Annotation of linguistic phenomena and topics in query logs

JEAN-LEON BOURAOUI, BENOIT GAILLARD, EMILIE GUIMIER DE NEEF, MALEK BOUALEM

Orange Labs
2 avenue Pierre Marzin
22300 Lannion cedex, France
{jeanleon.bouraoui, benoit.gaillard, emilie.guimierdeneef, malek.boualem}
@orange-ftgroup.com

Abstract. We present a thorough analysis of query logs of various search engines. We first propose a methodology to annotate these logs and explain how it is applied to the corpus. Finally, we report three main observations issued from this study: the distributions of categories of queries significantly vary according to the search engines they come from; Named Entities are very frequent; the queries are often ambiguous and include spelling errors.

Keywords: corpus analysis, query logs, corpus annotation

Analyses of query logs on research engines are quite rare, due to the fact that only search engines editors can access to this kind of corpus. However, categorizing the queries and thematic subjects that interest most of the users is useful for web providers to adjust their offer. From a more theoretical perspective, the categorizing queries also presents some interest since a good classification can yield significant performance gains for search engines. For example, Query Expansion (QE) or *Cross Language Information Retrieval* (CLIR) techniques can benefit greatly from accurate linguistic descriptions of query corpora. This was shown especially by Jensen (2006, p. 185).

In the first part of the paper, the different resources used to carry out the study are described. In a second section an annotation methodology is proposed. Finally, the most interesting phenomena that have been highlighted are presented.

1) PRESENTATION OF THE ANALYZED CORPORA

Five different corpora were used for this research. Four of them are composed of user's queries on various search applications.

The first corpus comes from the query log of a search engine prototype targeting video contents in the domain of news. The corpus contains 3 380 queries with some rare repetitions, submitted to the engine from July 2008 to January 2009. The whole corpus has been annotated.

The second corpus includes query logs of a search engine specialized in User Generated Content (UGC): videos that users generated and uploaded themselves (similar to *YouTube*).

The corpus includes 10 078 query patterns[1], ranked by frequency of occurrence (the most frequent query has been entered 695 times). 71.56% of the total number of queries were annotated.

The third corpus is related to the previous one, and corresponds to the index of the search engine (extracted single word patterns, ranked by frequency of occurrence). There are 123 335 words in this corpus; we annotated 39 392 of them, that is to say 27.07%. This resource was used to compare the classification that emerges from the queries to the one that can be observed in the index.

The fourth corpus contains query patterns submitted by users to the news section of 3 separate search engines. These queries bear on textual contents (and not video contents as for the second corpus).
The corpus contains 186 300 queries, ranked by frequency of occurrence (the most frequent query has been performed 29 395 times). The annotation was carried out on the 1000 most frequent query patterns, which corresponds to 45.36% of the total number of actual queries.

The fifth corpus is made of query patterns submitted from mobile phones, on a general search engine. It contains 4 655 433 queries; the most frequent was repeated 630 546 times, and the less frequent 26 times. The annotation was carried out on the 1000 most frequent query patterns, which corresponds to 85.07% of the total number of actual queries.

2) ANNOTATION METHODOLOGY

2.1) STATE OF THE ART FOR QUERY LOGS ANNOTATION

In 1999, some analysis of query logs have been carried out and described in (Silverstein and al., 1999). In 2000, a study by (Jansen and al., 2000) showed that most queries are made of 2 to 4 words; this study was carried out in the framework of Information Retrieval in multimedia data. A similar, more recent study is made by (Chau and al., 2005). In her Phd thesis (Léon, 2008), S. Léon introduces the notion "complex lexical units"[2], that gathers locutions, compound, Named Entities, and describes some methods of extraction and translation of such units. As we will show, the queries often include such units.
More recently (2006), the American provider AOL put on a web sites some query logs corresponding to 3 month (a total number of 20 millions queries, submitted by 650 000

[1] A query pattern corresponds to the same query entered several times by users. For example, in the second corpus, the query "noel" has been entered 332 times.

[2] In french : « unités lexicales complexes ».

different users). Initially, the goal of AOL was to provide data for the researchers. Each query from these logs is made of several fields: an anonymous ID for each single user, the query itself, the day and hour of the query, and, when applicable, the link clicked by the user among the answers to his queries. Consequently, these logs include all the necessary information for a detailed analysis of the queries submitted on a "common use" search engine. Several websites[3] propose various services devoted to the study of these logs: search engines, ranking of the most frequent sites or words used in queries, etc.

These lists show that the most frequent queries relate to other search engine (firstly Google) or sites with a large audience such as Myspace, Ebay, etc.

The methodology for annotating a given corpus, that is to say set of rules and categories used, is called an "annotation scheme". Such a scheme has to be validated to be considered as robust. In order to do it, it is necessary to compare the annotation of a same corpus by several annotators. It can be done by the use of the "kappa measure", described for example in ((Krippendorff (1980), cited in (Carletta 1996)); basically, it is based on the number of the inter-annotator differences. A more recent study of the methodologies of Named Entities annotation is described in (Fort and al., 2009)

2.1) DESCRIPTION OF OUR ANNOTATION METHODOLOGY

We initially have identified some *classification topics*, with corresponding *categories*. The goal was to take into account the main relevant linguistic phenomena and topics to represent the queries: morphological and syntactic features, semantic relations, etc. For example, we defined the classification topics for the *domain* concerned by the query, starting with categories such as: *TV*, *politics* …

From there, two annotators carried out a manual analysis and annotation of the two first corpora. Each time they found a query that highlighting an interesting phenomenon that they had not yet identified, they created a new category or topic in which to classify the query. The resulting annotations were frequently compared to each other. The goal was to obtain a unified annotation scheme. Thus, it was possible to assess the importance of topics and categories during the analysis process.

This methodology enabled to dynamically suppress, merge, or detail several categories and classification topics.

[3] For example, http://data.aolsearchlogs.com/ or http://www.seosleuth.com/site/

The resulting annotation scheme can be represented by trees (that are not necessarily deep). A query belongs to several trees. Thus, annotators and users of the analysis can choose which perspective they want to adopt to study the corpus or the results.

3) PRESENTATION OF THE CLASSIFICATION TOPICS AND CATEGORIES

In the process of annotating the logs presented in section 1, according to the methodology described in section 2, topics and categories were dynamically defined by the annotators. The final topics and categories already are in themselves a significant result as they highlight the salient features of the corpora. They constitute an annotation scheme that is very relevant to the corpus because it was defined by a bottom-up approach, based on the data.

There are 12 different first-level classification topics in the scheme; each one is subdivided into several categories. A category can become a classification topic; for example, the category "music" is also a classification topic that has "rock" as category.

We present in Table 1 an excerpt from the resulting set of topics and categories. Each column corresponds to a given category or topic, and contains some instances.

Lexical categories	**Grammatical categories**	**Categories of error**	**Domain**	**Linguistic phenomena**	**Ambiguities**
Named Entities	Name	Missing or added accentuation	Culture	SMS style[4]	2 different Named Entities
Expression[5]	Commun noun	Gender/number agreement between several words	Sport	Abbreviation	Named Entity or Commun noun
Single Word	Last name	Unrecognized character	Motor engines	Diminutive[6]	Polysemy
Date	Proper name	Deletion of one or several characters	Services	Implicit	Grammatical Ambiguity
Quote[7]	Noun phrase	Insertion of one or several characters	Policy	Play on words[8]	Correction alternatives
	First name	Transposition of one or several characters	General		Different senses according to the language

[4] Any query written using the same abbreviations than SMS. For example, the english word "before", when written "be4" will be labelled as "SMS style".

[5] We use this term in its linguistic sense: an expression is a set of words that is used as a single unit. For example, "grippe aviaire" ("asian influenza" in English)

[6] Usually used for any short nickname (for example, "Manu" for the French name "Emmanuel").

[7] Any query that correponds to a famous quote (example: "I have a dream").

[8] Any query that produce a humorous play on word.

	Nickname	Inversion of one or several characters	Economy		Different chunking available
	Verb	Segmentation	News		
	Adjective	More than one error in a word	Enterprises		
	Sentence	Phonetic spelling	Health		
	Acronym	Repetition of one or several characters	International		
	Key words		Geography		
	Miscalleanou		Miscalleanou		

Table 1 : Presentation of the most representative topics and categories of the proposed classification

Some of the categories are used as topics which are themselves divided into categories. This feature of the annotation scheme can be represented by trees. For example, one of the categories of the topic “*domain*” is “*culture*”, which is itself subdivided into categories such as “*music*”, which itself is divided into categories such as “*artist*” or “*title*”.

An instance of such trees is presented in Figure 1 below.

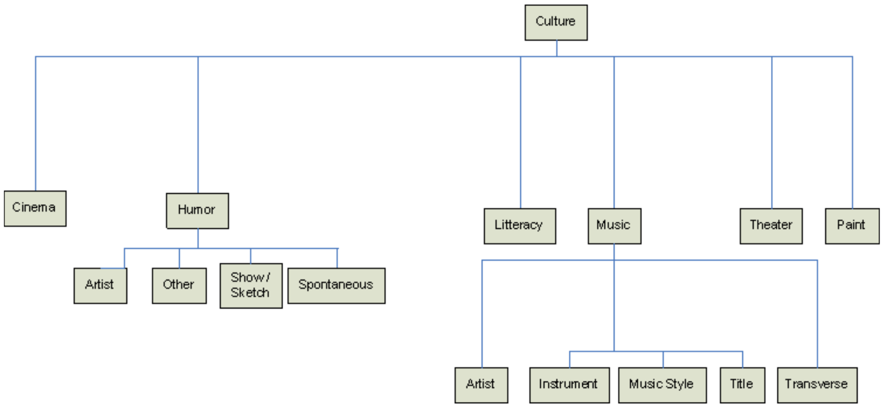

Figure 1 : Tree representation of the “culture” classification topic

3) Main observations

Although some observations are common to all corpora, there are also some significant discrepancies between the results of the annotation of the various logs. In consequence, we present in a first section the observations that are common to all corpora, and in a second one those which are specific.

3.1) *Observations common to all corpora*

A very large number of queries involve Named Entities. The majority of them concern people, places, and TV broadcasts. Figure 2 below displays the distribution, in the first corpus, of Named Entities categories (according to the classification described in section

3.1.); in this corpus, the Named Entities represent 41.06% of the total number of queries (the same trend is observed in the other corpus).

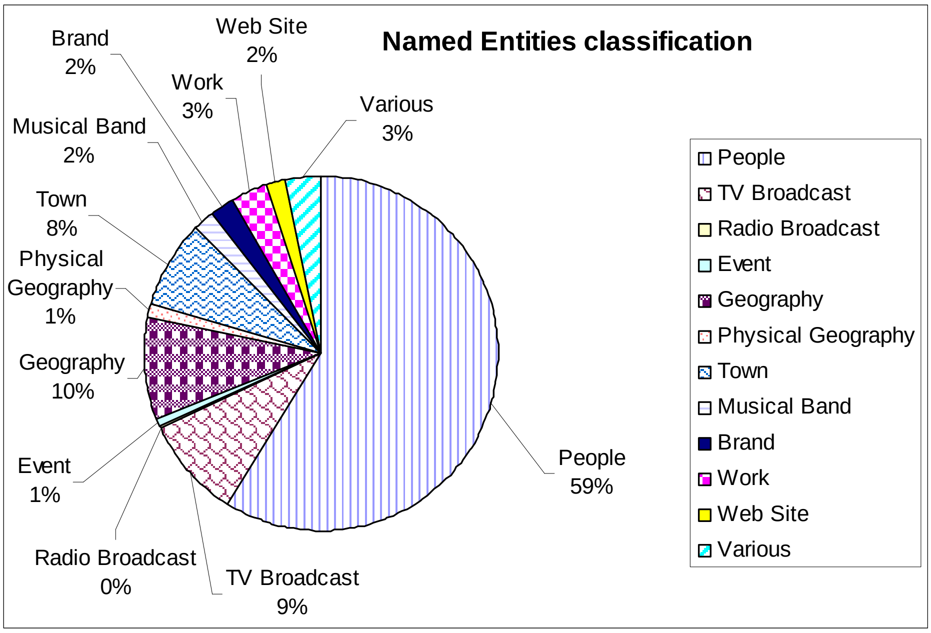

Figure 2 : Distribution of Named Entities categories in the first corpus, according to the proposed classification

In addition to Named Entities, a large number of queries contain phrases, words or compounds.

We noticed frequent spelling mistakes in the queries. Besides overwhelming accentuation and capitalization approximations, most of the mistakes are distributed among omissions, insertion and phonetic errors. These mistakes can cause difficulties while automatically processing queries for application such as CLIR or QE. For example the application can fail to recognize a Named Entity. The typology of errors and its distribution in corpus 1 is displayed in Figure 3 (the "0%" numbers correspond to only a couple of occurrences of the related categories)

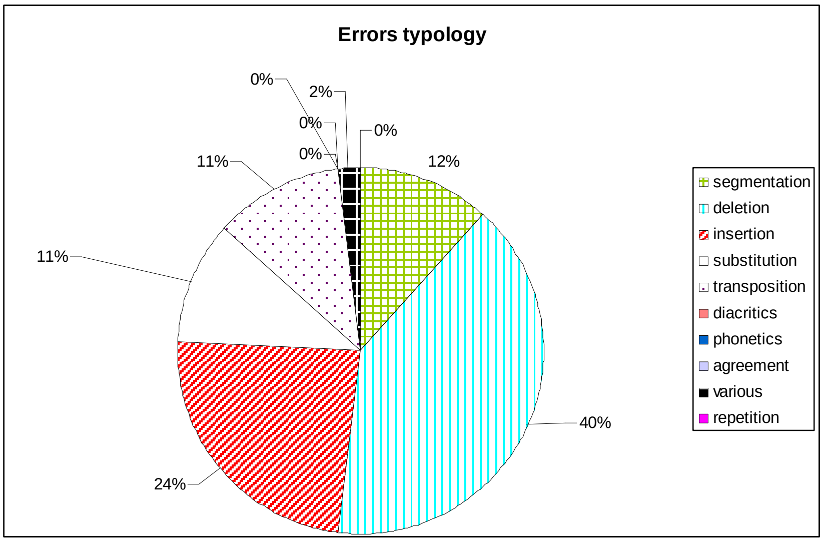

Figure 3 : Distribution of errors in the first corpus

An interesting observation resulting from this annotation work is the significant number of ambiguous queries. An ambiguous query is a query for which various alternative processing can be applied, and for which the choice between these various alternatives is not straightforward. A significant number of them arises from the fact that a word or (group of words) can refer to a Named Entity or to an usual word (for example, the word “cruise” in “Tom Cruise”). Another important cause of ambiguities is the misspelling. This is illustrated by the Figure 4Figure 3 below.

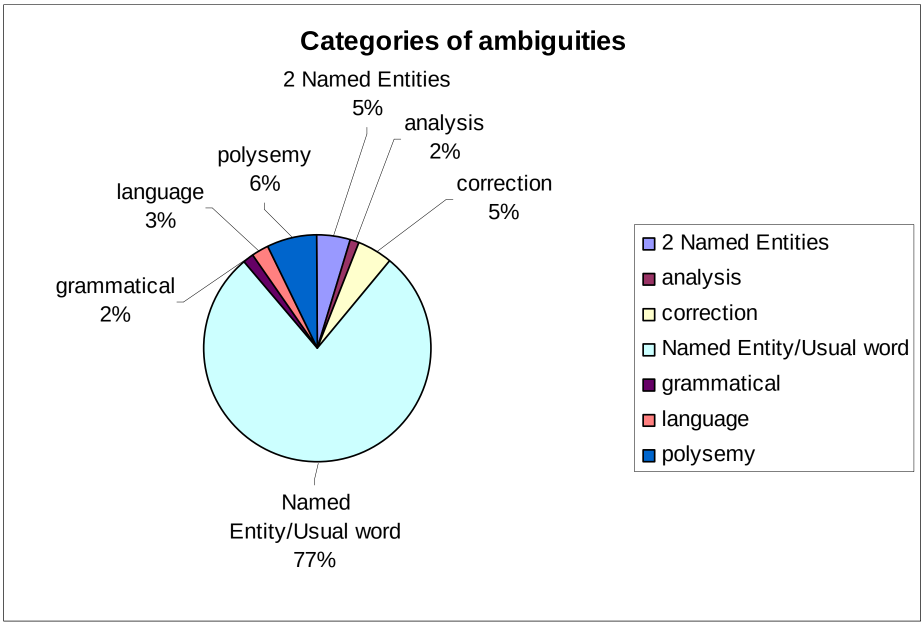

Figure 4 : Distribution of ambiguous queries

3.2) Corpus specific observations

The distribution of query categories clearly varies across the analyzed logs. Here, we present two examples of enlevé “the” corpus specific observations.

Since the first corpus results from “test” queries, submitted by researchers on a private prototype, it contains very little number of “adult” queries, whereas the other logs display a significant number of such queries.

We observed, on the mobile portal corpus, that a significant number of queries aim at directly finding the *url* of a known web service (e.g. “Google”); these queries seem to be used by the user as a shortcut or a bookmark to the service. This kind of queries is considerably less frequent in the other logs.

4) CONCLUSION

We presented a method of annotation of web-based query logs. This study involved several resources, following a methodology that uses the data to define the annotation scheme. We showed the benefits that this analysis and annotation scheme can bring to Information Retrieval. It was applied to a significant amount of the available resources. We analyzed and commented the various observed query categories.

From this analysis we can outline several needed processing for a better handling of queries: lemmatizing and interpreting them, using of orthographic correction and identifying their various components according to their typology. Our annotation scheme must also be validated by the use of the kappa coefficient, that we described in section 3.1. above. Finally, query logs analysis is sufficiently rich and adaptable to be used in a more systematic way. Some projects are in progress to automatically label the queries of the first corpus, with the *Tilt* platform (described in (Heinecke *and al.,* 2008)). Thus, it will be possible to compare the labelling done by Tilt with the same work by human annotators.

REFERENCES

Carletta J. (1996) "Assessing agreement on classification tasks: the kappa statistic", *Computational Linguistics,* Vol.2, Issue 2, p. 249-254

Chau M., Fang X., Liu Sheng O. R. (2005), "Analysis of the query logs of a Web site search engine", *Journal of the American Society for Information Science and Technology,* Volume 56 Issue 13, p. 1363-1376

Fort K., Ehrmann M., Nazarenko A. (2009), « Vers une méthodologie d'annotation des entités nommées en corpus ? », *TALN'09*, Senlis, France

Heineke J., Smits G., Chardenon C., Guimier De Neef E., Maillebuau E., Boualem M. (2008)« TiLT: plate-forme pour le traitement automatique des langues naturelles », *Traitement Automatique des Langues*, Volume 49 Numéro 2, p. 17-41.

Jansen J., Goodrum, A. and Spink, A., (2000) "Searching for multimedia: analysis of audio, video and image Web queries". *World Wide Web Journal 3(4).*

Krippendorff, K. (1980) *Content Analysis: An Introduction to Its Methodology*. Newbury Park, CA: Sage.

Léon S. (2008) « Acquisition automatique de traductions d'unités lexicales complexes à partir du web », PhD Thesis, 8 november 2008, Aix Marseille University.

Silverstein C., Henzinger M., Marais H., Moricz M. (1999), "Analysis of a very large web search engine query log", *ACM SIGIR Forum*, 33 (1), p. 6-12