance of these classes with the runs' performance scores.

3.2. Analyzing recall

Detailed results have been analysed using Principal Component Analysis [18]. The general goal of this data analysis method is to represent vectors (called characters) initially represented in a space of N dimensions (called variables) into a smaller space. Principal Component Analysis reduces data dimensionality into spaces which are the most important as determined by the eigen values of the variance/covariance matrix (using Euclidean distance). The eigen vectors are then known to be the most useful to visualize the maximum of information. Moreover, the most specific information will be the first displayed.

In the case of the analysis we report here, the recall obtained by each system is the measure, systems are variables, and topics are characters. Table 4 presents an extract of the resulting matrix.

	CIIRkl	CIIRnew	cmuAs	cmuBw	cmurCb	cmurCv	cmurCw	dumbrun
t305	0.33	0.33	0.2	0.07	0.2	0.2	0.2	0.27
t312	0	0	0	0	0	0	0	0
t314	0.72	0.72	0.6	0.32	0.88	0.88	0.88	0.16
t315	0.36	0.36	0.45	0	0.64	0.64	0.64	0.27
t316	0.89	0.89	0.06	0.06	0.11	0.11	0.11	0.67
t317	0.52	0.52	0.39	0.26	0.52	0.52	0.52	0.57
t322	0.26	0.26	0.09	0.18	0.15	0.15	0.15	0.41
t323	0.42	0.42	0.13	0.17	0.23	0.23	0.23	0.23
t325	0.48	0.48	0.05	0.1	0.24	0.24	0.24	0.38
t326	0.75	0.75	0.25	0.25	0.38	0.38	0.38	0.75

Table 4. Extract of the recall matrix.

Figure 3 presents the graphics resulting from the PCA based on the recall matrix. It is displayed according to axes 1 and 3; they correspond to 50% of the total inertia. Figure 3a) presents the characters whereas figure 3b) presents the variables. In the former graphic, a color and a specific form of plot have been associated to each class of topics detected section 3.1. For example the cluster that appears on the left side of the dendrogram figure 2 is represented in green and circles figure 3a).

A first interesting result that can be discovered visualizing figure 3a) is that the clusters of queries have a direct correlation with the behavior of the systems regarding recall. Indeed, the queries of each cluster resulting from the HAC on linguistic features are situated close each other on the PCA visualization.

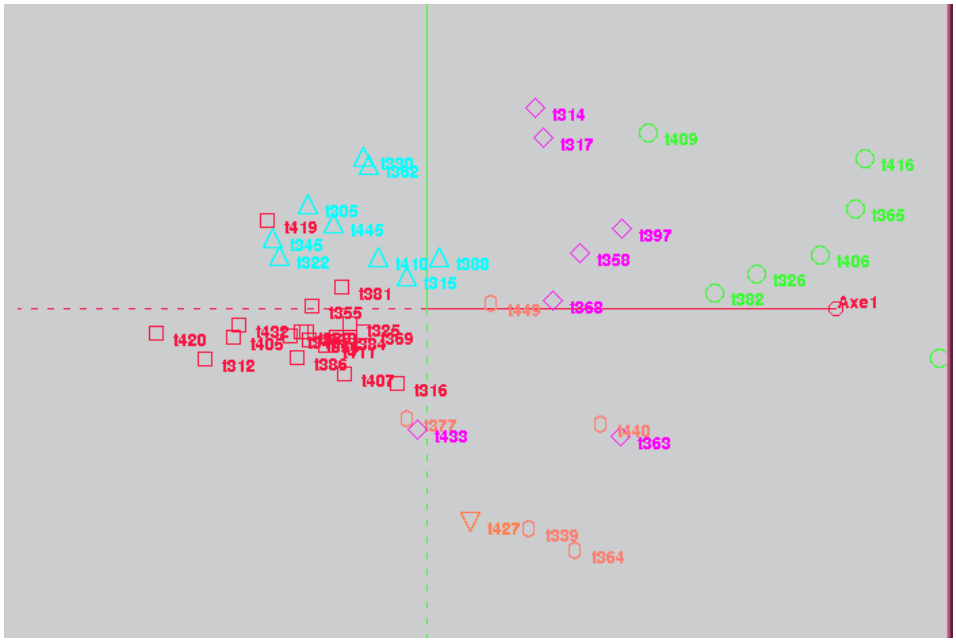
Variables (that represent systems) that appear on the periphery of the virtual hyper-sphere distinguish two groups of systems (figure 3b) that we arbitrary name Group 1 and Group 2. Group 1 consists of the following systems: *ntu1*, *ntu2*, *ntu3*, *colmerg* and *cmuBw*. Group 2 consists of *pircs01*, *pircs02*, *pircs03*, *thunv2* and *CIIRNew*. These groups are determined visually and chosen because they are orthogonal considering the first axes. Alternatively, we could have chosen to consider their coordinates on the first axes.

Each variable defines a vector in the new computed space. We drew three vectors to illustrate this on figure 3b. One corresponds to the *ntu1* run, the other to *ntu3* run and the third to *pircs02*. The contribution of the characters (topics) to a vector can be visualized on figure 3a). For example, topic 314 and 317 are positive contributions to *ntu1*; in other words, systems belonging to Group 1 obtain a high recall for these two topics whereas the other systems get a lower performance. This can be validated going back to the raw data: Run *ntu1* obtained recall 0.72 whereas the average recall over the systems for this topic is of 0.41. In the same way, regarding 317, ntu1 obtains 0.91 - the best recall for this topic- whereas the average recall over the systems for this topic is of 0.42. Similarly, topic 363 is a positive contribution to *pircs02*. Again, going back to the raw data, we found that *pircs02* obtains 0.9 for recall –which is again the maximum- whereas the average recall over systems for that topic is of 0.41.

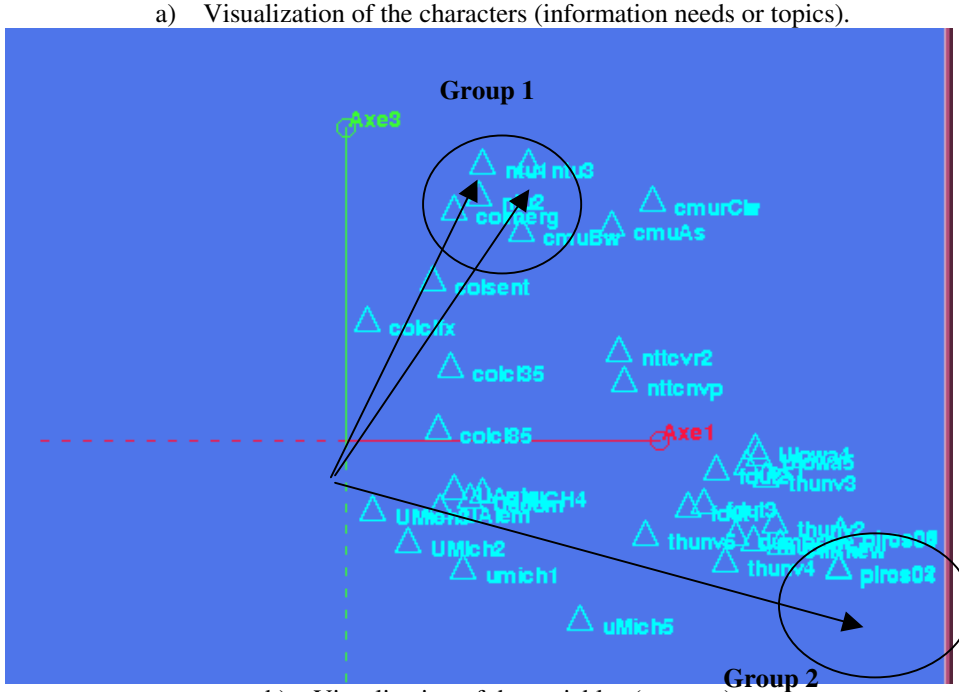
Analysing the two graphics simultaneously (Figures 3a and 3b), we discover that the cluster of *blue* queries (represented by a triangle and situated in the left-top corner of figure 3a) are easier for the systems belonging to Group 1 than for systems belonging to Group 2 (in term of recall). This information is extracted from the graphics where it can be seen that considering the origin of the axes, topics in blue have positive contribution on axis 3 – the axis of systems belonging to Group 1- and negative contribution on axis 1 –axis of systems belonging to Group 2). Table 5 presents the results obtained when averaging recall obtained for this cluster of queries by each group of systems. Recall is 0.42 for Group 1 of systems against 0.26 for Group 2. Averaging the results over all the systems for this cluster of queries leads to a recall of 0.29.

The opposite phenomenon occurs for queries belonging to the cluster in *orange* (and plotted using an oval figure 3a): systems from Group 2 get a better recall than systems belonging to Group 1. *Green* and *red* clusters of queries (plotted using a circle and a square respectively) behave in similar ways: green queries are easy for the two groups of systems whereas red queries are difficult whatever the group of systems.

These results are summarized table 5.



a) Visualization of the characters (information needs or topics).



b) Visualization of the variables (systems).

Figure 3. PCA (characters: topics, variables: systems, measure: recall) using axes 1 and 3.

	Class of queries				
	Blue (triangle)	Orange (oval)	Green (circle)	Red (square)	All queries
Group 1 of syst.	**0.42**	0.24	0.50	0.22	0.32
Group 2 of syst.	0.26	**0.62**	**0.77**	0.29	0.46
All systems	0.29	0.38	0.53	0.22	0.34

Table 5: Average recall over groups of systems and clusters of queries

4. Discussion and future work

We cluster topics according to the linguistic features they share. These clusters appear to be closely correlated to the success or failure of some systems (whereas previous studies showed the lack of correlation between features and average performance). Mining recall obtained by 42 systems on 49 topics representing users' information needs, we found that some clusters of topics can be

associated with types of systems. This is an important result as it opens a new track for data fusion. Data fusion relies on the fact that different strategies lead to different results and thus merging these results in a relevant way may improve the results. The literature of the domain reports studies that take into account features on the retrieved document set as good indicators of the prediction of the fusion effectiveness [19], [20]. If it was possible to decide which system would work in a given context; combining different systems could improve the results in a much more interesting way. This paper is a first contribution towards this direction. We show that it is possible to decide for each type of queries what would be the best system to use when recall is to optimise. Next step is to complete this study including precision measure, as it is well known that recall and precision vary in opposition and including more data from other IR tasks. Moreover, a more detailed analysis of individual runs can easily lead us to pinpointing which linguistic features is positive or negative for a given system. This can further leads to a better understanding of a given technique (known to be used by a system) when processing specific linguistic phenomena.

Another future work is to try to extract rules that could be applied to decide to which cluster a new query belongs to. Indeed, in this paper, we show that it is possible to cluster the topics but we did not extract the corresponding rules. Finally, an application to this work is to develop a fusing method that would be based on existing systems and on topic clusters we detected.

10. References

[1] Salton, G. (1971). The SMART retrieval system: Experiments in automatic document processing. Englewood Cliffs, N. J.: Prentice-Hall.

[2] van Rijsbergen, C. J. (1979). Information Retrieval. London: Butterworths.

[3] Robertson, S. E., and K. Spark Jones (1976). Relevance weighting of search terms. J. of the American Society for Information Science 27 (3), 129-146.

[4] Deerwester, S. et al. (1990), Indexing by latent semantic analysis, J. of the Society for Information Science, 41(6), 391-407.

[5] Eckart, C. & Young, G. (1936). The approximation of one matrix by another of lower rank. Psychometrika, Vol.1, 211-218.

[6] Ponte J.M., Croft W.B., (1998). A language modeling approach to information retrieval, 21nd Inter. Conf. on Research an Development in Inf. Retr., 275-281.

[7] Mitra M., Buckley C., Singhal A., Cardie C., (1997). An analysis of Statistical and Syntactic Phrases, RIAO, 200-214.

[8] Dkaki T., Mothe J., (2004). Combining Positive and Negative Query Feedback in Passage Retrieval, RIAO 2004, Coupling approaches, coupling media and coupling languages for information retrieval, 661-672.

[9] Sparck Jones K. (2003): Document Retrieval: Shallow Data, Deep Theories; Historical Reflections, Potential Directions, ECIR 2003, 1-11.

[10] Buckley, C and Harman, D. (2004) Reliable Information Access Final Workshop Report, Jan. 2004.

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

[12] Mothe, J., Tanguy, L., (2005), Linguistic features to predict query difficulty - A case study on previous TREC campaign. SIGIR workshop on Predicting Query Difficulty - Methods and Applications, pp. 7-10.

[13] Carmel D., Yom-Tov E., Darlow A., Pelleg D., (2006) What makes a query difficult?, 29th Inter. Conf. on Research an Development in Inf. Retr., 390 – 397.

[14] Harman D., (2002). Overview of the TREC 2002 novelty track, Text Retrieval Conference, 46-55.

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

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

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

[18] J.P. Benzecri, (1973) L'analyse de données, Tome 1 et 2, Dunod Edition.

[19] Lee, J.H., (1997). Analysis of multiple evidence combination, 20th Inter. Conf. on Research an Development in Inf. Retr., 267-275.

[20] Beitzel S.M., Jensen, E.C., Chowdhury, A. Grossman D., Goharian N. Frieder O. (2004), On Fusion of Effective Retrieval Strategies in the Same Information Retrieval System, J. of the American Society for Information Science and Technology, 55(10), p 859 – 868.

Acknowledgment Research outlined in this paper is part of the project WS-Talk that is supported by the European Commission under the Sixth Framework Programme (COOP-006026). However views expressed herein are ours and do not necessarily correspond to the WS-Talk consortium.