Evaluating the Impact of Personal Dictionaries for Cross-Language Information Retrieval of Socially Annotated Images

Diana Irina Tanase, Epaminondas Kapetanios
School of Computer Science,
University of Westminster,
London, UK

Abstract

These working notes focus on the users' actions in order to assist translations and on the usage of personal dictionaries (a feature which enables saving user added words). The special interest for this feature comes from a need to investigate to what extent users get actively involved in the query translation and contribute to overcoming the limitations of automatic translations. It is also our hope that by understanding the relationship between user language skills and the usage of the personal dictionary feature in the iCLIR context, we will be able to get at least a partial answer to a bigger question regarding collaborative translations in today's participatory web space.

ACM Categories and Subject Descriptors: H.3.5 [Online Information Services]: Web-based services; H.3.3 [Information Search and Retrieval]: Search process; General Terms: Experimentation
Free Keywords: personal dictionary, cross language information retrieval

1. Introduction

This year's iCLIR 2008 challenge was set to meet the need for a large-scale experiment, in a realistic naturally multilingual environment opened to all users in the web community. It comes as a follow-up to the iCLIR 2006 experiments that concentrated on developing a CLIR interface for Flickr (a widely used photo and video sharing service) and on devising evaluation measures to gauge the user experience. The 2006 task was completed by researchers from three universities, each investigating different aspects of the user-system interaction. The UNED group assessed how users with different language skills interacted with the system and if the CLIR facilities provided were used. The Swedish Institute of Computer Science group focused on user's perceptions namely satisfaction, completeness, and quality, while the University of Sheffield group focused both on user's behavior and search effectiveness.

Their results unveiled a certain level of reticence in using the assisted query translation functionality, more so when the target language was unknown, with automatic translation being favored by users. All experiments were run with relatively small numbers of users (less than 25) and with homogeneous types of users in terms of language skills.

To be able to draw broader conclusions regarding users' interaction with a CLIR system, this year's CLEF organizers made available for its participants a set of logs from a specially developed system with a CLIR front-end to the Flickr database. This system enabled monolingual and multilingual searches through a set of 180 images. The images were annotated in one or several of the following languages: English, Spanish, German, French, Dutch, and Italian, and it was promoted as a game, made available to any interested user. Hence, this year's experimental setting has a unique character by allowing researchers to learn from a larger and heterogeneous user group.

In these working notes, the focus is set on the users' actions in order to assist translations and also on the usage of personal dictionaries (a feature which enables saving user added words). The special interest for this feature comes from a need to investigate further to what extent users get actively involved in the query translation and contribute to overcoming the limitations of automatic translations. Previous research in (Petrelli 2006, Oard 2008) has acknowledged that a modern iCLIR system should incorporate the necessary functionality to allow users to type their own translations. Comparative studies have been set up to gauge a user's willingness to supervise the translation step by selecting or deselecting from a list of potential translations. Results indicated that in supervised mode, when users verify and refine the translated query, the system has performed better than in delegated mode, when users do not intervene in the translation process. Though differences were not statistically significant in terms of precision and recall. It was also discovered that the supervised mode helped some of the users to reformulate their initial query based on suggested translations (Petrelli 2006). This search pattern was observed by the experiments with MIRACLE (He 2007).

It is also our hope that by understanding the relationship between user language skills and the usage of the personal dictionary feature in the iCLIR context, we will be able to get at least a partial answer to a bigger question regarding collaborative translations in today's participatory web space. If ad-hoc web communities form to share information, communicate on events, stories, things to-do, and overall facilitate each other to find and identify relevant resources, *would it be possible to trigger such informal collaboration between users with different language skills and facilitate the identification of relevant web resources regardless of the document language*? A starting point in investigating this question is determining to what extent users do take a participatory role in creating personal dictionaries and what is the quality of these dictionaries. Ideally, these customized language resources could be shared inside multilingual web communities and capture up-to-date usage of languages.

Until the above hypothesis gains more weight, we will focus on describing the context and the experimental setup (Section 2) that generated the logs, the research questions we tried to answer while analyzing and interpreting the logs (Section 3). We will conclude with a discussion (Section 4) of the answers crystallized from our iCLEF 2008 participation.

2. The Flickling Game

The system created for the 2008 evaluation (Flickling) implements a baseline set of functionalities for a CLIR system. It relies on the Flickr API for the image retrieval task and on a term-to-term translation mechanism. The latter relies mainly on a set of free dictionaries, and on a selection algorithm for "best" target translation based on i) Flickr's related terms for the query (often multilingual) and on ii) string similarity between the source and the target words (Clough 2008). It is worth mentioning that the dictionaries selected for this task have limited coverage, and uneven sizes. In other words, Flickling was equipped with a basic set of language resources.

The game entails finding images by determining the correct query terms to describe a given image and suitable translations for these query terms when needed. The clear search task and setup of this system, attracted around 300 users, but after filtering the data only 176 have actually played the game. The participating users filled in a short pre-game questionnaire specifying their mother language, the active languages (fluent writing and reading), passive languages (some level of fluency) and unknown languages. The results of these questionnaires show that the participants had a good range of language skills from monolingual to polyglots.

3. Analysis of Log File Data

The log files distributed to all the participants of this year's evaluation recorded step-by-step the user's actions. In our analysis we focused on extracting the entries that related to the users' language profiles, the interactions with the translation mechanism, the addition of new entries in the personal dictionary and on the overall user's results for the game.

Research Questions

1. Does the degree of confidence with a language affect usage and creation of personal dictionary entries, i.e., do those users with little knowledge of a language make use of the personal dictionary and to which extent?
2. Does the degree of confidence with a language affect quality of personal dictionary?
3. Can it be inferred that the user's performance in the game results improved by using the personal dictionary and/or the assisted translation mechanism?
4. Is the personal dictionary a useful interface facility?

Based on the active and passive language skills, we created a language skill coefficient that describes the number of active languages and the number of passive languages. For example a user that specified "EN" as the active and "FR, DE" as passive would be assigned a coefficient of 3 (a count of the known languages giving equal weight to active and passive languages). This generated five classes of users with the following distribution: 37% are in class 1, respectively class 3, 20% are in class 2, 4% are in class 4, and 2% in class 5. This coefficient is a measure of user's degree of confidence with languages needed to play the game and will be used to classify the answers to the research questions above.

Our explorations concerning the questions above, starts with an overall look into the results of two questionnaires posed to the user during the game whenever an image was found or its search abandoned. This will provide more insight on the challenges of the game. Tables 1 a) and b), as well as Tables 2 a) and b) reflect the results grouped by language coefficient and the challenges of the Flickling game.

Found Image Questionnaire

What problems did you encounter while searching for this image?
0: It was easy
1: It was hard because of the size of the image set
2: It was hard because the translations were bad
3: It was difficult to describe the image
4: It was hard because I didn't know the language in which the image was annotated
5: It was hard because of the number of potential target languages
6: It was hard because I needed to translate the query

Language Skill Coefficient	Q/A	Q0	Q1	Q2	Q3	Q4	Q5	Q6
1	True	416	229	109	211	184	27	116
	False	680	867	987	885	912	1069	980
2	True	699	234	148	369	162	36	61
	False	847	1312	1398	1177	1384	1510	1485
3	True	584	190	125	265	126	34	33
	False	618	1012	1077	937	1076	1168	1169
4	True	168	78	9	56	29	10	6
	False	178	268	337	290	317	336	340
5	True	61	4	5	26	24	4	6
	False	51	108	107	86	88	108	106

Table 1a. Raw data counts from the found image questionnaire

Language Skill Coefficient	Q0	Q1	Q2	Q3	Q4	Q5	Q6
1	37.95%	20.89%	9.94%	19.25%	16.78%	2.46%	10.58%
2	45.21%	15.13%	9.57%	23.86%	10.47%	2.32%	3.94%
3	48.58%	15.80%	10.39%	22.04%	10.48%	2.82%	2.74%
4	48.55%	22.54%	2.60%	16.18%	0.00083%	2.89%	1.73%
5	54.46%	3.57%	4.46%	23.21%	21.42%	3.57%	5.35%

Table 1b. Percentages of questions answered TRUE

The results above indicate that users found the image searching process easy (Q0), and the task became harder when the size of the image set was large (Q3) (see language coefficient 1 and 4). Another challenging aspect can be identified as the describing of the image itself. This stands out across all types of users. It is in essence a classic problem of information retrieval regardless of the multilingual aspect. Previous iCLEF evaluations did emphasize that query formulation and re-formulation have a strong impact on search results. Bad translations and not knowing the language in which the image was annotated were subsequent problems. The actual translation process was problematic for people with one active language.

Give Up Questionnaire

Why are you giving up on this image?
0: There are too many images for my search
1: The translations provided by the system are not right
2: I can't find suitable keywords for this image
3: I have difficulties with the search interface
4: I just don't know what else to do
5: Other (please, comment below)

Language Skill Coefficient	Q/A	Q0	Q1	Q2	Q3	Q4	Q5
1	True	170	41	166	13	47	14
	False	206	335	210	363	329	362
2	True	72	18	90	11	25	26
	False	150	204	132	211	197	196
3	True	91	43	160	19	65	22
	False	236	284	167	308	262	305
4	True	12	3	13	5	6	1
	False	23	32	22	30	29	34
5	True	16	0	22	1	2	0
	False	21	37	15	36	35	37

Table 2a. Raw data counts from the give up questionnaire

Language Skill Coefficient	Q0	Q1	Q2	Q3	Q4	Q5
1	45.21	10.90	44.14	3.45	12.5	3.72
2	32.43	8.10	40.54	4.95	11.26	11.71
3	27.82	13.14	48.92	5.81	19.87	6.72
4	34.28	8.57	37.14	14.28	17.14	2.85
5	43.24	0	59.45	2.70	5.40	0

Table 2b. Percentages of questions answered TRUE

The results of the give up questionnaire complement the previous results. It is apparent that regardless of the user profile in terms of language skills, it is the problem of describing the image with the correct keywords that determined users to abandon an image search. A second issue for all users was the set of images to search through. Users would acknowledge not knowing what to do next, to be more of a problem than dealing with bad translation. The difference for each group of users between the answers of each question is only marginal, but it surprisingly reflects that users that know only one language trusted the translation and did not point out at translations as a major problem.

Interactions – Assisted Translations and Personal Dictionary

The Flickling system was equipped with a straightforward interface that allowed users to look at the list of translated words, to add or remove words from the suggested translation, and also to add their own dictionary entries. Table 3 a) and b) describe in terms of distributions the actions taken by each group of users.

log / lang. coef.	show transuggestion	type new translation	add transuggestion	remove transuggestion	total interactions
1	778	214	45	62	1099
2	385	99	16	39	539
3	497	94	27	59	677
4	169	52	4	14	239
5	49	1	8	9	67

Table 3a. Assisted Translations and Personal Dictionary Interactions

log /lang coef	number of users / number of users that used the personal dictionary	show transuggestion	type new translation	add transuggestion	remove transuggestion
1	33 / 14	70.79%	19.47%	4.09%	5.64%
2	17 / 4	71.43%	18.37%	2.97%	7.24%

3	35 / 18	73.41%	13.88%	3.99%	8.71%
4	6 / 3	70.71%	21.76%	1.67%	5.86%
5	4 / 1	73.13%	1.49%	11.94%	13.43%

Table 3b. Another view on Assisted Translations and Personal Dictionary interactions

Based on the results from Table 3b the main type of interactions with the Flickling system is viewing the existing translations, followed by typing new translations. The interesting aspect of the data showing in this table is that users with fewer language skills were quite active in terms of adding new translations. The most skilled set of users were most active in selecting or deselecting words when translations were needed.

Using the observations from "type new translation" column (normalized by number of users) we computed the correlation between usage of the personal dictionary and language skill (-0.5946), which indicates a decreasing linear dependency between the two, and provides a negative answer to the first research question in this section regarding the correlation between language skills and involvement in working with the personal dictionary. This is a slightly surprising result, but it can be explained by experimental attitude towards adding their new dictionary entries.

We have also plotted a user vs. personal dictionary graph, to identify the overall trend in the user base (Figure 1). Out of the 94 users (~ 53%) that did use the personal dictionary, the mean number of interactions was 18, with a median of 5. We isolated the users that have added entries to their personal dictionary more than 18 times, but the scatter plots did not show a positive correlation between overall score, precision, time and the usage of the personal dictionary. This result may be due to the fact that we considered the overall personal dictionary interaction and not individual image searches.

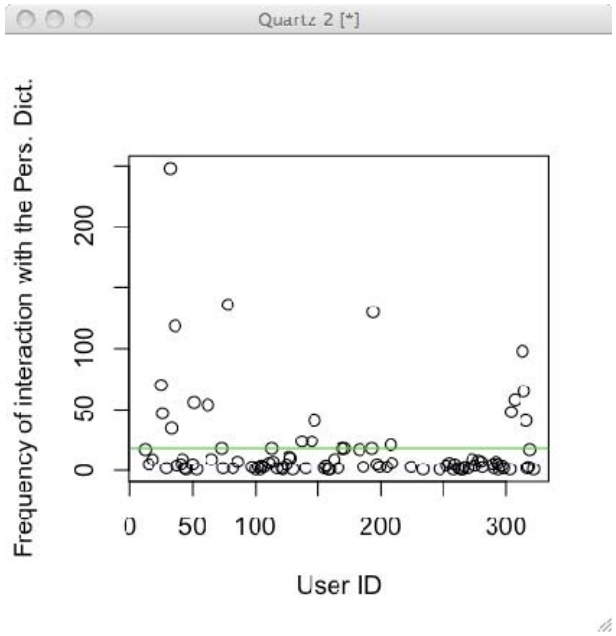

Figure 1. User vs. Personal Dictionary Interactions

Quality of the user added entries

The log recorded a total of 460 new entries to the Personal Dictionaries. By individually analyzing them we have noticed that there are several types of new entries. The most frequent are direct translations of the source query term, when there is no entry for it in the dictionaries. For the rest of the cases the users try to improve the provided translations list by adding synonyms ("antena" for "dish"), plural expressions ("anemone" for "anémona"), named entities ("London" for "Londres"), multiword expressions ("puente torre" for "tower bridge") or related concepts ("Africa" for "Uganda", "alligator" for "caiman").

Overall the contribution to the existing dictionaries was rather modest, and this can be explained by the fact images were annotated with words that existed in the dictionary and in very few cases there was a dictionary coverage problem, or by the fact almost 50% of the users confide completely in the automatic translation.

Also, due to an average number of just 18 entries per user, it is hard to assess an overall trend for each of the five groups of users in terms of the quality of the personal dictionary. In the context of this game, bad translations will quickly stand out, since judging relevance of the results set after a translation has been inputted is a quick visual task. Intuitively, the degree of confidence with a language would positively affect the quality of the personal dictionary.

Overall view of scores, precision, average time, and language skills

We have last to answer the third research question regarding the dependency between the user's performance in the game results and the interactions with the personal dictionary and/or the assisted translation mechanism. As seen in Figure 2, the language coefficient vs. distribution of translation related-actions shows a very weak correlation (correlation = -0.07266). For the second graph the correlation coefficient points to a medium strength between score and number of actions assisting translations (correlation = 0.3034) , while the third graph shows a very weak link between retrieval precision and number of actions assisting translations (correlation = 0.156).

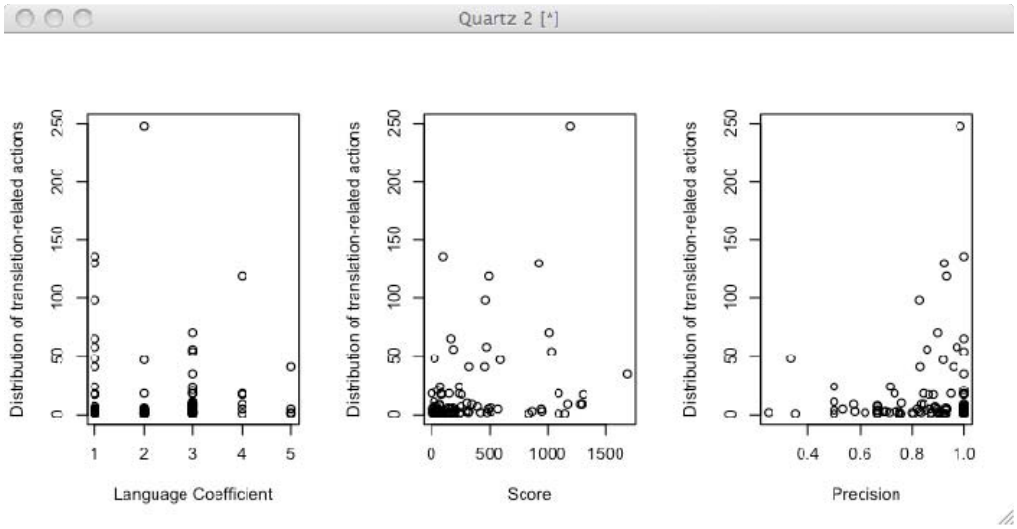

Figure 2. An overall view at language skills, scores, precision

These results support findings from previous research (Oard 2008) that acknowledged the user involvement in the translation as a positive interaction; as with all information retrieval, the quickest way to the relevant results is formulating a good query. This is in accordance with users' feedback from *Found Image Questionnaire* and *Give Up Questionnaire* that pointed at the difficulty in choosing good query terms for characterizing the searched image.

We conclude the presentation of these results, by looking at some of the answers from the overall questionnaire that was filled in by user after searching 15 images, and was completed by 63 users.

Overall Questionnaire

Out of the different questions comprised in this questionnaire, there are two main sets of questions that regard how translation was performed. The first question refers to the most useful interface facilities. The results show that automatic and assisted translations were perceived as equally important features. The second question investigates how the translations decisions are made. The dominant answers refer to using known languages or other language resources outside the game.

Which interface facilities were most useful?	Strongly Agree	Agree	Disagree	Strongly Disagree
A. *The automatic translation of query terms.*	13	31	17	2
B. *The possibility of improving the translations chosen by the system.*	7	38	14	4
C. *The additional query terms suggested by the system ("You might also want to try with... ").*	13	21	19	9
D. *The assistant to select new query terms from the set of results.*	9	28	18	6
How did you select the best translation for the query terms?	Frequently	Sometimes	Rarely	Never
A. *Using my knowledge of target languages whenever possible*	35	22	5	0
B. *Using additional dictionaries and other online sources.*	12	18	14	18
C. *I did not pay attention to the translations, I just trusted the system*	5	20	21	16

Table 4. Overall Questionnaire – Q7, Q9

4. Conclusions

Results obtained by previous iCLIR tasks at CLEF struggled to prove statistically the user's impact on an iCLIR system's precision and recall. The difficulty of making such assessments lies in the complexity of the interaction between a user and a CLIR system, and in defining suitable measures for interactive CLIR systems in general. During the log analysis, it became apparent that finding suitable measures for computing the different degrees of confidence in using a language is paramount for uncovering correlations. A finer grain analysis of user actions grouped by image search might have shed more light on the relationship between assisted translation, personal dictionary, and user's search performance.

Though we could not detect a clear link between usage of personal dictionaries and the efficiency of the gamers search, this year's experiment opens the path for furthering the research in using personalized language resources in the translation process.

Projects such as the Wiktionary, OmegaWiki, or Global WordNet are examples of global language resources that could prospectively be used for deploying large-scale CLIR systems. These resources are by design global and generic, and do not reflect the associated conceptualizations of specific groups of users. To compensate this aspect, a user's personal dictionary or maybe a community's custom dictionary, can keep translations in tune with a word's most frequent used sense or changes of meaning.

5. References

Gonzalo, J., Clough, P., Karlgren, J. (2008) Overview of iCLEF 2008: search log analysis for Multilingual Image Retrieval. In Borri, F., Nardi, A. and Peters, C., CLEF 2008 workshop notes.

He, D. and Wang, J. (2007). Information Retrieval: Searching in the 21st Century, chapter Cross-Language Information Retrieval. John Wiley & Sons.

Oard, D. W., He, D., and Wang, J. (2008). User-assisted query translation for interactive cross-language information retrieval. Inf. Process. Manage., 44(1):181–211.

Petrelli, D., Levin, S., Beaulieu, M., and Sanderson, M. (2006). Which user interaction for cross-language information retrieval? design issues and reflections. J. Am. Soc. Inf. Sci. Technol., 57(5):709–722.