Ambiguity of Queries and the Challenges for Query Language Detection

Juliane Stiller, Maria Gäde and Vivien Petras

Berlin School of Library and Information Science,
Humboldt-Universität zu Berlin, Dorotheenstr. 26,
10117 Berlin, Germany
{Juliane.Stiller, Maria.Gaede, Vivien.Petras} @ibi.hu-berlin.de

Abstract. In this paper, a sample set of 510 simple searches from the TEL action log 2009 is analyzed for query content and query language. More than half of the queries are for named entities, which has consequences for query language disambiguation. A manual identification of query language finds that often a definite language cannot be determined, because many named entities are not translated. Problems and challenges for query category and language identification are discussed. Further analysis shows that IP address and interface language are not very strong indicators for determining the query language.

Keywords: LogCLEF, log file analysis, query language, named entities, query language detection

1 Introduction

One of the challenges in cross-lingual information retrieval systems is to identify the user's information need which is expressed in short and decontextualized queries in multiple languages. To be able to process the query adequately (e.g. stem or translate correctly), it is essential to determine the language of the query.

In this paper, we explore what the challenges for query language identification and classification are, especially in the context of digital libraries. We extracted 510 search queries from the TEL action log corpus 2009 and analyzed them on a conceptual and linguistic level. Of special interest was the identification of query characteristics from the corpus. We examine the ambiguity of query terms and the resulting challenges in determining the query language.

We also looked at other signals, which might help to determine the language of a query such as the IP address or the interface language. An analysis of the relationship between interface language, IP address and country of origin of users and the query language was carried out.

With our analysis we follow up on results generated from the TEL corpus in the previous LogCLEF track in 2009[1], which is briefly summarized here. Data from The European Library (TEL) and Tumba! were evaluated with the aim to analyze and classify user queries in order to understand search behavior in multilingual contexts and to improve search systems [1]. In this context, Oakes and Xu [2] analyzed the query language used under certain interface languages. Furthermore, they found out that users rarely switch the query language during their sessions. The CELI research institute tried to identify translations of search queries, assuming that users of a multilingual digital library will repeat queries in different languages [3]. Ghorab et al. [4] looked on general statistics for comparing the behavior of users from different linguistic or cultural backgrounds and identifying communities. They observed that 20% of term changes involved language changes. Lamm et al. [5] investigated user search performance and interaction with the TEL interface. They defined successful and not successful user actions and discovered different search behaviors of users from different countries. Hoffmann et al. [6] pointed out the limitation of query logs and proposed to gain more context information through the semantic enrichment of queries by linking them to sources of background information such as Wikipedia since the most frequent queries are named entities.

The paper is organized as follows: Section 2 briefly describes the TEL simple search log corpus and provides some general statistics. In Section 3 we present and discuss the analysis of our sample query corpus introducing categories for this sample of digital library queries. Section 4 deals with the problems in query category and language detection and provides examples for characteristic difficulties. We conclude the paper by investigating signals for language identification analysis that includes information about the interface language and the language information of the country the searcher is from.

2 The TEL Simple Searches

For LogCLEF 2010, two corpora are provided. The first one contains logs from the Deutsche Bildungsserver, the second one logs from The European Library (TEL). We created a multilingual test corpus with TEL queries from the simple search interface. Log files from two different periods were available. In the first period from January 2007 to June 2008, action log files and server logs could be used for research. We analyzed the second set of data, which were action logs from the period of January to December 2009.

This one-year log data file contained 762,485 lines of log entries. We extracted queries, which either contained a simple search or an advanced search indicator (search_sim / search_adv) in the log file entry. Queries, which were entered on the result page or the full record view were not selected. Log entries, which contained a simple search made up 137,827 of the entries, advanced search 32,528 entries.

As the advanced search offers some categorizations of the query by adding certain facets like" title" or "author" we used only simple search queries for our sample

[1] http://www.uni-hildesheim.de/logclef/

corpus as they do not give any context information regarding the intent of the user. Therefore, the query entered by the user and the information saved in the logs are the only signals the system can get.

Figure 1: Example log entry from TEL action logs

```
2,guest,127.0,5E390977758E505C871AFB99E1342988,en,"(""t
oto"")",search_sim,"/en/search/collections/a0268,a0365/
",0,,,,2009-01-28 00:00:00.0
```

The query is shown in the sixth column field of the log entry (see Figure 1).

2.1 Query Length

As we extracted a sample corpus from the simple search queries, which initiated a search session, we also looked at the entire simple search queries to extract information about the average length of the queries. One important aspect is the length of the queries which can indicate the amount of textual information embedded in a query. Early studies investigated the query length in web search engines, for a comparison of the major studies see Jansen & Poosch [7]. To name one of these, Spink et al. [8] did a longitudinal study of Excite transaction logs and found that over the years there is a change in content users are intending to look for but not in the structure of the queries. They also found that over the years the average in query length was between 2.4 and 2.6 query terms.

A first analysis of queries in the cultural heritage domain was conducted by Jones et al. [9] who looked at queries from the New Zealand Digital Library. Similar patterns as for the web search queries were found. The average number of search terms in a query is 2.5.

We analyzed the words per query separated by white space in the simple search corpus. On average, the queries consist of 2.34 terms. Approximately, 43% of all simple search queries contain only one term, 27% two terms, 13% three terms, 7% four terms and only 10% of all queries contain 5 words and more. Our results validate again that query terms usually consist of a few keywords and therefore the correct language identification is very challenging.

2.2 Most Frequent Queries

To look at the most frequent queries of a retrieval system is an indication of trends and content people expect to find in the digital library. The most frequent queries in the TEL 2009 log files are *"toto",* followed by *"mozart"* and *"napoleon"* (see table 1). From the 10 most frequent queries, 7 contain named entities and expressed a search for a person. These queries already indicate the challenges in determining the language of the query as they cannot be assigned to one particular language.

Apparently, queries are not very often repeated since the most frequent one appears only 400 times.

Table 1: Top 10 frequent queries.

Query[2]	Frequencies
"toto"	400
"mozart"	400
"napoleon"	207
"bach"	198
"dante"	157
"einstein"	150
"chopin"	127
"a"	119
"music"	108
"harry potter"	106
"test"	98

3 A Sample Corpus for Query Analysis

From the simple searches, we extracted randomly 510 queries. The aim was to gain information about the query language, topic and intent of the queries. Another goal was to investigate the distribution of proper names and a categorization of different query types regarding their content.

3.1 Sample Corpus Query Statistics

In line with our findings about the entire simple search corpus our sample corpus showed an average in query length of 2.43 terms per query. More than ¾ of the queries consist of one term (41.5%), two terms (25%) or three terms (14.5%). 7% of the queries are composed of 4 terms, 12% have 5 and more terms. Ten queries contain only numbers such as ISBN /ISSN or dates and four queries consist of less than 3 characters.

Through a manual conceptual analysis of the extracted query terms we categorized the queries according to their content. We focused on flagging those queries, which might be problematic in terms of language identification and query translation. This includes proper names, uniform book titles and other entities requiring special recognition when processed during a cross-lingual search session. We subsumed these categories under the definition of named entities. Three different query types were defined in the context of named entities:

[2] The queries "toto", "a" and "test" are queries used for testing by the TEL office. It proves how important a thorough understanding of the data shown in log files is. To interpret user behavior by log file data correctly, it is necessary to exclude misleading log entries such as test data or search engine crawlers.

Table 2: Query types

Query type	# of queries	Example
Only NE	279	*egon schiele*
NE and other terms	37	*conrad huber coat of arms*
Non NE (topical)[3]	194	*translation*

This means that only 38% of our sample corpus queries could be readily translated with a dictionary-based translation approach - if the query language could be determined.

3.2 Development of Query Categories

Previous studies have dealt with search engine query classification according to their intent [10], search goals [11] or topics [12].

A log file analysis of English Altavista queries showed that 20% of the queries are navigational, 48% are informational and the rest (30%) are transactional queries (excluding all sexual oriented queries) [10]. Rose and Levinson [11] created a hierarchy of search goals where the first level resembled Broder's taxonomy changing the transactional query to a search goal for resources. They found a greater proportion (around 61%) of informational queries and a smaller of navigational ones (around 15%).

Other studies focused on an automatic query classification [13]. The shortness of queries poses great challenges for automated query classification [14].

For our sample corpus we identified the following 6 named entity categories and two for non-named entities terms (table 3).

Table 3: Query Categories

Category	Description	Example
person	artist, creator, scientists	*egon schiele, nicoals de bourgogne*
geo	monument, town, country	*germany, place de etoile*
work title	book, article, opera, pictures	*magna carta, radetzky marsch*
organization	institution	*turnbull aitken*
event	historical	*french revolution*
domain-specific	biology, medicine	*candida stellata, downbeat nystagmus*
topical	Navigational, non categorizable, ISBN,ISSN, dates	*studylounge.it, qualificações salário, 978-0-324-14459-7, "xviii*

[3] This category also contained 10 numbers expressing ISBN or ISSN, one URL.

As shown in table 4, besides the topical searches such as "*qualificações salário*" users are mainly searching for persons: "*dante*", followed by geo related topics, mainly countries or towns: "*japan*", and titles: *"social support and health status: a literature review (1997)"*. Queries that can be assigned to more than one class are often a combination of author (person) and work: *"all the russias by e. c. phillips"* and counted for multiple categories (see table 4).

Table 4: Number of queries per category

Category	# of queries
topical	194
person	181
work title	94
geo	49
domain-specific	4
event	2

3.3 Query Languages

Table 5 shows the languages in which queries were expressed more than 10 times. It is striking that most of the queries are ambiguous terms where it was not possible to identify the language. This was mainly the case for named entities such as persons or geographic terms. In different languages they have normally no spelling variants e.g. "*paris"*. Several queries are not named entities but still ambiguous across languages e.g. *"administration"* or *"culture"*. Besides the languages listed we found queries in 13 other languages. Additionally, 4 Latin terms appeared in the corpus which expressed a very specific information need, e.g. *"neuroptera myrmeleontidae"*.
We compared our manual language identification with an automated process using the Google Translate language detector[4].

Table 5: Query languages

Language	Manual analysis	Google language detection
Ambiguous	39.02%	-
English	31.18%	63.14%
French	6.86%	9.80%
German	5.49%	5.50%
Russian	3.33%	2.94%
Spanish	3.14%	2.35%
Italian	1.96%	3.14%
Other	9.02%	13.13%

[4] http://www.google.com/uds/samples/language/detect.html

The manual analysis showed that 39% of all queries cannot be assigned to a special language, which complicates the automatic language detection. In contrast to our analysis, Google did not detect any ambiguous queries with respect to language. With the Google API, more than 60% of the queries were detected to be English while our manual analysis identified 31% as English terms. The significant difference can be explained with a probable English language detection of the ambiguous queries because of an English language bias in the training or general Web data that Google uses.

4 Problems in Query Classification and Language Detection

It is a well known fact that the language identification of search engine queries is challenging but very important especially for multilingual information access. The correct language detection is necessary for further processing of the query such as stemming, spell checking, disambiguation or translation and the decision in which language the result list should be presented.

Web search queries are normally very short. Due to the large number of named entity queries – especially in the cultural domain - the automatic language detection has to deal with ambiguous terms or even terms that are not easily assigned to a certain language such as "Franz Kafka". In our sample corpus 61.96% of the queries contain named entities and 54.70% of the queries consist only of named entities. As table 6 shows, from the 279 named entity queries we determined 167 as ambiguous. These are terms that occur in many languages such as: *"Paris"* or *"Madonna"*. Of course there are also named entities that can be assigned to the different languages such as: *"Eiffelturm" (German), "Tour Eiffel" (France).*
It is also shown, that queries which contain a named entity and another word are less ambiguous than those that only contain a named entity.

The named entity recognition is a very important aspect concerning the identification of a query language. For example, the query term *"barber"* can either refer to the English word for *"hair dresser"* or to the composer *"Samual Barber"*. In this case the correct detection of language alone does not ensure the identification of the user information need or intent.

For search engines, there are also cases where correct language detection does not necessarily imply that the user wants to see the results in the same language. For example, although the identification of the language for the query "candida stellata" is Latin, a user entering this query from Germany, would most probably want to see German web pages, rather than web pages in Latin.

Table 6 shows the ratio of ambiguous terms in the sets of queries containing named entities and not containing named entities.

Table 6: Distribution of ambiguous terms in NE and non NE query sets

Queries	# of amb. terms	Percentage
316 containing NE	176	55.70%
279 only NE	167	59.85%
194 without NE	23	11.86%

This shows that many queries where the language cannot be clearly determined are expressing a search for a named entity. The 23 ambiguous terms which were not categorized as named entities are numbers and terms existing in several languages such as “culture” or “administration” or characters such as “a”. It is also worth to look at the ambiguity of different categories as shown in table 7.

Table 7: Ambiguity of different categories

NE category	# of amb. terms	Percentage
Person	147	81.21%
Geo	20	40.81%
work	11	11.70%

In the person category, the proportion of queries where the language cannot be identified is much higher than for geographic entities or titles of work. This is mainly due to the fact that names of persons do not change across languages, but it is fundamental that a CLIR system recognizes these entities. Standardized name authority files such as PND (Personennamendatei) or ULAN (Union List of Artist Names) are essential to fulfill this task.

5 IP Address, Interface Language and Query Language

As demonstrated before, language identification is very hard to implement correctly in an automated manner.

It is therefore reasonable to incorporate other aspects in the language detection that could hint at the language the user is searching in. Especially the correlation between the query language, the corresponding IP address and the interface language is of interest. Of course, the IP address might not be reliable in every case. The same user may use several IP addresses or several users can share one IP address. Furthermore, it is possible to hide the true location by using proxies. We are also aware of the fact that users rarely switch the interface language and that many of them work with the default English interface.

5.1 Interface and Query Language

Interface language as a signal to detect the query language was also analyzed during the last LogCLEF track by Oakes & Xu [2]. They found that for the most common interface languages, namely German, French, English, Portuguese, Dutch and Italian, the most common query language was identical to the interface language.

Since we also flagged terms, which could not be assigned to a particular language a slightly different dataset resulted. Looking at the number of queries which were entered under the same interface language as the query language the proportion of the total number of queries in these languages is very small. This is probably due to the fact that not many users switch their interface language.

Table 8: Relationship between interface and query language

	ambiguous	en	fr	de	es	it	nl	pl	pt	ru	other	total
en	160	139	24	21	12	6	4	3	4	7	19	**399**
fr	4	4	5	0	1	1	0	0	0	0	1	**16**
de	4	3	0	2	0	0	0	0	0	0	0	**9**
es	2	2	2	0	1	1	0	0	1	0	0	**9**
it	5	1	2	1	0	1	0	0	0	0	0	**10**
pl	6	1	0	0	0	0	0	4	0	0	0	**11**
pt	1	2	1	0	1	0	0	0	1	0	0	**6**
ru	4	0	1	3	0	1	0	0	0	7	1	**17**
other	13	7	0	1	1	0	2	0	0	3	6	**33**
total	**199**	**159**	**35**	**28**	**16**	**10**	**6**	**7**	**6**	**17**	**27**	**510**

Table 8 shows the relationship between the selected interface language (rows) and the query language (columns). For better overview only part of the data is shown, missing languages are substituted under “other”. Under the French interface, 16 queries out of our sample were entered. These queries were French (5), English (4), Spanish (1), Italian (1), other (1) and 4 queries were entered, which could not be assigned to a specific language. French is the most common language under the French interface but looking at the whole set of queries, which were entered with a French user interface the proportion is relatively small.

5.2 IP Addresses and Query Language

For the shortened IP addresses, which were given in the log files, the respective country was identified. To make a statement about the relationship regarding country of origin and query language we determined the official language in these countries.

Table 9: Countries derived from IP addresses and their respective languages

Country	# of queries	Respective language
Germany	40	German
Italy	35	Italian
USA	33	English
France	32	French
Russia	28	Russian
Netherlands	26	Dutch
Poland	26	Polish
Spain	25	Spanish
United Kingdom	20	English
Austria	9	German

Table 9 shows the countries where most queries originated from according to their IP address and the official languages spoken in these countries.

Table 10: Relationship between languages spoken in countries derived from IP and query language

	ambiguous	en	fr	de	es	it	nl	pl	pt	ru	other	total
English	23	30	4	1	4	3	0	0	0	0	3	**68**
French	10	7	12	2	1	1	0	0	1	1	1	**36**
German	20	17	0	7	3	0	0	1	0	0	2	**50**
Greek	2	7	0	0	0	0	0	0	0	1	3	**13**
Spanish	15	9	4	0	5	1	0	0	1	0	2	**37**
Italian	15	8	3	4	0	1	0	0	1	0	3	**35**
Dutch	12	5	2	2	0	1	3	0	0	1	0	**26**
Polish	12	6	0	0	0	0	0	6	0	1	1	**26**
Portuguese	3	7	1	0	0	0	0	0	3	0	0	**14**
Russian	11	2	2	4	0	1	0	0	0	6	2	**28**
other	76	61	7	8	3	2	3	0	0	7	10	**177**
total	**199**	**159**	**35**	**28**	**16**	**10**	**6**	**7**	**6**	**17**	**27**	**510**

The rows of table 10 show the languages spoken in the countries where queries originated from. This signal seems to be less strong than looking at the interface language. Looking for example at the 50 queries originating from German speaking countries[5] like Germany or Austria, only 7 were German whereas the other languages of these queries were English (17), Spanish (3), Polish (1), other languages (2) and 20 queries, which could not be assigned to a single language.

[5] Included countries are: Germany, Austria and Liechtenstein. We excluded countries with several spoken languages such as Switzerland and Belgium.

6 Conclusion

The correct query language identification is decisive for language-dependent retrieval, the disambiguation and translation of query terms. It is also used in many retrieval systems to determine the language of the results presented. Our analysis shows that search query language identification and named entity recognition need to come together especially within a cultural heritage context. Many queries are expressing a search for proper names of persons, geographic entities or titles of work. Most of these queries cannot be assigned to a certain language. We also showed that signals commonly assumed to give indications about a user‘s preferred language are not as strong as expected.

The retrieval system, however, should be able to identify named entities and language preferences and be able to present the users results in a language they can understand and enable them to judge the relevance of the documents. More research is therefore needed not only on language detection, a problem that might not be solved entirely – but also on named entities and their presentation in the search process.

References

1. Mandl, T., Agosti, M., Nunzio Di, G., Yeh, E., Mani, I., Doran, C., Schulz, J.M.: LogCLEF 2009: the CLEF 2009 Multilingual Logfile Analysis Track Overview. In: Working Notes of the Cross Language Evaluation Forum (CLEF). (2009)
2. Oakes, M.P., Xu, Y.: A Search Engine Based on Query Logs, and Search Log Analysis at the University of Sunderland, In: Working Notes, LADS Workshop, Cross-Language Evaluation Forum 2009. (2009)
3. Bosca A., Dini,L.: Cacao Project at the LOGCLEF Track, In: Working Notes, LADS Workshop, Cross-Language Evaluation Forum 2009. (2009)
4. Ghorab, M. R., Leveling, J., Zhou, D., Jones, G. J. F., and Wade, V.: TCD-DCU at LogCLEF 2009: An Analysis of Queries, Actions, and Interface Languages. In: Working Notes, LADS Workshop, Cross-Language Evaluation Forum 2009. (2009)
5. Lamm, K., Mandl, T., Kölle, R.: Search Path Visualization and Session Performance Evaluation with Log Files from The European Library. In: Working Notes, LADS Workshop, Cross-Language Evaluation Forum 2009. (2009)
6. Hofmann, K., de Rijke, M., Huurnink, B., Meij, E.: A Semantic Perspective on Query Log Analysis. In: Working Notes, LADS Workshop, Cross-Language Evaluation Forum 2009. (2009)
7. Jansen, B.J., Pooch, U.: A Review of Web Searching Studies and a Framework for Future Research. Journal of the American Society for Information Science and Technology. 52(3), 235-246 (2000)
8. Spink, A., Jansen, B.J., Wolfram, D., Saracevic, T.: From E-Sex to E-Commerce: Web Search Changes. Computer. 35(3), 107-109 (2002)
9. Jones, S., Cunningham, S.J, Mcnab, R.: An Analysis of Usage of a Digital Library. In: Proceeding of Second European Conference on Digital Libraries, 261-277 (1998)
10. Broder, A.: A Taxonomy of Web Search. SIGIR Forum, 36(2), 3-10 (2002)
11. Rose, D. E., Levinson, D.: Understanding User Goals in Web Search. In: WWW ‘04: Proceedings of the 13th International Conference on World Wide Web, 13–19. New York, NY, USA: ACM (2004)

12. Jansen, B. J., Spink, A., Bateman, J., Saracevic, T.: Real Life Information Retrieval: A Study of User Queries on the Web. SIGIR Forum: A Publication of the Special Interest Group on Information Retrieval, 32, 5-18 (1998)
13. Baeza-Yates, R. A., Calderón-Benavides,L., González-Caro, C.N.: The Intention Behind Web Queries. In: Lecture Notes in Computer Science. Vol. 4209, String Processing and Information Retrieval. 13th International Conference, SPIRE 2006, Glasgow, UK, October 11 - 13, 2006, 98-109 (2006)
14. Kang, I., Kim, G.: Query Type Classification for Web Document Retrieval. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '03. ACM, New York, NY, 64-71 (2003)