

FITE-TRT: A High Quality Translation Technique for OOV Words

Ari Pirkola, Jarmo Toivonen*, Heikki Keskustalo, Kalervo Järvelin

Department of Information Studies
33014 University of Tampere, Finland
{ari.pirkola, heikki.keskustalo,
kalervo.jarvelin}@uta.fi

Institute of Signal Processing*
Tampere University of Technology
Tampere, Finland
jarmo.toivonen@cs.tut.fi

ABSTRACT

We devised a novel statistical technique for the identification of the translation equivalents of source words obtained by transformation rule based translation (TRT). The effectiveness of the devised FITE (frequency-based identification of translation equivalents) technique was tested using biological and medical cross-lingual spelling variants and OOV words in Spanish-English and Finnish-English TRT. For Spanish-English, translation recall was 89.2%-91.0% and for Finnish-English 71.9%-72.9%. For both language pairs FITE-TRT achieved high translation precision, i.e., 97.0%-98.8%. The technique also reliably identified native source language words, i.e., source words that cannot be correctly translated by TRT. Dictionary-based CLIR augmented with FITE-TRT performed substantially better than dictionary-based CLIR where OOV keys were kept intact.

Categories and Subject Descriptors

H.3.3 [Information Systems]: Information Search and Retrieval

General Terms

Algorithms, Performance, Experimentation

Keywords

Cross-language information retrieval, OOV words, TRT

1. INTRODUCTION

Out-of-vocabulary (OOV) words constitute a major problem in cross-language information retrieval (CLIR) and machine translation (MT). In those cases where equivalent terms in different languages are etymologically related technical terms (*cross-lingual spelling variants* - as German *konstruktion* and English *construction*) it is possible to use transliteration type of translation to recognize the target language equivalents of the source language words. In [5] we generated automatically large collections of character correspondences in several language

pairs for the translation of cross-lingual spelling variants. We call the regular correspondences augmented with statistical information *transformation rules* and the translation technique based on the generated rules *transformation rule based translation* (TRT).

It is obvious that a technique where words not found in a dictionary are translated by transformation rules would be useful in many information systems where automatic translation is part of the system. However, the TRT technique may be useless if it just indicates a set of translation equivalent candidates for a source word but is not able to indicate the one correct equivalent, which was the case in [5] as well as in [7]. In the present research we combat this problem, and move TRT from what is called fuzzy translation towards dictionary-like translation where for each source word either one translation equivalent rather than a set of words possibly containing the equivalent is indicated, or the source word is indicated not to be translatable by means of TRT. For this we developed a novel statistical equivalent identification technique called *frequency-based identification of translation equivalents* (FITE). The identification of equivalents is based on regular frequency patterns associated with the target word forms obtained by TRT.

In this paper we also present a novel feature of TRT, viz., translation through indirect translation routes. If a direct translation from a source language into a target language fails to find an equivalent the source word is retranslated into a target language through intermediate languages. As in the case of direct translation the equivalents are searched for from TRT's translation set by means of the novel FITE technique.

We study Spanish-English and Finnish-English TRT. For both language pairs German and French serve as intermediate languages. As test words we use terms in the domains of biology and medicine. The terms were selected from texts and real information requests of biomedical researchers.

The novel FITE-TRT technique is fundamentally different from other OOV methods/systems presented in the literature. For instance, Cheng et al. [1] and Zhang and Vines [8] both developed a Web-based translation method for Chinese-English OOV words where the OOV words were extracted from bilingual Chinese-English texts found in Chinese Web pages using word co-occurrence statistics and syntactic structures. Fujii and Ishikawa [2] used character-based rules to establish mapping between English characters and romanized Japanese katakana characters. They also utilized probabilistic character-based language models, which can be seen as a variation of the fuzzy

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SAC'06, April, 23-27, 2006, Dijon, France.

Copyright 2006 ACM 1-59593-108-2/06/0004...\$5.00.

matching technique. The technique is different from FITE-TRT but bears some resemblance to fuzzy translation reported in [5], however focusing on languages with different orthographies and having thus different focus.

2. TRT AND TRANSFORMATION RULES

The idea of TRT and the automatic method to generate transformation rules is described in [5]. A *transformation rule* contains source and target language characters that are transformed and their context characters. In addition, there are two important numerical factors associated with a rule, i.e., frequency and confidence factor, which may be used as thresholds to select the most common and reliable rules for TRT. *Frequency* refers to the number of the occurrences of the rule in the dictionary data that was used for the rule generation. *Confidence factor* (CF) is defined as the frequency of a rule divided by the number of source words where the source substring of the rule occurs.

Below we present an example of a German-English rule:

ekt ect center 191 214 89.25

The rule is read as follows: the letter *k*, prior to *t* and after *e*, is transformed into the letter *c* in the center of words, with the confidence factor being 89.25% ($100\% * 191/214$).

Examples of target word forms obtained in TRT are shown in Section 4.3.

3. RESEARCH PROBLEMS AND EVALUATION MEASURES

We examine the following research questions:

- How to effectively identify the correct equivalent of a source word among the many word forms produced by TRT when most of the transformation rules available for a language pair are used in TRT?
- How to reliably identify native source language words, i.e., source words that cannot be correctly translated by TRT?
- What are the translation recall and precision and indication precision (see the definitions below) of the proposed FITE-TRT method?
- What is the contribution of each step in the FITE-TRT process to its overall effectiveness?
- What is the effectiveness of a standard CLIR system boosted by the use of FITE-TRT in comparison to standard CLIR and monolingual baselines?

The effectiveness of FITE-TRT was evaluated by using the measures of *translation recall*, *translation precision*, and *indication precision*. For spelling variants *translation recall* is defined as the proportion of source words for which FITE identifies correct equivalents among all source words and *translation precision* is defined as the proportion of correct equivalents among all words for which an equivalent is indicated. For native words the question what share of them are translated by TRT is an irrelevant question, and naturally recall

is not measured for them. For native source words *indication precision* is defined as the proportion of words (correctly) indicated to be untranslatable.

When incorporating FITE-TRT into a retrieval system, its document retrieval effectiveness was evaluated using the standard evaluation measures of *mean average precision* (MAP) and *precision at 20 documents* (e.g., <http://trec.nist.gov>).

4. METHODS AND DATA

4.1 Transformation Rule Collections

The transformation rules were generated using the rule generation method described in [5]. The total number of rules in the generated rule collections as well the number of rules at or above the applied thresholds of CF=4.0% and frequency=2 are presented in Table 1. We applied the confidence factor and frequency thresholds because TRT may give very large translation sets. At the present stage of development the TRT program is not efficient enough to process very large word sets.

As shown in Table 1, for Finnish-English there were two collections. The second collection was constructed because the first collection (which was constructed first) was missing many important rules.

4.2 Training and Test Word Sets and Translation by TRT

We used a *training word set* for the development of the FITE technique. The set contained the title words ($n=75$) of the Spanish CLEF topics numbered 91 to 109. In addition to native Spanish words the titles contain Spanish-English spelling variants, native English words and English acronyms. The effectiveness of FITE-TRT was evaluated using four sets of *test words*. For each source language word set there was a corresponding English word set that contained the equivalents of the source words.

The first two sets contained Spanish and Finnish spelling variants of English biological and medical terms ($n=89$) gathered from the index of CLEF's LA times collection. The Spanish terms formed the first and the Finnish terms the second test word set. These terms are called *bio-terms*.

The third test word set contained Spanish ($n=98$) and the fourth set Finnish ($n=53$) OOV keys of the Spanish/Finnish-English CLIR runs (Section 4.4). Among the OOV keys there were, in addition to spelling variants, native Spanish/Finnish words as well as English words and English acronyms. The difference in the number of the OOV words reflects the different sizes of the CLIR system's Spanish-English and Finnish-English dictionaries. These test word sets are here called *OOV-CLIR-SPA-ENG* and *OOV-CLIR-FIN-ENG*. The total number of unique source words translated by TRT was $89+98$ (Spanish) + $89 + 53$ (Finnish) = 329 words.

Words containing four or less letters were not translated by TRT. This restriction was set because the short words were English acronyms and they need not be translated. (Generally, acronyms cannot be translated by means of TRT which only handles spelling variants.) On the other hand, cross-lingual spelling

variants are not very short words. Within the four test word sets there were two short (4-letter) spelling variants, which were removed from the sets according to the short word restriction.

The source words were translated by a TRT program through direct and indirect translation routes using the thresholds described above.

In the translation sets word forms were sorted by their document frequencies (Section 4.3). In indirect translation only five top German and French forms in a translation set were further

translated into English. Different English translation sets representing the same German/French translation set were combined. Document frequencies for source words and the generated intermediate and target word forms were taken from the Web using Altavista's search engine which shows the number of retrieved documents and allows to retrieve in a desired language. The equivalents of source words were identified from TRT's translation sets by means of the FITE technique as described in Section 4.3.

Table 1. The number of rules in the rule collections.

Rule collection	# Rules	# Rules CF≥4.0%, Freq. ≥2
Spanish-English	8800	1295
Spanish-German	5412	984
Spanish-French	9724	1430
German-English	8609	1219
French-English	9873	1170
Finnish-English/collection 1	1582	557
Finnish-English/collection 2	5423	1229
Finnish-German	4686	1108
Finnish-French	3463	877

4.3 The FITE Technique

4.3.1 Frequency Pattern

The core of FITE is that except for the translation equivalents the word forms yielded by TRT are malformed rather than real words, or they are rare words, e.g., foreign language words in the target language text. The equivalents belong to a language's basic lexicon and are much more common in the language than the other word forms. This regular *frequency pattern* allows the identification of the equivalents.

The example in Table 2 shows the document frequency pattern associated with the word forms obtained by TRT for the Spanish word *biosintesis* in Spanish-English TRT.

The word forms are sorted by document frequency (DF). In $df_T(i)$ i refers to a word form and T to a target language collection. The DF value associated with a word form refers to the frequency of English Web pages that contain the word form. It is seen that the DF of *biosynthesis*, the equivalent of *biosintesis*, is remarkably higher than the DFs of the other the word forms. This type of frequency distribution is very common for word forms within a translation set of TRT. The magnitude of difference between the document frequency of the first word form ($df_T(1)$) and the document frequency of the second word form ($df_T(2)$), or between $df_T(2)$ and $df_T(3)$ (see Section 4.3.4) forms the basis of the equivalent identification. We used the coefficient value (the magnitude of difference) of 10 for the identification of equivalents.

4.3.2 Relative Frequency

There are situations where the highest DF is possessed by a word that is not the correct equivalent. For example, the source word may occur frequently in a target language collection and if TRT

fails to translate the source word it may appear at the first position in a translation set. (As a special case, a source word is always included in the translation set because source and target language words may be identical.)

Table 2. An example of generated word forms and their document frequencies $df_T(i)$ (partial).

Generated Word Form i	$df_T(i)$
biosynthesis	2 230 000
biosintesis	909
biosyntesis	634
biosinthesis	255
biosynthessis	3
biosintessis	0
biosinthesiss	0
biosyntessiss	0

As a solution for this problem, we compute *relative document frequency rel-df*, as follows: $rel-df(i, sw, T, S) = df_T(i) / (\alpha \times df_S(sw))$. In the formula i = target word form; sw = source word; T = target language collection; S = source language collection.

The coefficient α is a corpus dependent normalizing factor. It is assigned such a value that $rel-df > 1$ indicates that the target word form is an equivalent, and $rel-df < 1$ indicates the equivalent is not found in the translation set. The coefficient values reflect the relative sizes of the subwebs or other frequency sources in relation to each other. In our case $\alpha = 2$ was used in all test conditions. The values of $\alpha =$ from 1 to 2 are appropriate for

the conditions where the target corpus is much larger than the source corpus, which was the case in our experiments.

The example in Table 3 illustrates the case where the word with the highest DF is not the correct equivalent. The translation set contains the word forms and the associated frequencies of English Web pages for a Spanish source word *fraccionamiento*.

A typical frequency pattern is found. However, *fraccionamiento*, the word with the highest DF, is the Spanish source word not translated into English. Its DF in the Spanish portion of Web is 416 000. It is not accepted as an equivalent since $df_T(\textit{fraccionamiento}) / (2 \times df_S(\textit{fraccionamiento})) < 1$. We considered two highest ranked word forms, and naturally also for the second form, *fraccionamento*, $rel-df < 1$.

Table 3. Generated word forms and their frequencies for the source word *fraccionamiento* (partial).

Generated Word Form i	$df_T(i)$	$df_S(sw)$ sw= <i>fraccionamiento</i>	rel-df
fraccionamiento	58 000	416 000	0.07
fraccionamento	95	416 000	< 0.01
fraccionament	31	-	-
fraccionamient	7	-	-
fraccionamente	3	-	-
fraccionamiento	0	-	-
fraccionamyent	0	-	-

4.3.3 Length Factor

Cross-lingual spelling variants are close to each other in word length. A great difference between the length of a target word form and the source word is an indication of a wrong equivalent. The length factor is taken into account as FITE identifies equivalents.

The length criteria for accepting an equivalent candidate as an equivalent are shown in Table 4. It is seen, for example, that when a source word contains 7 characters the target word form has to have 5-9 characters in order to be accepted as an equivalent.

Table 4. FITE's length criteria.

# characters in the source word	Accepted # characters in the target word form
5	4-7
6	5-8
7-10	length difference 0-2 characters
> 10	length difference 0-3 characters

4.3.4 The application of FITE

All the three criteria described in Sections 4.3.1- 4.3.3 – (1) the comparison of the DFs of the top word forms with one another, (2) the *rel-df* formula, and (3) length criteria - have to be met to accept a word form as an equivalent. The criteria were applied in steps as described below. A given step was applied only if the preceding step did not yield a solution. If no step yielded a solution the source word was indicated to be *untranslatable* by means of TRT. The process proceeded as follows:

1. The highest ranking target word $w_{T,1}$:
 - Step 1A direct translation
 - Step 1B second direct translation (for Finnish only)
 - Step 1C the first pivot language translation
 - Step 1D the second pivot language translation
2. The second highest ranking target word $w_{T,2}$:
 - Step 2A direct translation

- Step 2B second direct translation (for Finnish only)
- Step 2C the first pivot language translation
- Step 2D the second pivot language translation

In Steps 1A-1D the first word form was accepted as an equivalent if the next three criteria were fulfilled: (A) $df_T(1) \geq 10 df_T(2)$; (B) $df_T(1) / (2 \times df_S(sw)) > 1$; (C) the length of sw is close to the length of $w_{T,1}$, as defined in Table 4.

In Steps 2A-2D the corresponding criteria were used for the second highest ranking target word $w_{T,2}$.

We conclude this section by summarizing in Figure 1 the FITE-TRT process. The left side of the figure describes the production of transformation rules and the translation of OOV words by the TRT technique. The FITE technique is described on the grey background. FITE-TRT effectiveness was evaluated using the measures of translation recall and precision and indication precision.

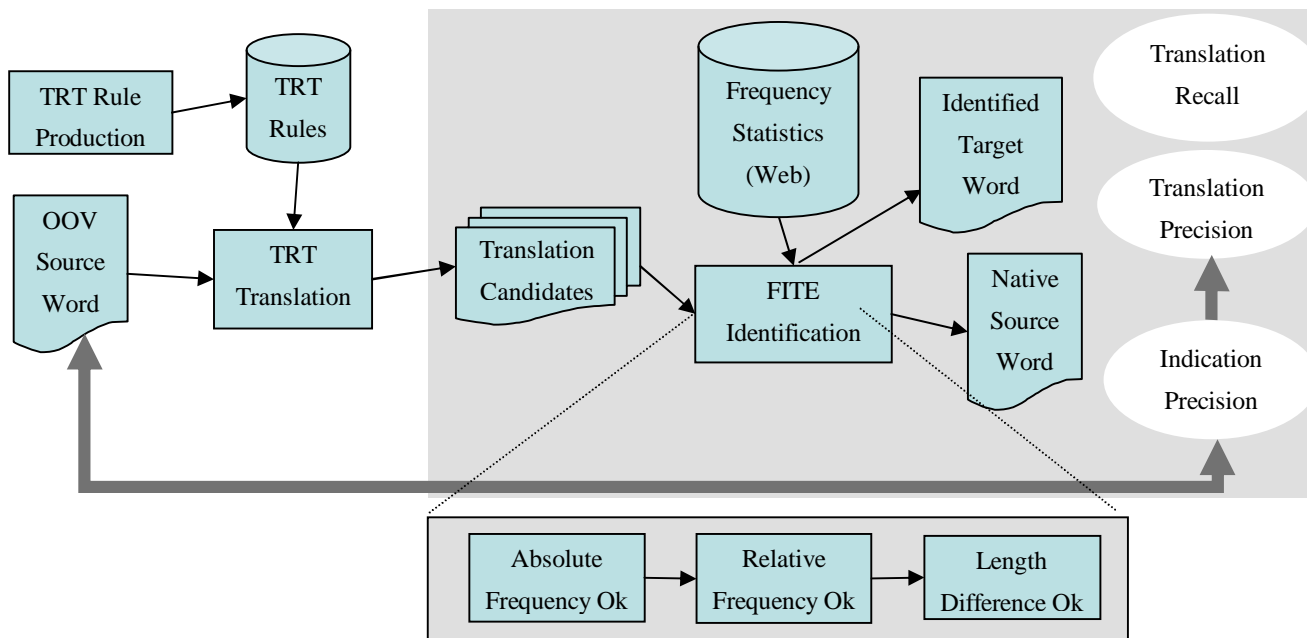


Figure 1. The FITE-TRT process.

4.4 CLIR Experiments

FITE-TRT was applied as part of an actual CLIR system. As a test data we used TREC Genomics Track 2004 data. The data consist of 50 test topics, a subset of the Medline collection containing around 4.5 million documents, and relevance judgments. Queries were formulated on a basis of the Title and Need fields of the topics. The data are well suited for investigating FITE-TRT since the topics are rich in technical (mainly biological and medical) terms. The topics were translated manually into Spanish and Finnish by one of the authors. The final Spanish topics were formulated by a knowledgeable Spanish speaker (a university teacher of Spanish). The Finnish translator is a native Finnish speaker and has expertise in medical informatics.

The Spanish and Finnish topics were translated back into English using the UTACLIR system [3] and the queries were then run on the Genomics Track test collection. UTACLIR's output without FITE-TRT provides a cross-lingual baseline for the FITE-TRT queries for which the OOV keys were translated by means of TRT and equivalents were identified using FITE. We also run the original English queries to show the performance level of the translated queries.

All inflected query words were rendered into base form for a dictionary look-up either using a morphological analyzer or manually. Manual normalization was necessary because at this stage of development TRT only translates base forms. Thus, the results show CLIR performance in a situation where a searcher gives query keys in base form.

The test system was the *InQuery* retrieval system [4]. Queries were formulated using the structured query method described in [6].

5. FINDINGS

Table 5 reports the translation recall and precision results for the bio-terms, and Table 6 the contribution of different translation routes to the recall for the bio-terms. Table 7 presents the translation recall and precision and indication precision results for the OOV-CLIR-SPA-ENG and OOV-CLIR-FIN-ENG words. The results of the CLIR experiments are presented in Table 8.

Table 5 shows that Spanish-English FITE-TRT reaches 91.0% recall and Finnish-English FITE-TRT 71.9% recall. While Spanish-English FITE-TRT achieves higher recall precision is approximately the same and remarkably high for both language pairs, i.e., 97.0%-98.8%.

The contribution of direct translation to recall is substantial for both language pairs (Table 6). For Finnish-English the contribution of the second direct route is high. Indirect translation adds recall only for Spanish-English. The first pivot language adds recall by 6.7% while the second one adds recall only by 2.2%.

For Spanish-English recall and precision are high and approximately the same for the two test word sets (tables 5 and 7). Recall is 89.2% - 91.0% and precision 97.6% - 98.8%. Also for Finnish-English precision is high (97.0% - 97.2%) but recall is lower (71.9% - 72.9%) than for Spanish-English. The higher recall for Spanish-English is the result of the better quality of the Spanish-English rule collection compared to the two Finnish-English collections.

For the native words indication precision is 100% in all test situations. There were only 10 native Spanish and Finnish words in all, however the results are reasonable since the cases where TRT gives accidentally correct words are not common.

The results of the retrieval experiments are presented in Table 8.

As expected the queries where OOV keys are translated by FITE-TRT perform substantially better than the baseline queries where OOV keys are retained untranslatable. In Spanish-English CLIR MAP improvement percentage is 40.3%. Precision at 20 documents is improved by 48.8%. For Finnish-English performance improvements are smaller. These findings are in

agreement with the high number of OOV keys in the CLIR runs and FITE-TRT's high translation recall and precision. The higher performance of English queries wrt. to the performance of Spanish-English and in particular Finnish-English queries is mostly caused by irrelevant keys yielded by UTACLIR.

Table 5. FITE-TRT effectiveness. Translation recall and translation precision for *bio-terms*.

<i>Source language</i>	<i>Translation Recall</i>	<i>Translation Recall %</i>	<i>Translation precision</i>	<i>Translation precision %</i>
Spanish	81/89	91.0	81/82	98.8
Finnish	64/89	71.9	64/66	97.0

Table 6. FITE-TRT effectiveness. The contribution of different steps to translation recall for *bio-terms*.

<i>Translation route</i>	<i>Spanish-English</i>		<i>Finnish-English</i>	
	<i>Recall</i>	<i>Recall %</i>	<i>Recall</i>	<i>Recall %</i>
First direct route (Steps 1A and 2A)	73/89	82.0	49/89	55.1
Second direct route (Steps 1B and 2B for Finnish)	-	-	15/89	16.9
First indirect route (Steps 1C and 2C)	6/89	06.7	0/89	00.0
Second indirect route (Steps 1D and 2D)	2/89	02.2	0/89	00.0
All	81/89	91.0	64/89	71.9

Table 7. FITE-TRT effectiveness. Translation recall and translation/indication precision for *OOV-CLIR-SPA-ENG* and *OOV-CLIR-FIN-ENG* keys.

<i>Source language</i> <i>Word type</i>	<i>Translation Recall</i>	<i>Translation Recall %</i>	<i>Translation/indication precision</i>	<i>Translation/indication precision %</i>
Spanish				
Spelling variants	83/93	89.2	83/85	97.6
Native words	-	-	5/5	100.0
Finnish				
Spelling variants	35/48	72.9	35/36	97.2
Native words	-	-	5/5	100.0

Table 8. CLIR performance.

Query type	MAP	% change wrt Utaclir	Pr. at 20 docs	% change wrt Utaclir
Baselines				
English queries	0.3195	-	0.5152	-
Utaclir Spa-Eng baseline	0.2018	-	0.3009	-
Utaclir Fin-Eng baseline	0.1971	-	0.3047	-
FITE-TRT queries				
Spanish - English	0.2832	+40.3	0.4477	+48.8
Finnish - English	0.2491	+26.4	0.3981	+30.7

6. DISCUSSION AND CONCLUSIONS

The aim of this research was to devise a technique that effectively identifies the correct equivalents of source words among the many word forms produced by TRT and which reliably identifies native words not translated by TRT.

The effectiveness of the devised FITE technique was tested for Spanish-English and Finnish-English spelling variants and actual OOV words in the domains of biology and medicine. It was demonstrated that the FITE technique identifies effectively the English equivalents of the Spanish and Finnish source words obtained by TRT as well as native words. For both language pairs FITE-TRT achieved very high precision. For Spanish also recall was high. We also showed that CLIR performance is improved when a CLIR system is augmented with FITE-TRT.

The TRT program we used in this study was not able to process a high number of word forms in a reasonable time frame. Therefore we had to apply CF and frequency thresholds. We observed that for many source words equivalents were not found in translation sets because the CFs and frequencies of the relevant rules were below thresholds. We therefore expect that recall values can still be improved by using a more sophisticated TRT program. It is possible that comprehensive rule collections and a sophisticated TRT program would make indirect translation unnecessary.

In this study we used Web search engine results as a frequency source, which is not a good solution when practical applications of the FITE-TRT technique are developed. The next main step in the FITE-TRT research is to collect large word frequency lists using Web mining and use the lists as a frequency source, which makes FITE-TRT independent of Web search engines.

7. ACKNOWLEDGMENTS

The Multilingual Medical Technical Dictionary (<http://www.interfold.com/translator/>) was provided by Andre Fairchild, of Denver, Colorado, USA. We would like to thank Andre Fairchild for permission to use the dictionary.

This work was financed by the Finnish Academy projects no. 1209960 (Multilingual and Task-based Information Retrieval) and no. 1206568 (NLP-based Information Retrieval Systems for the Biological Literature).

8. REFERENCES

- [1] Cheng, P.-J., Teng, J.-W., Chen, R.-C., Wang, J.-H., Lu, W.-H. and Chien, L.-F. (2004). Translating unknown queries with Web corpora for cross-language information retrieval. *Proceedings of the 27th ACM SIGIR Conference*, pp. 146-153.
- [2] Fujii, A. and Ishikawa, T. (2001). Japanese/ English cross-language information retrieval: Exploration of query translation and transliteration. *Computers and the Humanities*, 35(4), 389-420.
- [3] Hedlund T., Airio E., Keskustalo H., Lehtokangas R., Pirkola A. and Järvelin K. (2004). Dictionary-based cross-language information retrieval: Learning experiences from CLEF 2000-2002. *Information Retrieval*, 7, 99-119.
- [4] Larkey, L.S. and Connell, M.E. (2005). Structured queries, language modeling, and relevance modeling in cross-language information retrieval. *Information Processing & Management*, 41(3), 457-473.
- [5] Pirkola, A., Toivonen, J., Keskustalo, H., Visala, K. & Järvelin, K. (2003). Fuzzy translation of cross-lingual spelling variants. *Proceedings of the 26th ACM SIGIR Conference*, pp. 345 - 352.
- [6] Sperer, R. and Oard, D. (2000). Structured translation for cross-language IR. *Proceedings of the 23rd ACM SIGIR Conference*, pp. 120-127.
- [7] Toivonen, J., Pirkola, A., Keskustalo, H., Visala, K. & Järvelin, K. (2005). Translating cross-lingual spelling variants using transformation rules. *Information Processing & Management*, 41(4), 859-872.
- [8] Zhang, Y. and Vines P. (2004). Using the Web for automated translation extraction in cross-language information Retrieval. *Proceedings of the 27th ACM SIGIR Conference*, pp. 162-169.