UNED at iCLEF 2008: Analysis of a large log of multilingual image searches in Flickr

Víctor Peinado, Julio Gonzalo, Javier Artiles and Fernando López-Ostenero
NLP & IR Group, ETSI Informática, UNED
c/ Juan del Rosal, 16, E-28040 Madrid, Spain
victor@lsi.uned.es, javart@gmail.com, {julio, flopez}@lsi.uned.es

Abstract

In this paper, we summarise our analysis of the large log of multilingual image searches in Flickr provided to iCLEF 2008 participants. We have studied (a) correlations between the language skills of searchers in the target language and other session parameters, such as success (was the image found?), number of query refinements, etc.; (b) learning effects over time; (c) usage of specific cross-language search facilities and (d) users perceptions on the task (questionnaire analysis).

We have identified 5101 complete search sessions (searcher/target image pairs) in the logs provided by the organisation. Our analysis shows that when users have active competence in the target language, their success rate is 12% higher than if they do not know the language at all. If the user has passive competence of the language (i.e. can partially understand texts but cannot make queries), the success rate equals those with active competence, but at the expense of executing more interactions with the system.

The most remarkable learning effect is that users carry out fewer interactions when they are familiarised with the task and the system, keeping the success rate and the number of hints invariant. Finally, the usage of specific cross-language facilities (such as refining translations offered by the system) is low, but significantly higher than standard relevance feedback facilities, and is perceived as useful by searchers.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.4 [**Information Systems Applications**]: H.4.m Miscellaneous

General Terms

interactive information retrieval, cross-language information retrieval

Keywords

iCLEF, Flickr, log analysis, multilingual image search, user studies

1 Introduction

In this paper, we summarise our analysis of the large log of multilingual image searches in Flickr provided to iCLEF 2008 participants [1].

In this search log, every session consists of a searcher (a registered user with a profile that includes her native language and her proficiency in English, Spanish, Italian, German, Dutch

and French) and a target image (from the Flickr image database, annotated in one or more of that six languages). When the session starts, the user does not know in which language(s) the image is annotated. The interface provides facilities to perform queries simultaneously in up to six languages (via dictionary translation of query terms), to provide controlled relevance feedback (clicking on suggested terms and terms from the images found) and to refine the translations provided by the system (changing the selection of the system or adding new translations). The task is, therefore, a multilingual known-item retrieval task. If the user gives up, she can ask for hints; the first hint is the target language (which turns the task into bilingual or monolingual search, depending on the language profile of the user). The rest of the hints are keywords used to annotate the image, which is aimed at preventing users from being discouraged with difficult images.

The log consists of more than 5,000 search sessions by more than 200 users with a wide range of skills in the interface languages, coming from four continents. The size of this corpus permits studying the behaviour of users in a multilingual search scenario at a scale that had not been possible before.

The UNED team has focused on studying (a) correlations between the language skills of searchers in the target language and other session parameters, such as success (was the image found?), number of query refinements, etc.; (b) learning effects over time; (c) usage of specific cross-language search facilities and (d) users perceptions on the task (questionnaire analysis). This paper is a summary of our study.

The structure of the rest of the paper is as follows: Section 2 describes the process performed to regularise the logs and characterise each user's search sessions. In Section 3 we search for correlations between language skills of searchers and other parameters of the search sessions. In Section 4 we study learning effects over time. In Section 5 we report on other aspects of our study, focusing on the usage of cross-lingual refinement facilities and users' perceptions on the task. Finally, in Section 6 we draw some general conclusions.

2 Log Processing and Characterisation of the Search Sessions

We have processed the logs provided by the iCLEF organisation in order to identify and characterise search sessions. A search session starts when the user is given a target image and finishes when the user either finds the image or gives up and stops searching. In the meantime, the user may log out and log in (even several times) and, essentially, interact with the interface: launch queries, explore the rank of results, ask for hints, read descriptions associated to images, manipulate the translations suggested by the system and therefore improve her personal dictionary, etc.

Once search sessions are identified and open sessions are filtered out (those that were active when the log was produced or those that died because of user inactivity for more than 24 hours), we retained 5101 search sessions.

We have processed the logs to provide a rich characterisation of each session. The essential features are the user's profile (in particular her language skills), the use of the different interface facilities (including translation features), the session number (when was the image searched in the search history of the user), etc. We have also distinguished between the behaviour before and after the first hint, which is the language in which the image is annotated, because it represents the frontier between fully multilingual search (the image can be annotated in any of six languages) and bilingual or monolingual search.

See Appendix A for a comprehensive list of the features that we have extracted.

3 Analysis Considering Language Skills

In our first analysis we have divided search sessions in three groups: "active" is the group of sessions where the image was annotated in a language in which the user can read and write. Sessions in "passive" are those where the target language was partially understandable by the user, but the user could not make queries in that language (think, for instance, of French for most Spanish or Italian speakers). Finally "unknown" stands for images annotated in languages completely unfamiliar for the user. In our pool of sessions we found 2809 for active, 726 for passive and 1566 for unknown. These figures are large enough to reach quantitatively meaningful conclusions.

Table 1 shows average values for some session features, for each of these three groups. The most notable result is the degree of success (was the image found?) for each of the groups: Active and passive speakers successfully found the image 82% and 81% of the times. Users with no competence in the annotation language obtained 72%, performing 12% worse. It is somehow surprising that users which only have a passive knowledge of the target language perform as well as those with active knowledge, because the first group must necessarily use the translation capabilities of the system to express their query. The unknown group performs only 12% worse, which reveals a difference but not a large gap. Note that the translation capabilities of the interface were not optimal: they used only freely available dictionaries with some coverage gaps, and they were not tailored to the domain (the Flickr database).

Note that, as users could ask for hints, it can be the case that "passive" cases reach the same success as "active" ones because they simply ask for more hints. This is not the case: the average number of hints hardly varies between the three groups, ranging from a minimum of 2.13 hints per session to a maximum of 2.18.

competence	**success**	**# hints**	**# queries**		**refinements**		**ranking exploration**	
			mono	**multi**	**mono**	**multi**	**mono**	**multi**
active	82%	2.13	3.96	3.8	3.91	3.76	2.58	2.69
passive	81%	2.18	4.32	4.05	4.27	4.01	3.4	3.14
unknown	72%	2.15	4.5	4.38	4.41	4.28	2.93	3.06

Table 1: User's behaviour according to language skills

The rest of the columns of Table Table 1 show the average number of queries launched, direct query refinements and the number of times that the user explored the ranking beyond the first page of results (containing 20 items).

In general, there is a clear ordering between active, passive and unknown sessions: active sessions need less interactions, passive more, and unknown even more. For instance, the average number of queries posed in the multilingual search mode is 3.8 for active sessions, 4.06 for passive sessions, and 4.38 for unknown sessions. Therefore, passive sessions achieve similar success than active sessions, but with a higher effort. Unknown sessions have even higher effort, but still with a 12% loss in effectiveness. The only feature in which this tendency is broken is with rank exploration: passive sessions tend to explore the rank further than unknown sessions, perhaps because the textual information in the images can be more easily used to do relevance feedback.

Note that we have not included search time in the tables. Although the logs provide time stamps, we have discarded them because there is no way of knowing when the user was actively engaged in the task or performing some other task while the session remained open. Therefore, time is less reliable as an activity indicator than the number of interactions with the system.

4 Analysis Considering Time. The Learning Effect

For many users, this might be the first occasion in which they search simultaneously in several languages. It seems interesting to check if there is a learning effect once they have search for a certain number of images and they are familiarised with the possibilities and difficulties of searching

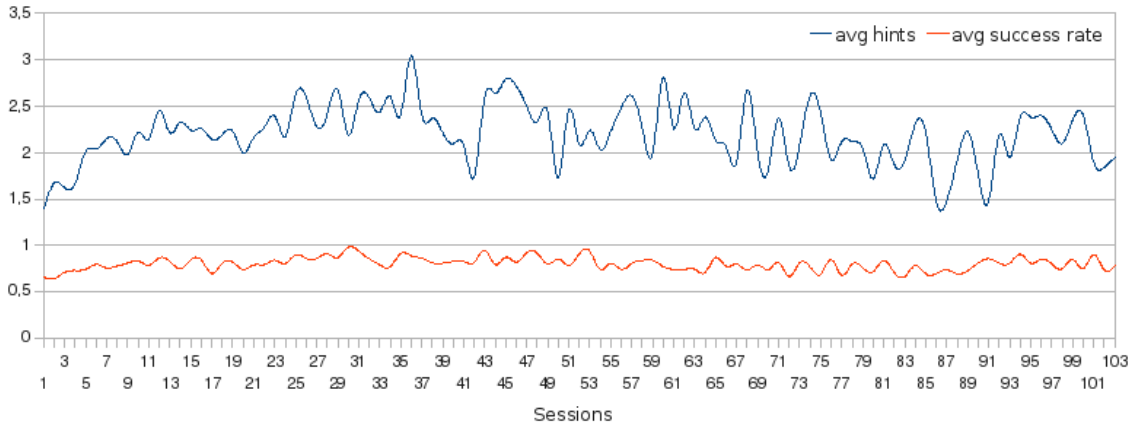

Figure 1: Average hints requested and success rate per sessions performed.

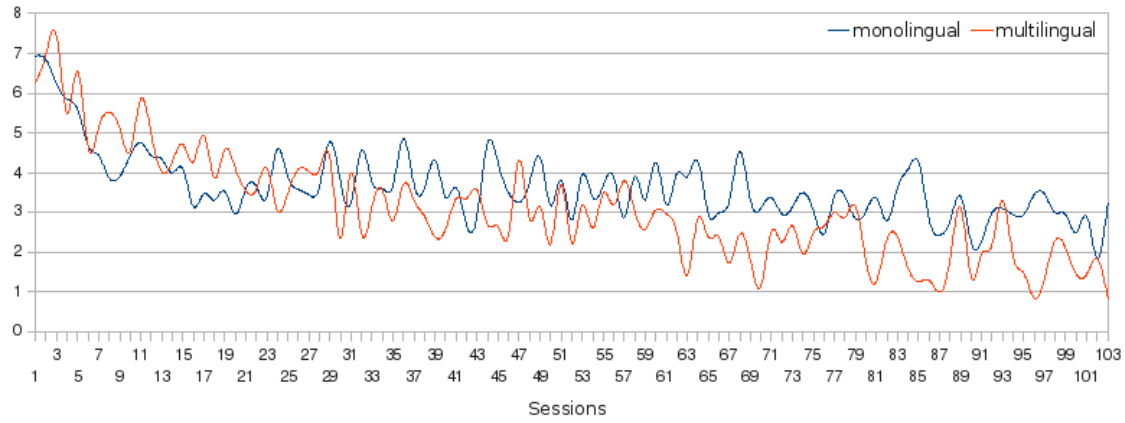

Figure 2: Average queries launched per sessions performed.

in a cross-language setting. For instance, will they learn to refine the translations chosen by the system when they are not appropriate?

With this goal in mind, we have also analysed users' behaviour considering the number of search sessions completed, and extracted some trends which are depicted in the following figures. First, Figure 1 shows how the success rate remains stable regardless of the number of sessions performed previously. Also the number of hints requested, in spite of showing a wider variability, remains stable too.

The next figures show that, the more time users spend interacting with the search engine, three features decrease: the average number of queries launched (see Figure 2), the average number of direct query refinements (see Figure 3) and the average number of ranking exploration beyond the first page of results (see Figure 4) decrease, both in the monolingual and the multilingual environments. In other words, it takes users less effort to find the images.

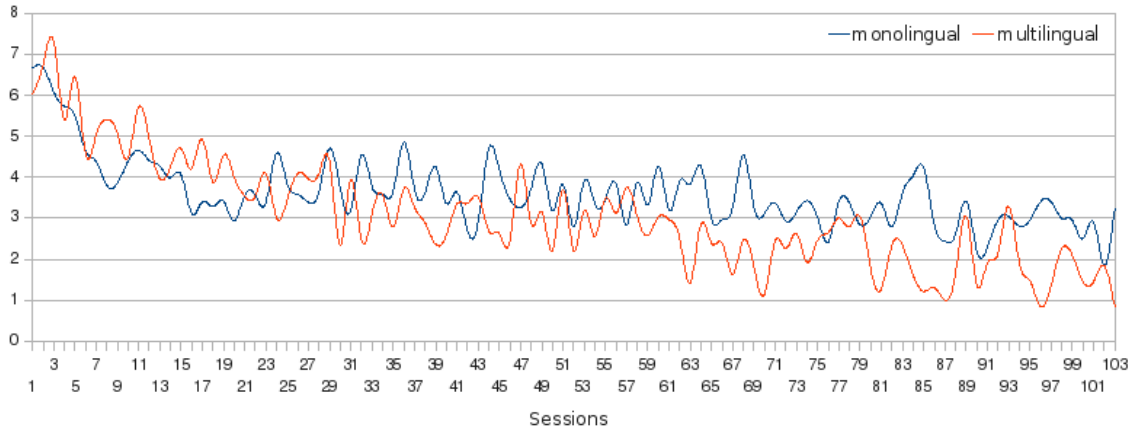

Figure 3: Average direct refinements per sessions performed.

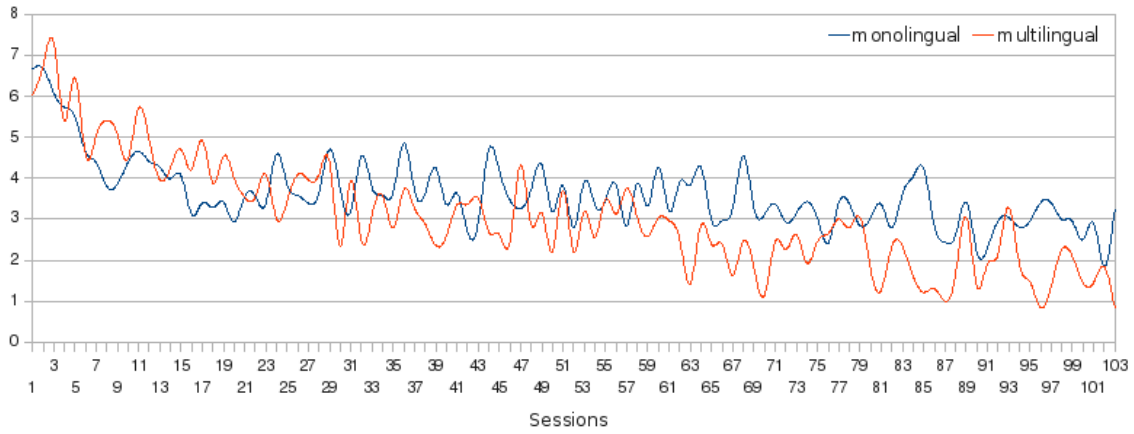

Figure 4: Average navigations beyond the first page of results per sessions performed.

5 Other issues

5.1 Usage of Specific Cross-Lingual Refinement Facilities

FlickLing[2] search interface provides some functionalities which take advantage of some of the Flickr's services[1]. Flickr services suggest new terms related to a given query and FlickLing allows to use these terms to launch a new query or to refine a previous one. This functionality was only used by less than 2% of users, as shown in Table 2. This is in agreement with the common place that relevance feedback facilities are rarely used in search engines (at least in non-specialised search scenarios), even if they can provide more search effectiveness.

The percentage of queries where users change the translations chosen by the system or add new translations to the dictionary is also quite low in absolute terms, but it is relevant if we compare it with the use of standard relevance feedback mechanisms. Usage ranges from 4% to 11% are much higher than relevance feedback (2%) and can be taken as a positive indication of their usefulness at certain points of the search process.

competence	**Flickr related terms**				usage of **personal dictionary**	
	new query		**query refinement**			
	mono	**multi**	**mono**	**multi**	**manipulations**	**adding new terms**
active	1%	1%	2%	2%	4%	7%
passive	1%	1%	3%	1%	6%	9%
unknown	2%	1%	4%	5%	9%	11%

Table 2: Usage of Specific Cross-Lingual Refinement Facilities

5.2 User Perceptions on the Task

Although the primary source of information are the activity logs of the users, the logs also collect the answers to two types of questionnaires: one is presented after each session (in two forms: one if the search failed and another one if the search succeeded), and another one is presented only once, when the user has performed fifteen search sessions (and therefore has a rather complete overall impression of the task).

5.2.1 Post-session questionnaires

Let us start with the results of post-session questionnaires, which are depicted in Figure 5 (after success questionnaire) and Figure 6 (after failure).

In cases of success, the task is perceived as easy in more than 1800 cases, and hard in over 2200 cases. The two most popular sources of difficulty are not related with the cross-language nature of the task: "it was difficult to describe the query" and "it was hard because of the size of the image set". Not knowing the target language is mentioned as a difficulty in slightly over 400 cases, and bad translations in around 350 cases. 200 answers thought that having to translate the query was a problem.

In cases of failure, the pattern is very similar: the two most common cases of failure are "I can't find suitable keywords for this image" and "There are too many images for my search".

Overall, the perception of users is that multilinguality is a difficulty, but not as relevant as other aspects of the search task.

5.2.2 Final questionnaires

These questionnaires are not exactly answered when the experience is over, but when the user has performed fifteen search sessions and is therefore familiarised with the task. These are the results:

[1]See `http://www.flickr.com/services/api` for further information about Flickr API.

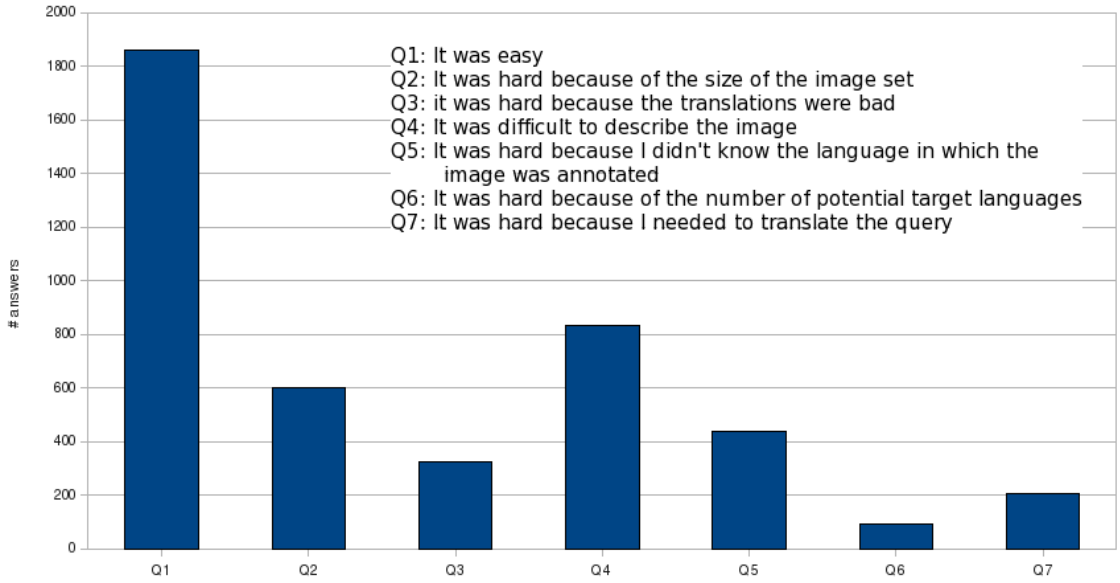

Figure 5: Post-image questionnaires after finding an image: overall results.

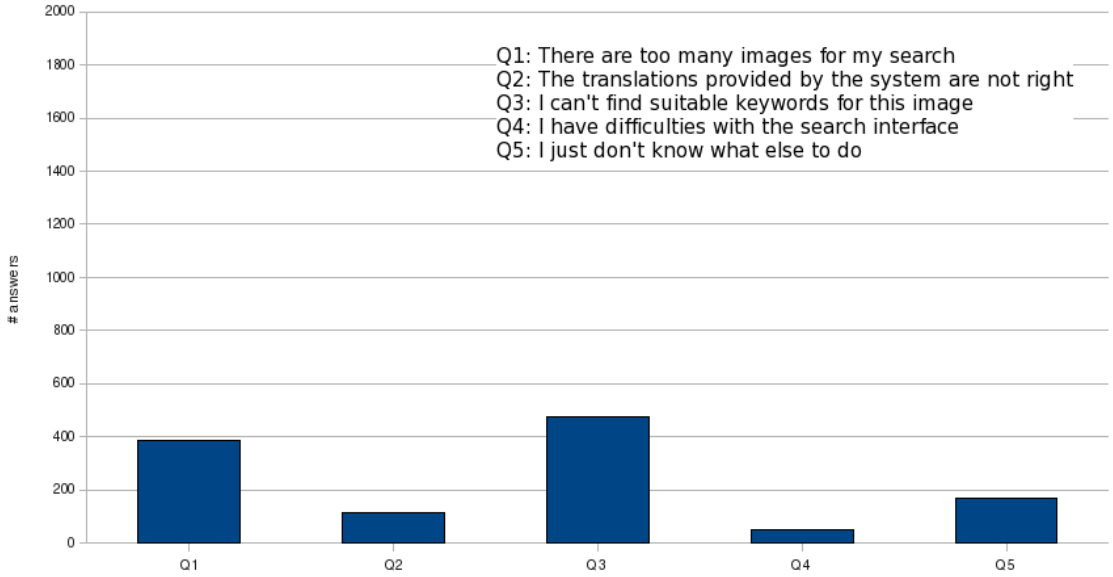

Figure 6: Post-image Questionnaires after giving up an image: overall results.

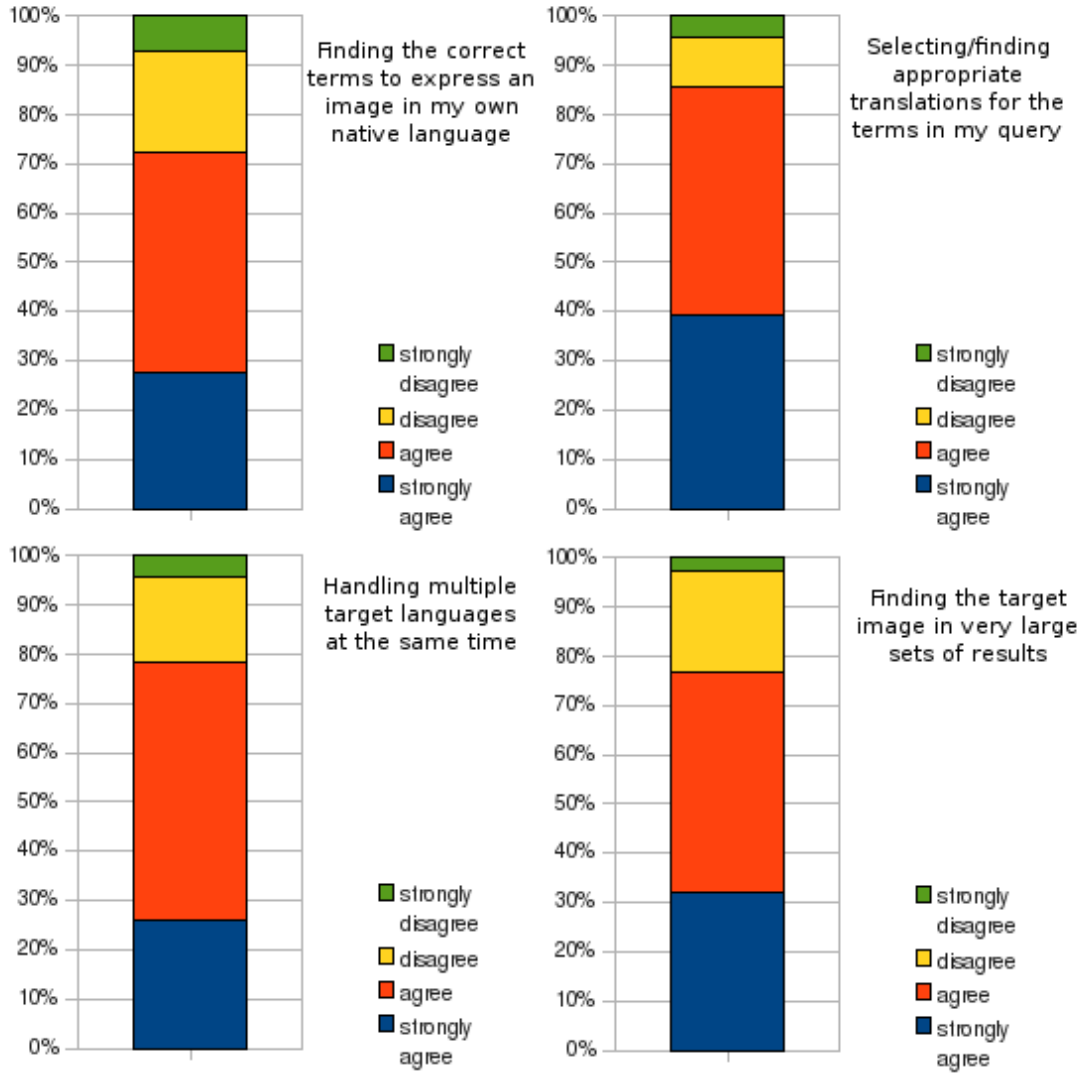

Figure 7: Which, in your opinion, are the most challenging aspects of the task?

Challenging aspects of the task

Which, in your opinion, are the most challenging aspects of the task? Figure 7 shows the answers to this question. Notably, when we restrict this question to experienced users, which has searched at least for fifteen images, the results change drastically. Over 85% of the users agree or strongly agree that "Selecting/finding appropriate translations for the terms in my query" is a challenging aspect of the task, which makes it the most challenging aspect.

Interface facilities

Two questions were addressed at how users perceived the interface facilities. The first one is "Which interface facilities did you find most useful?" and the results are depicted in Figure 8. Note that cross-language facilities (automatic translation of query terms and possibility of improving the translations chosen by the system) are much more valued than standard feedback facilities (the assistant to select new terms from the set of results and the additional query terms suggested by Flickr). This is in agreement with the proportional usage of these two kinds of facilities, although we must remark that the actual usage of those facilities is lower than what would be expected

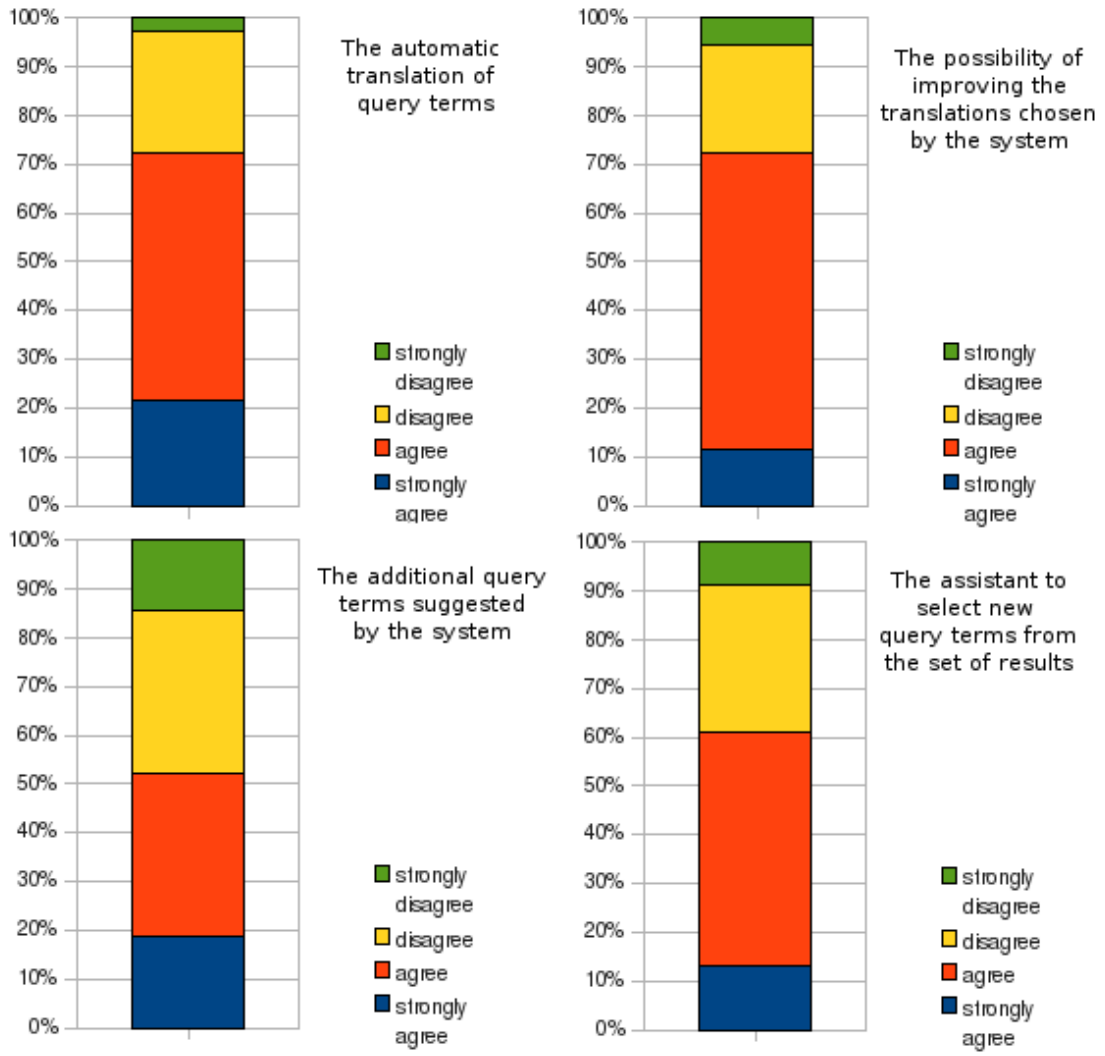

Figure 8: Which interface facilities did you find most useful?

from the questionnaire. Do users learn that they are useful late in the search history? According to our study of learning effects, this is not the case.

The second question is "Which interface facilities did you miss?", with a list of explicit facilities. The results are shown in Figure 9. Three facilities have an agreement rate (agree or strongly agree) above 70%: "a system able to select the translations for my query better", "The classification of search results in different tabs according to the image caption language", and "the possibility to search according to the visual features of the image". Other choices have slightly lower agreement rates: "an advanced search mode giving more control on how Flickr is queried", "bilingual dictionaries with a better coverage", and "more support to decide what the possible translations mean and therefore which ones are more appropriate". The least valued option (yet with an agreement rate above 50%) is "detection and translation of multi-word expressions", perhaps due to the nature of the task and the annotations (tags are frequently single words).

It is difficult to extract conclusions from the answers to this question, apart from the fact that users seem to appreciate all features that can seemingly improve the search experience, even if interactive features are not frequently used in practice.

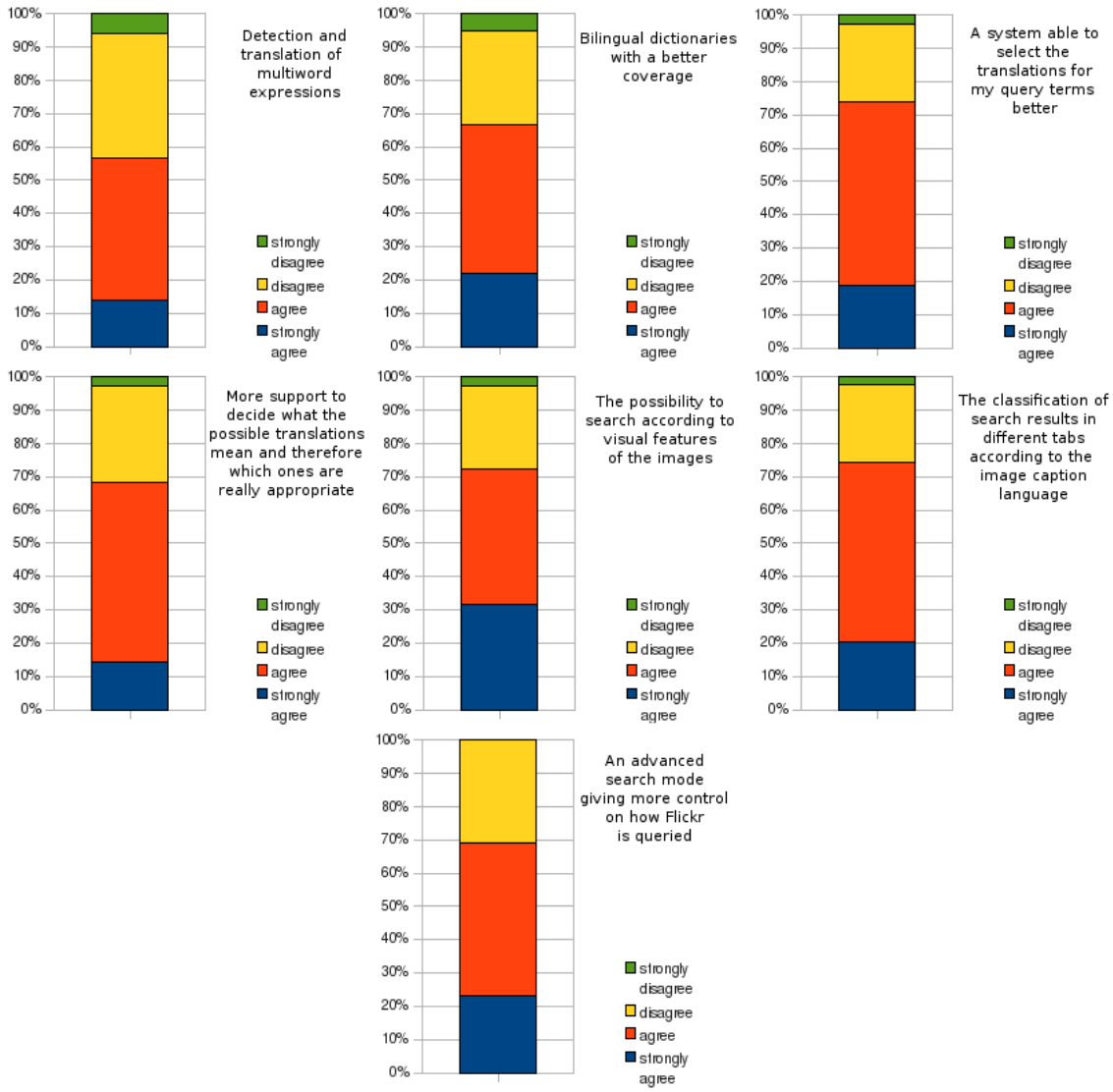

Figure 9: Which interface facilities did you miss?

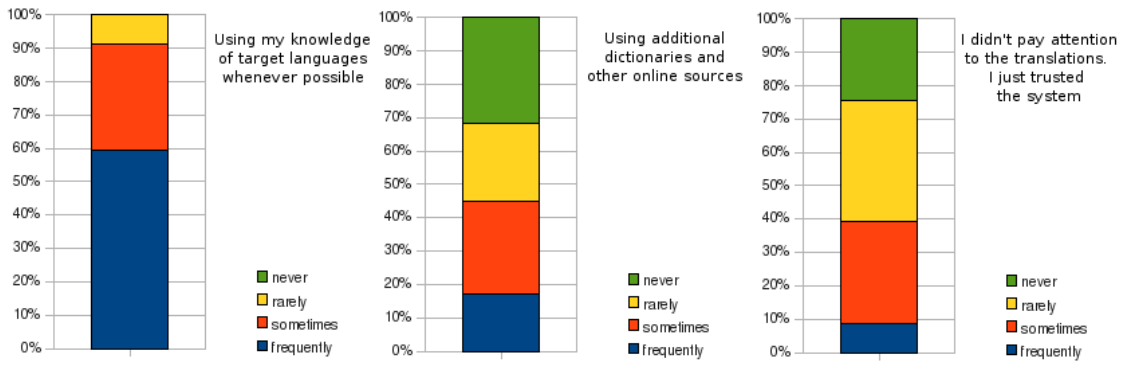

Figure 10: How did you select/find the best translations for your query terms?

Finding appropriate translations

The last question was "How did you select/find the best translations for your query terms?". By far the most popular answer is "using my knowledge of target languages whenever possible", which was frequently used by 60% of the users and sometimes by another 30%. In contrast, less than 10% frequently "did not pay attention to the translations. I just trusted the system". This is in sharp contrast with the average behaviour of users, which rarely modify the translations chosen by the system, and deserves further investigation. Finally, "using additional dictionaries and other online sources" is used frequently by less than 20% of the users, and "sometimes" by another 20%.

6 Conclusions

The search logs under study in the iCLEF 2008 task provide a more solid base to extract conclusions about the behaviour of users in multilingual search scenarios than most previous experiments, which were mostly performed under laboratory conditions and therefore more restricted in size.

At UNED we have identified 5101 complete search sessions (searcher/target image pairs) in the logs provided by the organisation. Our analysis shows that when users have active competence in the target language, their success rate is 12% higher than if they do not know the language at all. If the user has passive competence of the language (i.e. can partially understand texts but cannot make queries), the success rate equals those with active competence, but at the expense of executing more interactions with the system.

The most remarkable learning effect is that users carry out fewer interactions when they are familiarised with the task and the system, keeping the success rate and the number of hints invariant. Finally, the usage of specific cross-language facilities (such as refining translations offered by the system) is low, but significantly higher than standard relevance feedback facilities, and is perceived as useful by searchers.

Finally, the perception of experience users about cross-language retrieval interactive facilities is very positive, in spite of the fact that they are not frequently used. This is an indication that advanced search features - in this case, manipulation of translations offered by the system - might not be used frequently, but when they are used they become critical for the success of the task. A consequence is that query translation assistance should be hidden in the default settings of a cross-language search interface, but should be possible to invoke it for certain advanced users or specific search situations.

Acknowledgements

This work has been partially supported by the Regional Government of Madrid under the MAVIR Research Network (S-0505/TIC-0267) and the Spanish Government under project Text-Mess

(TIN2006-15265-C06-02).

References

[1] Gonzalo, J., Clough, P., Karlgren, J.: Overview of iCLEF 2008: search log analysis for Multilingual Image Retrieval. This volume.

[2] Peinado, V., Artiles, J., Gonzalo, J., Barker, E., López-Ostenero, F.: FlickLing: a multilingual search interface for Flickr. This volume.

A Characterization of the Search Sessions

In order to characterize our users' behavior, we decided to represent every search session capturing the user's profile, the use of the translation capabilities, and the state of every single feature of the interface. In most of the cases, we emphasize a crucial moment in the development of the search session: the first hint reveling the language of annotation of the image, since it may greatly delimite the problem. Thus, we extracted the following 111 features.

```
## SEARCH SESSION AND USER PROFILE
1: user ID
2: user's mother language
3: interface language used
4: German is (active|passive|unknown)
5: English is (active|passive|unknown)
6: Spanish is (active|passive|unknown)
7: French is (active|passive|unknown)
8: Italian is (active|passive|unknown)
9: Dutch is (active|passive|unknown)
10: Session number for the individual user
11: target image ID
12: image language annotations
13: image language is is (active|passive|unknown)
14: hints requested
15: success of failure in the session

## MONOLINGUAL INTERFACE
16: queries
17: queries before asking the first hint
18: queries after asking the first hint
19: direct query refinements
20: direct query refinements before asking the first hint
21: direct query refinements after asking the first hint
22: query refinements from a related term suggested by Flickr
23: query refinements from a related term suggested by Flickr before asking the first hint
24: query refinements from a related term suggested by Flickr after asking the first hint
25: query refinements by adding a related term to a previous query
26: query refinements by adding a related term to a previous query before asking the first hint
27: query refinements by adding a related term to a previous query after asking the first hint
28: query refinements from an image tag
29: query refinements from an image tag before asking the first hint
30: query refinements from an image tag after asking the first hint
31: query refinements by adding an image tag to a previous query
32: query refinements by adding an image tag to a previous query before asking the first hint
33: query refinements by adding an image tag to a previous query after asking the first hint
34: exploration of the ranking beyond the first page (20 results)
35: exploration of the ranking beyond the first page (20 results) before asking the first hint
```

36: exploration of the ranking beyond the first page (20 results) after asking the first hint
37: clicks to enlarge an image
38: clicks to enlarge an image before asking the first hint
39: clicks to enlarge an image after asking the first hint
40: wrong clicks on "I found it" icon
41: wrong clicks on "I found it" icon before asking the first hint
42: wrong clicks on "I found it" icon after asking the first hint
43: clicks on "show description" icon
44: clicks on "show description" icon before asking the first hint
45: clicks on "show description" icon after asking the first hint

MULTILINGUAL INTERFACE
46: queries
47: queries before asking the first hint
48: queries after asking the first hint
49: direct query refinements
50: direct query refinements before asking the first hint
51: direct query refinements after asking the first hint
52: query refinements from a related term suggested by Flickr
53: query refinements from a related term suggested by Flickr before asking the first hint
54: query refinements from a related term suggested by Flickr after asking the first hint
55: query refinements by adding a related term to a previous query
56: query refinements by adding a related term to a previous query before asking the first hint
57: query refinements by adding a related term to a previous query after asking the first hint
58: query refinements from an image tag
59: query refinements from an image tag before asking the first hint
60: query refinements from an image tag after asking the first hint
61: query refinements by adding an image tag to a previous query
62: query refinements by adding an image tag to a previous query before asking the first hint
63: query refinements by adding an image tag to a previous query after asking the first hint
64: exploration of the ranking beyond the first page (20 results)
65: exploration of the ranking beyond the first page (20 results) before asking the first hint
66: exploration of the ranking beyond the first page (20 results) after asking the first hint
67: clicks to enlarge an image
68: clicks to enlarge an image before asking the first hint
69: clicks to enlarge an image after asking the first hint
70: wrong clicks on "I found it" icon
71: wrong clicks on "I found it" icon before asking the first hint
72: wrong clicks on "I found it" icon after asking the first hint
73: clicks on "show description" icon
74: clicks on "show description" icon before asking the first hint
75: clicks on "show description" icon after asking the first hint

LANGUAGES ENABLED AND PERSONAL DICTIONARY
76: German as target language
77: German as target language before asking the first hint
78: German as target language after asking the first hint
79: English as target language
80: English as target language before asking the first hint
81: English as target language after asking the first hint
82: Spanish as target language
83: Spanish as target language before asking the first hint
84: Spanish as target language after asking the first hint
85: French as target language
86: French as target language before asking the first hint
87: French as target language after asking the first hint
88: Italian as target language
89: Italian as target language before asking the first hint

90: Italian as target language after asking the first hint
91: Dutch as target language
92: Dutch as target language before asking the first hint
93: Dutch as target language after asking the first hint
94: modifications of the translations suggested by the system
95: modifications of the translations suggested by the system before asking the first hint
96: modifications of the translations suggested by the system after asking the first hint
97: new translations added to the personal dictionary
98: new translations added to the personal dictionary before asking the first hint
99: new translations added to the personal dictionary after asking the first hint

POST-IMAGE QUESTIONNAIRES AFTER SUCCESS
100: It was easy
101: It was hard because of the size of the image set
102: It was hard because the translations were bad
103: It was difficult to describe the image
104: It was hard because I didn't know the language in which the image was annotated
105: It was hard because of the number of potential target languages
106: It was hard because I needed to translate the query

POST-IMAGE QUESTIONNAIRES AFTER GIVING UP
107: There are too many images for my search
108: The translations provided by the system are not right
109: I can't find suitable keywords for this image
110: I have difficulties with the search interface
111: I just don't know what else to do