Authors:	Pirkola Ari, Toivonen Jarmo, Keskustalo Heikki, Järvelin Kalervo
Name of article:	Frequency-based identification of correct translation equivalents (FITE) obtained through transformation rules
Year of publication:	2007
Name of journal:	ACM Transactions on Information Systems
Volume:	26
Number of issue:	1
ISSN:	1046-8188
Discipline:	Natural sciences / Computer and information sciences
Language:	en
School/Other Unit:	School of Information Sciences

Frequency-based Identification of Correct Translation Equivalents (FITE) Obtained through Transformation Rules

ARI PIRKOLA, JARMO TOIVONEN, HEIKKI KESKUSTALO, AND KALERVO JÄRVELIN
University of Tampere

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 technique called *frequency-based identification of translation equivalents* (*FITE*) was tested using biological and medical cross-lingual spelling variants and out-of-vocabulary (OOV) words in Spanish-English and Finnish-English TRT. The results showed that - depending on the source language and frequency corpus - FITE-TRT (i.e., the identification of translation equivalents from TRT's translation set by means of the FITE technique) may achieve high translation recall. In the case of the Web as the frequency corpus, translation recall was 89.2%-91.0% for Spanish-English FITE-TRT. For both language pairs FITE-TRT achieved high translation precision, i.e., 95.0%-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 basic dictionary-based CLIR where OOV keys were kept intact. FITE-TRT with Web document frequencies was the best technique among several fuzzy translation / matching approaches tested in cross-language retrieval experiments. We also discuss the application of FITE-TRT in the automatic construction of multilingual dictionaries.

---

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 Pirkola et al. [2003] we generated automatically large collections of character correspondences in several language pairs for the translation of cross-lingual spelling variants. Equivalent term pairs in two languages were first extracted automatically from translation dictionaries, and then regular character correspondences between the words in the two languages were identified using an edit distance measure. Large sets of transformation rules augmented with statistical information were generated for automatic translation of spelling variants. We call the translation technique based on the generated rules *transformation rule based translation* (TRT). TRT is similar to transliteration except that no phonetic elements are involved in it. The term *fuzzy translation* is used in connection with TRT. It refers to the fact that TRT often gives for a source word many possible equivalents, not one equivalent or several alternatives like regular translation.

In Toivonen et al. [2005] we showed that high translation recall (i.e., the proportion of source words for which TRT yields equivalents among all source words) may be achieved when most of the rules available for a source word are used in TRT. However, high translation recall is associated with low translation precision (i.e., the proportion of equivalents among all word forms yielded by TRT). In other words, the translation set containing the target language word forms often includes the correct translation equivalent of a source word and a large number of other word forms.

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, in many cases the TRT technique may be useless if it just indicates a set of possible translations for a source word but is not able to indicate the one correct equivalent, which was the case in Pirkola et al. [2003] as well as in Toivonen et al. [2005]. In the present research we combat this problem, and move TRT from fuzzy translation towards dictionary-like translation where for each source word either one translation 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 (pivot) languages. As in the case of direct translation the equivalents are identified from TRT's translation set by means of the novel FITE technique. Transitive translation through a pivot language is a well-known technique in CLIR used to address the problem of limited availability of translation resources [Ballesteros 2000; Gollins and Sanderson 2001; Lehtokangas et al. 2004]. Also in TRT indirect translation could be used in cases where direct translation is not possible due to the lack of translation resources (transformation rules). In this study, however, we investigate whether indirect translation improves FITE-TRT effectiveness. It may compensate the failures of direct translation and thereby increase translation recall.

We explore the effectiveness of FITE in 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.

FITE-TRT is also applied as part of an actual CLIR system. The effectiveness of dictionary-based CLIR augmented with FITE-TRT is compared to the effectiveness of dictionary-based CLIR augmented with plain TRT and skipgram [Keskustalo et al. 2003] OOV word methods. We also run dictionary-translation-only (i.e., no OOV word technique is applied) and monolingual English queries as baselines.

In Pirkola et al. [2006] we presented the main features of FITE-TRT and the first results on FITE-TRT effectiveness and the effectiveness of CLIR augmented with FITE-TRT. In this paper we describe the FITE-TRT technique in more detail, present the *find-equivalent* algorithm, and extend the first study by using large word frequency lists mined from the Web as FITE-TRT's frequency source and by comparing in cross-language retrieval experiments FITE-TRT to other OOV word methods.

The novel FITE-TRT technique is fundamentally different from other OOV word methods / systems presented in the literature. For instance, Cheng et al. [2004] and Zhang and Vines [2004] 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. Meng et al. [2000] employed TRT type rule-based approach to the OOV word problem. Phonetic mappings were derived from English and Chinese (Mandarin) pronunciation rules for English-Chinese spoken document retrieval. The

researchers also considered Chinese name variation. An English proper name may have several character sequence variants and pronunciations in Chinese. To combat this problem the transliteration approach should involve approximate matches between the English and Chinese pronunciations. Fujii and Ishikawa [2001] 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 fuzzy matching. The technique is different from FITE-TRT but bears some resemblance to fuzzy translation reported in Pirkola et al. [2003], however focusing on languages with different orthographies and having thus different focus. The skipgram fuzzy matching approach to OOV words by Keskustalo et al. [2003] is discussed in Section 5.2.1.

The rest of this paper is organized as follows. Section 2 presents the TRT technique, its background research, the transformation rule collections, and the dictionary data that was used in the rule generation. In Section 3 we define the terms *cross-lingual spelling variant* and *native word*, and present the research problems and evaluation measures used in the experiments. The novel FITE technique is described in Section 4. Section 5 presents the methods and data used in the experiments and the findings. Section 6 contains the discussion and conclusions.

2. TRT, TRANSFORMATION RULES AND BACKGROUND RESEARCH

The idea of TRT and the automatic method to generate transformation rules is described in Pirkola et al. [2003]. 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   middle191   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 middle of words, with the confidence factor being 89.25% (100% * 191/214). Examples of target word forms obtained in TRT are shown in Sections 4.2-4.3.

In Pirkola et al. [2003] we studied TRT in combination with fuzzy matching, i.e., digram and trigram matching. We investigated five source languages, with English being a target language for all the source languages. The results showed that for Finnish, German and Spanish the combined technique performed better than digrams and trigrams alone. For French and Swedish performance changes were slight.

In Toivonen et al. [2005] we studied how effective TRT is as such without fuzzy matching. We found that translation recall was high when low frequency and confidence factor were used as thresholds to select the rules for TRT. However, at low confidence factor and frequency levels translation precision was low. The FITE-TRT technique addresses this problem and, as we will show in this paper, it achieves both high recall and precision.

The transformation rules used in TRT in this study were generated using the rule generation method based on the use of dictionary data described in Pirkola et al. [2003]. The dictionary data consisted of a multilingual medical dictionary by Andre Fairchild (http://members.interfold.com/translator/) for the language pairs of Spanish-English, Spanish-German, Spanish-French, German-English, and French-English. For Finnish-English the data for rule generation was obtained by translating (1) a list of Finnish medical terms into English using a medical dictionary by Kielikone Inc. and (2) a list of Finnish terms in various domains into English using Kielikone's general-purpose dictionary. Thus, for Finnish-English we constructed two collections. The second collection was constructed because the first collection missed many important rules.

Table 1 shows the number of entries and the average number of translations per an entry for each dictionary used in the rule generation (columns 2 and 3). Table 1 also shows the total number of rules in the generated rule collections as well the number of rules at or above the thresholds of CF=4.0% and frequency=2 applied in this study (columns 4 and 5). We applied the confidence factor and frequency thresholds because TRT may give very large translation sets, and at the present stage of development the TRT program is not efficient enough to process very large word sets. Due to the efficiency issues we also applied a limit of 40 word forms: if there were more than 40 word forms in a translation set of an intermediate language the source word was retranslated by TRT with the confidence factor of 10.0% and frequency of 10 to yield a smaller translation set.

Table 1. Dictionary and transformation rule collection statistics

Dictionary/Rule collection	# Entries	Avg. # translations	# Rules	# Rules CF≥4.0%, Freq. ≥2
Spanish-English	19029	1.58	8800	1295
Spanish-German	14252	1.70	5412	984
Spanish-French	15183	1.66	9724	1430
German-English	18917	1.70	8609	1219
French-English	17089	1.91	9873	1170

As can be seen in Table 1 each rule collection contains a high number of rules which suggests that the rule generation method captured effectively spelling variation between the language pairs.

3. RESEARCH PROBLEMS AND EVALUATION MEASURES

We distinguish between two kinds of words in a language with respect to the words in another language: cross-lingual spelling variants and native words. *Cross-lingual spelling variants* are etymologically related words and therefore similar in the two languages, differing only slightly in spelling. A *native word* and its target language equivalent are not related to each other morphologically even if they share the same meaning. The words have different origins and etymologies in the history of respective languages. The words do not have morphological or phonetic resemblance - or if there is some, it is purely accidental.

As examples, consider the English words *computer* and *chemotherapy* and their Finnish equivalents *tietokone* and *kemoterapia*. The first pair *computer-tietokone* do not have morphological or phonetic resemblance, *computer* originating from Latin (*computare*, to calculate) and *tietokone* being a compound of *knowledge~* and *~machine*. The Finnish words *tieto* and *kone* are old words in the language. In the second pair *chemotherapy-kemoterapia* both words originate from Greek (*chemeia + therapeia*) and, albeit having been modified to fit the style of their present languages, still have not lost their morphological or phonetic resemblance.

TRT is intended to translate spelling variants and FITE is intended to identify the translation equivalents of spelling variants and to indicate the native source language words – the source words that cannot be correctly translated by TRT. Using test word sets containing both types of words we examine the following research questions:

- In the case of spelling variants, 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?

- Are word frequency lists mined from the Web competitive with the Web as a collection of documents as FITE-TRT's frequency source?

- 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 a CLIR system augmented with TRT and fuzzy matching OOV word methods, and in comparison to dictionary-translation-only 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. For example, if there are 10 source words and TRT gives for these 100 target language word forms among which there are correct equivalents for 8 source words then translation recall is 8/10=80%. *Translation precision* is defined as the proportion of correct equivalents among all words which are indicated as equivalents. For example, if FITE identifies 10 translation equivalents of which 9 are correct equivalents translation precision is 9/10=90%. For native words the question what share of them is 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 by TRT. For example, if there are 5 native source words and FITE indicates that for none of these translation equivalents are contained in the translation sets indication precision is 5/5=100%

Retrieval effectiveness was evaluated using the measures of *mean average precision* (MAP) and *precision at 20 documents*. MAP is a standard evaluation measure used in TREC (http://trec.nist.gov), and it refers to the average of the precision values obtained after each relevant document is retrieved. MAP is a system-oriented measure while precision at 20 documents is important from the practical IR standpoint. The probability of a searcher scanning further down a ranked result list decreases as (s)he scans down and we use a document cut-off value of 20 as a rule of thumb for the stopping point of scan.

4. THE FITE TECHNIQUE

4.1 Frequency Data

FITE identifies the correct translation equivalents among the TRT generated word forms by their frequency distribution in some corpus. Frequencies for FITE were taken from the *Web* and *word frequency lists*. In the case of the Web we consider document frequency (DF) and in the case of frequency lists word frequency (WF). DF statistics were collected using the Altavista search engine and its language selection feature. A research assistant fed the word forms into the search engine which reported for each word form the number of documents containing the word form.

The word frequency lists were mined from the Web using a Web mining technique which is described next. In the first step of the Web mining process a query script based on the use of a text-based Web browser *Lynx* was run to fetch medical and biological documents in a desired language from the Google search engine. The query script described in [Zhang and Vines 2004] was modified for this purpose. We used the following parameters and parameter values in the script: language [=English / Finnish / German / Spanish]; the number of documents to fetch [=700]; query keys [the words *medicine*, *biology*, and *disease* (conjuncted by the AND-operator) and the corresponding words in Finnish, German, and Spanish]. The use of these keys directed the actual Web mining towards medical and biological sub-webs. In the second step, URLs were extracted from the fetched documents and were saved in a file. In the third step, the URL file was cleaned by removing duplicates so that only URLs with unique domain names were kept in the file. The URL file served as an input for the fourth step, the actual Web mining stage where documents were downloaded from each Web site represented in the URL file using a *wget* program (http://www.gnu.org/software/wget/). Wget's parameter "directory depth" was set at 3, i.e., on each Web site documents at directory depths 1-3 were downloaded. In the fifth step of the process all downloaded documents were combined into one large file. In the sixth step, word frequency lists were constructed from the combined document file.

The number of documents downloaded from the Web varied depending on the language. For example, for German 35 000 documents were downloaded. The total size of these documents was 2.26 GB.

The numbers of unique words contained in the frequency lists are as follows:

- English: 762 000 words
- Finnish: 886 000 words
- German: 470 000 words
- Spanish: 386 000 words

We did the Web experiments prior to the word frequency experiments. Since the results of the Web experiments indicated that the contribution of the second intermediate language (French) was small it was not considered in the frequency list experiments, and we did not construct a word frequency list for French. Its minor contribution was probably due to the fact that it was used as the second intermediate language, rather than its linguistic features.

In statistical MT the choice between translation alternatives depends on the translation probabilities of the alternatives and their context [Al-Onaizan et al. 1999; Brown et al. 1990]. Translation probabilities are computed on a basis of aligned corpora. In contrast to this, the source and target language corpora used by FITE are independent of each other. In statistical MT bi-gram and tri-gram language models are typically used to capture the context. FITE is based on a unigram language model, i.e., no context dependence of translations is assumed.

4.2 Frequency Pattern

In order to avoid several long function definitions not precisely in the focus of our paper, we introduce below some notational conventions used in the definition of the FITE method.

Notational Convention 1. Let *SL* be a source language and *TL* a target language, and *sw* be some source language word in the source language collection *S*. We denote this by $sw \in SL$. We denote the word set produced by our TRT translation by $TRT_{SL\text{->}TL}(sw)$ using the transformation rules for *SL->TL* translation. The result is a set, i.e., its elements *tw* hold the relationship $tw \in TRT_{SL\text{->}TL}(sw)$. If the TRT translation is performed using a *strict* confidence factor (=10%) and a *strict* rule frequency (=10), see Section 2, we denote this by $TRT_{SL\text{->}TL|strict}(sw)$.

Notational Convention 2. Let *sw* be some word in source language *SL*, i.e. $sw \in SL$. We denote its document frequency in the source language collection *S* by $df_S(sw)$. Note that if *sw* does not appear in any documents of *S* then $df_S(sw) = 0$. If *S* is a source language wordlist containing word frequencies, we denote the frequency of *sw* in *S* by $wf_S(sw)$.

Notational Convention 3. Let *tw* be some word of the target language *TL* in the target language document collection *T, i.e. $tw \in TL$*. We denote its document frequency in *T* by $df_T(tw)$. It refers to the frequency of target language documents that contain the word *tw*. Note that if *tw* does not appear in any documents of *T* then $df_T(tw) = 0$. If *T* is a target

language wordlist containing word frequencies, we denote the frequency of *tw* in *T* by $wf_T(tw)$.

Notational Convention 4. Let *sw* be some source language *SL* word, i.e. $sw \in SL$, *TL* a target language, $TWS = TRT_{SL->TL}(sw)$ the word set produced by our TRT translation, and *T* a target language document collection or word list. The frequency-ranked list of words of *TWS* in *T* is denoted by R= *trt-frank(TWS, T)*. Table 2 is an example of such a list with the frequency data added. For a given source language word *sw* we obtain this list by *trt-frank($TRT_{SL->TL}(sw)$, T)*. The elements of this list are denoted by the usual notation, e.g., *trt-frank(TWS, T)*[3] gives its third component.

As an example, in the case of Table 2, *trt-frank($TRT_{SPA->ENG}$(biosintesis), EngWeb)*[3] = biosyntesis. Its frequency is df_{EngWeb}(*trt-frank($TRT_{SPA->ENG}$(biosintesis), EngWeb)*[3]) = df_{EngWeb}(biosyntesis) = 634.

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 in the English sub-web. The word forms are sorted by document frequency, i.e., by *trt-frank($TRT_{SPA->ENG}$(biosintesis), EngWeb)*. We can see 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. Given a target word form ranking R = *trt-frank(TWS, T)*, the magnitude of difference between the document frequency of the first word form (df_T(R[1])) and the document frequency of the second word form (df_T(R[2])), or the frequencies between R[2] and R[3] (see Section 4.5) forms the basis of the equivalent identification. We used the coefficient value (the magnitude of difference) of 10 for the identification of equivalents (both for Web and word frequency lists).

Table 2. An example of generated word forms for R = *trt-frank($TRT_{SPA->ENG}$(biosintesis), EngWeb)* and their document frequencies $df_{EngWeb}(tw)$ in the English sub-web (partial).

Generated Word Form *tw* ∈ *R*	$df_{EngWeb}(tw)$
biosíntesis	2 230 000
biosintesis	909

biosyntesis	634
biosinthesis	255
biosynthessis	3
biosintessis	0
biosinthessis	0
biosyntessis	0

The same pattern holds for the Web and word frequency lists, with the main differences being in that in the case of frequency lists *word* frequencies instead of document frequencies are considered and in that Web gives more malformed words than the frequency lists. The following definition of the function *freq-pattern-ok* checks whether the frequencies of two target language words tw_i and tw_j have the required pattern.

Definition-1. Let tw_i and tw_j be two candidate word forms in the target language (sub-) web document collection, or word frequency list, T as given by TRT. Let $df_T(tw_i)$ and $df_T(tw_j)$ be their frequencies in T. Let β be a corpus dependent normalizing factor, $\beta > 1$. The Boolean function *freq-pattern-ok* gives the value *true* if the frequency of tw_i in T is at least β times the frequency of tw_j in T.

$$freq\text{-}pattern\text{-}ok(tw_i, tw_j, \beta, T) = \quad \textit{true, if } df_T(tw_i) \geq (\beta \times df_T(tw_j))$$
$$\textit{false, otherwise.}$$

Typically, the function *freq-pattern-ok* is applied on two consecutive words in a frequency-ranked order. For example, *freq-pattern-ok*(biosynthesis, biosintesis, 10, *EngWeb*) yields the value *true* – cf. Table 2. In the tests, the coefficient β was set experimentally at $\beta = 10$.

4.3 Relative Frequency

There are situations where the highest DF (WF) 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. (A source word is always included in TRT's translation set because source and target language words may be identical.) Also, in the case of Web as a document collection there are mixed language pages some of which a search engine may consider target language pages which wrongly increases the target DF of a source word found in the mixed language pages. TRT may also accidentally give high DF words which are not correct equivalents. As a solution for this problem, we compute *relative document frequency (rel-df)* and *relative word frequency (rel-wf),* defined as follows.

Definition-2. Let *sw* be a source language word in the source language collection *S,* and *tw* a target language word form in the target language (sub-)web document collection *T* as given by TRT. Let $df_S(sw)$ be the frequency of *sw* in *S* and $df_T(tw)$ the frequency of the target language word *tw* in *T*. Let α be a corpus dependent normalizing factor, $\alpha > 0$. The function *rel-df* gives the relative document frequency for *tw* in *T*.

Definition-3. Let *sw* be a source language word in the source language word list *S,* and *tw* a target language word form in the target language word list *T* as given by TRT. Let $wf_S(sw)$ be the frequency of *sw* in *S* and $wf_T(tw)$ the frequency of the target language word *tw* in *T*. Let α be a corpus dependent normalizing factor, $\alpha > 0$. The function *rel-wf* gives the relative word frequency for *tw* in *T*.

$$rel\text{-}wf(tw,\ sw,\ \alpha,\ T,\ S) = wf_T(tw)\ /\ (\alpha \times wf_S(sw))$$

The coefficient α is a corpus dependent normalizing factor. It is assigned such a value that *rel-df* and *rel-wf* > 1 indicate that the target word form is an equivalent, and *rel-df* and *rel-wf* < 1 indicate the equivalent is not found in the translation set. The coefficient values reflect the relative sizes of the subwebs / word frequency lists in relation to each other. In our case $\alpha = 2$ was used in all test conditions. The value $\alpha = 2$ was determined experimentally. The values of α 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 Finnish word frequency list contains more words than the English list (Section 4.1). However, the sum frequency over all words is substantially higher in the English than Finnish list. This allows the use of $\alpha = 2$ in the *rel-wf* formula also for Finnish-English.

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

Generated Word Form $tw \in R$	$df_{EngWeb}(tw)$	df_{SpaWeb}(fraccionamiento)	*rel-df*
fraccionamiento	58 000	416 000	0.07
fraccionamento	95	416 000	< 0.01
fraccionament	31	-	-
fraccionamient	7	-	-
fraccionamente	3	-	-
fraccionamyento	0	-	-
fraccionamyent	0	-	-

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 *rel-df(fraccionamiento, fraccionamiento, 2, EngWeb, SpaWeb)* < 1. We considered two highest ranked word forms, and naturally also for the second form, *fraccionamento, rel-df* < 1.

4.4 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 can be seen, for example, that when a source word contains 7 characters the target word form has to have from 5 to 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

Definition-4. Let *sw* be a source language word and *len*(*sw*) its length in characters. Likewise, let *tw* be a target language word and *len*(*tw*) its length in characters. The Boolean function *tw-len-ok* gives the value *true* if the length of the target word *tw* is within the range defined in Table 4.

$$
\begin{aligned}
\textit{tw-len-ok}(tw,\ sw) = \ &\textit{true, if } 4 \le len(tw) \le 7 \text{ and } len(sw) = 5 \\
&\textit{true, if } 5 \le len(tw) \le 8 \text{ and } len(sw) = 6 \\
&\textit{true, if } |\ len(tw) - len(sw)\ | \le 2 \text{ and } 7 \le len(sw) \le 10 \\
&\textit{true, if } |\ len(tw) - len(sw)\ | \le 3 \text{ and } len(sw) > 10 \\
&\textit{false, otherwise}
\end{aligned}
$$

4.5 The Application of FITE

In the empirical experiments the source test words (Section 5.1.1) were translated into English by the TRT translation program. The applied thresholds were described in

Section 2. The equivalents were searched for from the translation sets using the FITE technique. As described in Sections 4.2-4.4 the main criteria of equivalent identification of FITE are the following: (1) the frequency patterns of the top word forms tested by the function *freq-pattern-ok*, (2) the relative frequency criterion tested by the *rel-df* / *rel-wf* functions, and (3) length criterion tested by the function *tw-len-ok*.

The basic idea is to apply the criteria 1-3 in three steps A – C: First in Step A direct translation is tried and then in Steps B and C the pivot language translations one after the other. The criteria are first applied to the highest-ranking target word candidate as given by the function *trt-frank*. If these steps do not yield a solution, then basically the same steps are repeated to the second highest-ranking target word candidate. This process is specified as Algorithm *find-equivalent* which is presented in the Appendix. The algorithm is for the case of word-frequency lists S and T for the source and target languages. In the case of web document collections, the word-frequency lists are replaced by a function that gives the web document frequency for a given source word. The TRT rule bases for the source, pivot and target languages, as described above, need to be available but are not precisely defined (see notational conventions in Section 4.2). We use the notations and functions of preceding sections in the definition of the algorithm.

The algorithm *find-equivalent* first tries the first candidates produced by the TRT direct or pivoted processes. It calls the procedure *direct-trans* to produce the frequency-based ranking of the direct TRT candidates. The procedure generates them as the list R and then the first component of R is tested for the criteria 1-3 by the procedure *test-cand*. If the first component stands the test it is given as the equivalent. If not, the algorithm *find-equivalent* then uses the first and finally the second pivot language translation, given by the procedure *pivot-trans*, which first checks the number of pivot language word forms obtained. The TRT rules are used liberally (the thresholds of CF=4.0% and frequency=2 are used; see Section 2), if there are at most 40 candidates and otherwise strictly (the thresholds of CF=10.0% and frequency=10 are used). Either way produces a target language word candidate list *TWS,* which then is ranked by frequency and the first component tested for the criteria 1-3.

If the first pass, focusing on the first-ranked components, is not successful, then the algorithm *find-equivalent* tries the second equivalent candidate produced by the TRT processes. The <u>*first component*</u> is still selected as the equivalent if the three criteria are fulfilled as follows: the second component passes the frequency pattern and relative frequency criteria and the first one the length criterion. The second word form is selected as the equivalent only if the first word form does not meet the length criterion and the

second form meets all the three criteria. Otherwise the source word is indicated to be *untranslatable* by means of TRT – the string "nil" is returned. We found empirically the need to compare the second candidate word form to the third form to find out if there are more than one correct target language words (i.e., high frequency word forms) in the translation set. If there are exactly two acceptable words the first word rather than the second one is selected as the equivalent based on our observations that also in these cases the equivalent tends to be at the first position. The second word form is accepted as the equivalent only if the first form does not meet the length criterion as explained above.

In the actual experiments described in Section 5, the algorithm *find-equivalent* was applied / modified as follows. In the case of frequency lists the second pivot language (French) was not considered. Finnish-English experiments differed from the Spanish-English experiments in that there were two direct translation routes thanks to two Finnish-English rule collections. The order of the use of the translation routes for Finnish-English was as follows: Finnish-English / collection 1, Finnish-English / collection 2, Finnish-German-English, and Finnish-French-English.

All the source words were in base form, and only base form equivalents were accepted as correct equivalents. Thus, equivalents in plural form and the derivatives of the actual equivalents were not accepted as correct equivalents. This is because our aim is to develop a dictionary-like rule-based translation method, which indicates the precise equivalents of source words.

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 means of TRT. The FITE technique - the focus of the present paper - is the grey shaded area. FITE-TRT effectiveness was evaluated using the measures of translation recall and precision and indication precision.
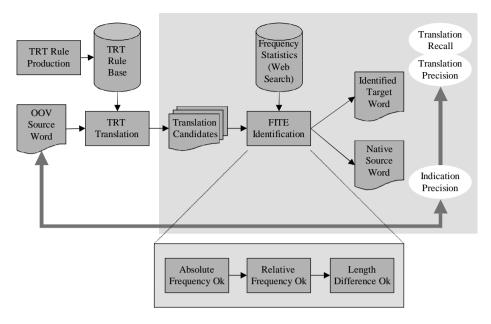
Figure 1. The FITE-TRT process.

5. EXPERIMENTS AND FINDINGS

In this section we present the methods and data used in the FITE-TRT and CLIR effectiveness experiments and the experimental results. Subsection 5.1 presents the training and test words, describes how the words were translated by means of TRT, and presents the findings of the FITE-TRT effectiveness experiments. The CLIR experiments are dealt with in Subsection 5.2.

5.1 FITE-TRT Effectiveness

5.1.1 Training and Test Word Sets and Translation by TRT

FITE-TRT is intended to handle both spelling variants and native source language words. The training and test words sets contained both types of words. Next we describe the selection of training and test words. Then we characterize quantitatively the difference between cross-lingual spelling variants and native words.

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.

For the first two sets a list of English biological and medical terms was gathered manually from the index of CLEF's [Peters 2005] LA Times collection. The English terms were translated into Spanish and Finnish by means of translation dictionaries and monolingual (Spanish and Finnish) medical dictionaries. From these words we selected for our tests Spanish-English and Finnish-English spelling variants. The identification of spelling variants was done based on the similarity of the Spanish-English and Finnish-English word pairs judged by a researcher. The similarity feature used as a selection criterion is discussed in Section 3 and later in this section. The Spanish terms formed the first and the Finnish terms the second test word set. Both contained the same terms (n=89) albeit in different languages. These terms are called *bio-terms*.

For the third and fourth test word sets, TREC Genomics Track 2004 topics in Spanish and Finnish (Section 5.2.1) were translated into English using the UTACLIR system, an automatic dictionary-based query translation and construction system developed in the Information Retrieval Laboratory at the University of Tampere [Hedlund et al. 2004]. The OOV keys of the UTACLIR runs were used as test words. Among the OOV keys there were, in addition to spelling variants, native Spanish / Finnish words as well as English words and English acronyms. The Spanish word set contained 98 and the Finnish set 53 OOV keys (after the removal of short words, see below). The difference in the number of the OOV keys reflects the different sizes of UTACLIR's Spanish-English and Finnish-English dictionaries. These test key sets are here called *OOV-UTACLIR-SPA-ENG* and *OOV-UTACLIR-FIN-ENG*.

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 total number of unique source words translated by TRT was 89+98 (Spanish) + 89 + 53 (Finnish) = 329 words.

To characterize quantitatively the difference between cross-lingual spelling variants and native words we computed for both types of source language test words their degree of similarity with respect to their English equivalents using a simple measure of longest

common subsequence divided by the mean length of source and target words (LCS/MWL). LCS is defined as the length of the longest subsequence of characters shared by two words. The closer to 1.0 LCS/ MWL is, the more similar the words are. As an example, for the Spanish bio-term *omnivoro* LCS/MWL = 7/((8+8)/2) = 0.875 w.r.t. the English equivalent *omnivore*. For the native Spanish OOV word *vinculante* LCS/MWL = 0.353 w.r.t. the equivalent *binding*.

Table 5 shows the results of LCS/MWL calculations. From the viewpoint of TRT, the target language (English) words as source words are similar cases as spelling variants and they were thus regarded as spelling variants (in the sets *OOV-UTACLIR-SPA-ENG* and *OOV-UTACLIR-FIN-ENG*). (As mentioned in Section 4. 3, a source word is included in TRT's translation set because source and target language words may be identical.) In cases where native Spanish and Finnish words had multiple meanings in English the meaning that appeared in the Genomics Track topic was selected for LCS/MWL calculation. Both the Spanish-English and Finnish-English OOV word sets contained five native words.

Table 5. LCS/MWL for the test words.

Word type	# words	Average LCS/MWL	Standard Deviation
Spa-Eng spelling variants			
Bio terms	89	0.839	0.114
OOV words	93	0.911	0.110
Spanish native words	5	0.339	0.105
Fin-Eng spelling variants			
Bio terms	89	0.784	0.114
OOV words	48	0.853	0.117
Finnish native words	5	0.361	0.085

It can be seen in Table 5 that for bio-terms and spelling variant OOV words the average LCS/MWLs are 0.839 and 0.911 (Spanish-English) and 0.784 and 0.853 (Finnish-English). For native Spanish and Finnish words average LCS/MWLs are much lower: 0.339 for Spanish and 0.361 for Finnish. Low standard deviation figures show that the LCS/MWL values are clustered around the average values.

In the experiments the source words were translated by the TRT program through direct and indirect translation routes into English using the confidence factor and rule frequency

thresholds (Section 2). The equivalents of source words were identified from TRT's English translation sets by means of the FITE technique as described in Section 4.5. Like target language translation sets the intermediate translation sets are often large, and only five top German and French forms in a frequency-ranked translation set were further translated into English. Different English translation sets corresponding to the same Finnish / Spanish source word were combined.

5.1.2 Findings

Table 6 reports the translation recall and precision results for bio-terms, and Table 7 the contribution of different translation routes to the recall for bio-terms. Table 8 presents the translation recall and precision and indication precision results for OOV-UTACLIR-SPA-ENG and OOV-UTACLIR-FIN-ENG words.

Table 6 shows that Spanish-English FITE-TRT reaches 91.0% recall in the case of Web document frequencies and 82.0% recall in the case of word frequency lists. Finnish-English FITE-TRT reaches 71.9% (Web) and 67.4% (frequency lists) recall. While Spanish-English FITE-TRT achieves higher recall, precision is approximately the same and it is remarkably high for both language pairs, i.e., 97.0%-98.8%. The same trends hold for the OOV words (Table 8): for Spanish-English recall is higher than for Finnish-English, and for both language pairs precision is very high (95.0%-97.6%).

Table 7 shows that the contribution of direct translation to the recall is substantial for both language pairs. For Finnish-English FITE-TRT the contribution of the second direct route (collection 2) to the recall is high. Indirect translation adds recall only for Spanish-English. In the case of Web as a frequency corpus the first pivot language adds recall by 6.7% while the second one adds recall only by 2.2%.

For the native words indication precision is 100% in all test situations (Table 8). There were only 10 native Spanish and Finnish words in all, however the results are reasonable since the cases where TRT accidentally gives correct words are not common.

Table 6. FITE-TRT effectiveness. Translation recall and translation precision for *bio-terms*.

Source language/ Frequency corpus	Translation Recall	Translation Recall %	Translation Precision	Translation Precision %
Spanish				
Bio terms/Web	81/89	91.0	81/82	98.8
Bio terms/frequency lists	73/89	82.0	81/82	98.8

Finnish				
Bio terms/Web	64/89	71.9	64/66	97.0
Bio terms/frequency lists	60/89	67.4	73/75	97.3

Table 7. FITE-TRT effectiveness. The contribution of different steps to translation recall for *bio-terms*.

Frequency corpus / Translation route	Spanish-English		Finnish-English	
	Recall	Recall %	Recall	Recall %
Web				
First direct route	73/89	82.0	49/89	55.1
Second direct route	-	-	15/89	16.9
First indirect route	6/89	06.7	0/89	00.0
Second indirect route	2/89	02.2	0/89	00.0
All	81/89 (91.0%)	90.9	64/89 (71.9%)	72.0
Frequency lists				
First direct route	70/89	78.7	45/89	50.6
Second direct route	-	-	15/89	16.9
First indirect route	3/89	3.4	0/89	0.0
All	73/89 (82.0%)	82.1	60/89 (67.4%)	67.5

Table 8. FITE-TRT effectiveness. Translation recall and translation/indication precision for *OOV-UTACLIR-SPA-ENG* and *OOV- UTACLIR-FIN-ENG* keys.

Source language/ Frequency corpus/ Word type	Translation Recall	Translation Recall %	Translation /Indication Precision	Translation /Indication Precision %
Spanish (OOV-UTACLIR-SPA-ENG)				
Web				
Spelling variants	83/93	89.2	83/85	97.6
Native words	-	-	5/5	100.0
Frequency lists				
Spelling variants	81/93	87.1	81/83	97.6
Native words	-	-	5/5	100.0
Finnish (OOV-UTACLIR-FIN-ENG)				
Web				
Spelling variants	35/48	72.9	35/36	97.2
Native words	-	-	5/5	100.0
Frequency lists				
Spelling variants	38/48	79.2	38/40	95.0
Native words	-	-	5/5	100.0

5.2 CLIR Effectiveness

5.2.1 Methods and Data

FITE-TRT was applied as part of an actual CLIR system. As test data we used TREC Genomics Track 2004 data [Hersh et al. 2005]. 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 a researcher. The final Spanish topics were formulated by a knowledgeable Spanish speaker (a university teacher of Spanish). The researcher is a native Finnish speaker and has expertise in medical informatics.

The test system was the *InQuery* retrieval system [Allan et al. 2000; Larkey and Connell 2005]. InQuery is a probabilistic retrieval system based on the Bayesian inference network model. Queries can be presented as a bag of word queries, or they can be structured using a variety of query operators. In this study the translated queries were structured using InQuery's #syn-operator as described in Pirkola [1998] and Sperer and Oard [2000]. The keys within the #syn-operator are all treated as instances of one key in weight computation.

The Spanish and Finnish topics were translated back into English using the UTACLIR system and the queries were then run on the Genomics Track test collection. UTACLIR's output without any OOV word technique provides cross-lingual baseline for the FITE-TRT queries for which UTACLIR's OOV words were translated by means of TRT and equivalents were identified using FITE. The original English queries were also run to show the performance level of the translated queries. We also compared the effectiveness of FITE-TRT queries to the effectiveness of *plain TRT* and *skipgram* queries. Plain TRT is the TRT part of FITE-TRT, and for plain TRT queries UTACLIR's OOV words were translated by TRT with CF=4% and frequency=2. All translations of a source word were included in a query and were wrapped in the #syn-operator. In skipgram queries the OOV words were translated using a skipgram fuzzy matching technique [Keskustalo et al. 2003]. This string matching technique is a generalization of the n-gram technique where words are split into digrams on the basis of both consecutive as well as non-consecutive characters (see below). In these comparison experiments only direct translation / matching was examined since it is not sensible to study indirect fuzzy matching. Also,

indirect TRT without frequency based selection of intermediate word forms for further translation would give very long queries that would be hard to manage.

The skipgram fuzzy matching technique constructs digrams both of consecutive and non-consecutive characters of words [Keskustalo et al. 2003]. The generated digrams are put into comparable categories based on the number of skipped characters as digrams are constructed. The character combination index (CCI) indicates the number of skipped characters as well as the comparable categories. Here we used the CCI=([0], [1, 2]). This means that digrams formed of consecutive characters form one comparable category and digrams with one and two skipped characters the other. CCI=([0], [1, 2]) was very effective in the study conducted by Keskustalo et al. [2003] who explored the same general problem as we do in this paper, i.e., the identification of translation equivalents of cross-lingual spelling variants. Skipgrams with CCI=([0], [1, 2]) outperformed conventional digrams formed of consecutive characters of words.

In the skipgram experiments each OOV word was matched against each string in the index of the TREC collection. Two types of queries were constructed: for each OOV word (1) two best matches and (2) five best matches were selected for a query. In the query the best matches were linked to each other with InQuery's #syn-operator.

All inflected query words were rendered into base form for a dictionary look-up. For Finnish, UTACLIR's morphological analyzer gave base forms for most inflected words, and those that the analyzer was not able to handle were lemmatized manually. All Spanish inflected words were lemmatized manually. Manual lemmatization of the inflected keys was necessary because at this stage of development TRT only translates lemmas. Thus, the results show CLIR performance when a searcher gives query keys in base form.

The results were tested for significance by the Wilcoxon signed rank test [Conover 1980]. The Wilcoxon test takes into account both the direction and magnitude of change between each comparable result of a query.

In summary, we run the following queries in Spanish-English and Finnish-English CLIR experiments:

- Original English queries
- UTACLIR baseline (no OOV word technique)
- UTACLIR + FITE-TRT (Web)
- UTACLIR + FITE-TRT (frequency lists)
- UTACLIR + TRT
- UTACLIR + skipgrams with two best matches
- UTACLIR + skipgrams with five best matches

5.2.2 Findings

The results of the retrieval experiments are presented in Tables 9 - 11. The statistical significance of the test queries was tested against UTACLIR baseline (Tables 9-10) and against UTACLIR + FITE-TRT/Web (Table 11). In tables the statistical significance is indicated by asterisks.

As expected, the queries where OOV keys are translated by FITE-TRT perform better than the baseline queries where OOV keys are retained untranslatable (Tables 9 and 10). In Spanish-English CLIR MAP improvement percentages are 40.3% (Web) and 35.2% (frequency lists). Precision at 20 documents is improved by 50.5% (Web) and 49.2% (frequency lists). Also Finnish-English FITE-TRT queries remarkably outperform the CLIR baseline although performance improvements are smaller than in the case of Spanish-English (Table 10). All Spanish-English and Finnish-English results are statistically significant at p=0.001. These findings are in agreement with the high number of OOV keys in the UTACLIR runs and FITE-TRT's high translation recall and precision. It should be noted that some OOV words may only be marginally topical, in which case correct FITE-TRT identification may result in a performance drop. Therefore FITE-TRT performance does not always correlate linearly with CLIR+ FITE-TRT performance.

Spanish-English FITE-TRT queries perform better than Finnish-English FITE-TRT queries, which were much longer and more ambiguous than Spanish-English queries due to the larger coverage of UTACLIR's Finnish-English dictionary. The higher degree of translation ambiguity and FITE-TRT's lower performance resulted in lower CLIR performance.

The higher performance of English queries w.r.t. to the performance of Spanish-English and in particular Finnish-English queries is mostly caused by translation ambiguity.

Table 11 shows the performance of FITE-TRT, TRT, and skipgram queries. The results are ranked based on MAP values. It can be seen that for both language pairs the best OOV word method is FITE-TRT with Web document frequencies. However, it shows significantly better results only against the cases of Spanish-English/skipgram/2 and Finnish-English/TRT, and the results are significant only at p=0.05. In the latter case, difference in MAP is small (0.2447-0.2393) but systematic and hence significant. In comparison to UTACLIR baselines (Tables 9 and 10) all queries perform well. It was expected that plain TRT queries yield good results since TRT with CF=4% and frequency=2 very often gives a source word's correct equivalent while the other

translations typically are malformed word forms not occurring in the database index and having no effects whatsoever on retrieval results. However, we expect that the effectiveness of FITE-TRT based CLIR can still be improved e.g. by augmenting FITE-TRT with word class information while TRT-based CLIR showed here its near upper limit.

Table 9. Spanish-English CLIR performance. Indirect translation is involved.
(*** statistical significance level 0.001)

Query type	MAP	% change wrt Utaclir	Pr. at 20 docs	% change wrt Utaclir
Baselines				
English queries	0.3195	-	0.5152	-
Utaclir baseline	0.2018	-	0.3009	-
FITE-TRT queries				
Web	0.2832***	+40.3	0.4530***	+50.5
Frequency lists	0.2728***	+35.2	0.4490***	+49.2

Table 10. Finnish-English CLIR performance. Indirect translation is involved.
(*** statistical significance level 0.001)

Query type	MAP	% change wrt Utaclir	Pr. at 20 docs	% change wrt Utaclir
Baselines				
English queries	0.3195	-	0.5152	-
Utaclir baseline	0.1971	-	0.3047	-
FITE-TRT queries				
Web	0.2491***	+26.4	0.4420***	+45.1
Frequency lists	0.2480***	+25.8	0.4460***	+46.4

Table 11. The performance of FITE-TRT, TRT, and skipgram queries. Indirect translation is not involved. (* statistical significance level 0.05)

Spanish-English	MAP	Finnish-English	MAP
FITE-TRT/Web	0.2814	FITE-TRT/Web	0.2447
TRT	0.2796	FITE-TRT/Frequency lists	0.2436
FITE-TRT/Frequency lists	0.2691	skipgram/5 best matches	0.2411

skipgram/5 best matches	0.2611	skipgram/2 best matches	0.2395
skipgram/2 best matches	0.2473*	TRT	0.2393*

6. DISCUSSION AND CONCLUSIONS

In this study we first examined the following two basic questions. Regarding spelling variants, 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? We devised the FITE-TRT technique - a novel OOV word translation technique. Its effectiveness was tested for Spanish-English and Finnish-English spelling variants and actual OOV words in the domains of biology and medicine. Here the research questions were as follows. What are the translation recall and precision and indication precision of the proposed FITE-TRT method? Are word frequency lists mined from the Web competitive with the Web as a collection of documents as FITE-TRT's frequency source? What is the contribution of each step (direct and indirect translation routes) in the FITE-TRT process to its overall effectiveness?

We found that depending on the source language and frequency source, FITE-TRT may achieve high translation recall. When equivalents were identified on the basis of Web document frequencies, Spanish-English FITE-TRT achieved 89.2%-91.0% recall. For Finnish-English and for frequency lists mined from the Web, recall was lower but still substantial (67.5 – 72 % ??). The results indicated that FITE-TRT achieves high precision. FITE indicates precisely the equivalents of source words as well as the native words. For Spanish-English and Finnish-English test words translation precision was 95.0%-98.8%. All native OOV words were correctly indicated to be untranslatable by TRT. The test requests only contained 10 native OOV words, and the results reported here need to be corroborated using a larger set of native words. The contribution of direct translation to the overall recall was substantial for Spanish-English bio-terms. Direct translation achieved 82.0% recall while the overall recall was 91.0%. We expected that Finnish-English FITE-TRT through pivot languages would have compensated failures of direct translation (only 72 %) but it did not happen. Indirect translation did not help at all. These findings suggest that that the costs of indirect translation against its benefits are high. Because a pivot language increases FITE-TRT complexity it does not seem sensible to use two pivot languages as part of a FITE-TRT system.

Lastly, we examined what is the effectiveness of a standard CLIR system boosted by the use of FITE-TRT in comparison to a CLIR system augmented with TRT and fuzzy matching OOV word methods, and in comparison to dictionary-translation-only CLIR and monolingual baselines. It was shown that FITE-TRT with Web document frequencies was the best technique among several approaches to handling OOV words tested in the experiments. Dictionary-based CLIR augmented with FITE-TRT performed substantially better than the baseline where OOV keys were kept intact. In Spanish-English CLIR MAP improvement percentages were 40.3% (Web) and 35.2% (frequency lists). Also Finnish-English FITE-TRT queries remarkably outperformed the CLIR baseline although performance improvements were smaller than in the case of Spanish-English (about 26% for both Web and frequency lists).

The TRT program we used in this study was not able to process a high number of word forms in a reasonable time frame, and we had to apply CF and frequency thresholds. In the preliminary tests we translated without using any thresholds, and for long words we had to end the program because it was not able to complete the translation process within a day. We observed that for many source words equivalents were not found in translation sets because the CFs and/or frequencies of the relevant rules were below the thresholds. We therefore expect that recall values can still be improved by using a more sophisticated TRT program which allows efficient TRT even without the use of thresholds. For example, in the case of Spanish / bio-terms / Web there were 81 correct identifications, one wrong identification, and seven words for which TRT did not identify equivalents. For five of the seven words equivalents were not contained in the translation sets because of low CF or frequency rules. The five words and the rules are shown in Table 12.

Table 12. Low CF/frequency rules in the test situation of Spanish-English/bio-terms/Web.

Spanish word	English equivalent	Rule
bromo	bromine	mo mine e 1 195 0.51
alucinogeno	hallucinogen	a ha b 4 1891 0.21
seudoefedrina	pseudoephedrine	s ps b 13 907 1.43
espinaca	spinach	ca ch e 3 832 0.36
estricnina	strychnine	ric ryc m 1 284 0.35

Also deficiencies in the Finnish-English rule collections and word frequency lists caused recall errors. FITE-TRT effectiveness was better for Spanish-English than Finnish-English. The better effectiveness can be attributed to the higher quality of

Spanish-English rule collection. Deficiencies in the Finnish-English transformation rules can be overcome by using more data in rule generation. The frequency lists we constructed using Web mining turned out to be good frequency data for FITE. However, for some source word / equivalent pairs frequencies were too low for rel-wf formula, which resulted in a decrease in recall performance. This deficiency can be overcome by adding more data to the existing lists.

The main advantage of using frequency lists is that after the lists have been constructed FITE-TRT is independent of the Web. This is important when a practical FITE-TRT implementation is developed.   The frequency lists we used contain in particular biological and medical terms in accordance with the test terminology used in the study. The lists are large and contain terms in various domains. The application of FITE-TRT in the other domains may require lists with different types of terms. However, for each domain, the lists are concise enough to be held in main memory for efficient implementation.

Overall the percentage of wrong equivalents indicated by FITE was small. The identification of derivatives and plural forms of the correct equivalents was the primary cause of precision errors. As an example, for the Finnish word *leukosyytti* and the Spanish word *bacteria* the correct equivalents are *leucocyte* and *bacterium* while FITE gave the words *leucocytic* and *bacteria*. Many of these types of cases could be solved by augmenting the transformation rules with word class information, e.g. that a rule is likely to refer an adjective rather than a noun. Information on OOV word's word class is achieved when the sentential context of the OOV word is known.

At present TRT is only intended to translate singular base forms. Word class information is needed if FITE-TRT will be applied to running texts containing inflectional word forms. This would imply the supplementation of rule collections with word class information. The next main challenge in the FITE-TRT research is to improve the rules such that FITE-TRT can handle words in a running text.

The CLIR experiments showed that the best fuzzy translation / matching based query was FITE-TRT with Web document frequencies. The FITE component of FITE-TRT was the focus of this paper, and here an important issue is its contribution in CLIR. The results showed that in the case of Finnish-English, FITE-TRT was significantly better than plain TRT but only at p=0.05. In the case of Spanish-English, FITE-TRT did not show significantly better results. Overall, the results are inconclusive, and the issue needs to be investigated more thoroughly in future research. Such factors as query structure and other than OOV keys affect the effectiveness of FITE-TRT and TRT queries. Also

efficiency needs to be taken into account in future research. TRT queries are much longer than FITE-TRT queries and thus require more processing power. On the other hand, the FITE component of FITE-TRT increases computational expenses.

In many other information systems than retrieval systems, in particular in MT fuzzy translation (TRT) does not come into question. The good quality of translations achieved through FITE-TRT suggests that it can contribute to better MT performance.

There is still one CLIR-related application of FITE-TRT which is worth mentioning, that is, an automatic construction of multilingual dictionaries of technical terms and proper names by means of FITE-TRT. The dictionary construction process could be designed to be largely automatic thanks to FITE-TRT's high effectiveness. Given a list of words and a set of transformation rule collections the process would automatically yield the translation equivalents of the words in different languages – the result would essentially be a multilingual dictionary. The construction of dictionaries is a non-time-critical task, and given enough time it would be possible to construct large multilingual dictionaries. The cost benefits of an automatic method are obvious. It can easily be seen that there is a difference in the cost of automatic vs. manual construction of a dictionary of, say, 10 languages and 50 000 dictionary entries.

ACKNOWLEDGMENTS

REFERENCES

AL-ONAIZAN, Y., CURIN, J., JAHR, M., KNIGHT, K., LAFFERTY, J., MELAMED, D., OCH, F.J., PURDY, D., SMITH, N.A., AND YAROWSKY, D. 1999. Statistical machine translation: Final report. *Johns Hopkins University 1999 Summer Workshop on Language Engineering*.

ALLAN, J., CONNELL, M.E., CROFT, W.B., FENG, F-F., FISHER, D., AND LI, X. 2000. Inquery and TREC-9. In *Proceedings of the Ninth Text REtrieval Conference (TREC-9)* (Gaithesburg, MD). http://trec.nist.gov/pubs/trec9/t9_proceedings.html

BALLESTEROS, L.A. 2000. Cross language retrieval via transitive translation. In *Advances in Information Retrieval: Recent Research from the CIIR*. Kluwer Academic Publishers, 203-234.

BROWN, P., COCKE, J., DELLA PIETRA, S., DELLA PIETRA, V., JELINEK, F., LAFFERTY, J., MERCER, R., AND ROOSIN, P. 1990. A statistical approach to machine translation. *Comp. Ling. 16*, 79-85.

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. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Sheffield, UK). ACM, New York, 146-153.

CONOVER, W.J. 1980. *Practical nonparametric statistics*. John Wiley & Sons, 493 pages.

FUJII, A., AND ISHIKAWA, T. 2001. Japanese/English cross-language information retrieval: Exploration of query translation and transliteration. *Comput. Humanit. 35*, 4, 389–420.

GOLLINS, T., AND SANDERSON, M. 2001. Improving cross language retrieval with triangulated translation. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New Orleans, Louisiana). ACM, New York, 90 – 95.

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. *Inf. Ret. 7*, 99-119.

HERSH, W.R., BHUPTIRAJU, R.T., ROSS, L., JOHNSON, P., AND KRAEMER, D.F. 2005. TREC 2004 genomics track overview. In *Proceedings of the Thirteenth TExt REtrieval conference (TREC-13)* (Gaithersburg, MD).
http://trec.nist.gov/pubs/trec13/t13_proceedings.html

KESKUSTALO, H., PIRKOLA, A., VISALA, K., LEPPÄNEN, E., AND JÄRVELIN, K. 2003. Non-adjacent digrams improve matching of cross-lingual spelling variants. In *Proceedings of the10th International Symposium on String Processing and Information Retrieval (SPIRE 2003)* (Manaus, Brazil), 252 – 265.

LARKEY, L.S., AND CONNELL, M.E. 2005. Structured queries, language modeling, and relevance modeling in cross-language information retrieval. *Inf. Proc. Manage. 41*, 3, 457-473.

LEHTOKANGAS, R., AIRIO, E., AND JÄRVELIN, K. 2004. Transitive dictionary translation challenges direct dictionary translation in CLIR. *Inf. Proc. Manage. 40*, 6, 973-988.

MENG, H., CHEN, B., GRAMS, E., KHUDANPUR, S., LO, W-K., LEVOW, G-A., OARD, D., SCHONE, P., TANG, K., WANG, H-M., AND WANG, J. 2000. Mandarin English Information (MEI): Investigating translingual speech retrieval. *John Hopkins University Summer Workshop 2000*. 64 pages.

PETERS, C. 2005. CLEF - Cross-Language Evaluation Forum. http://clef.isti.cnr.it/

PIRKOLA, A. 1998. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia). ACM, New York, 55-63.

PIRKOLA, A., TOIVONEN, J., KESKUSTALO, H., VISALA, K., AND JÄRVELIN, K. 2003. Fuzzy translation of cross-lingual spelling variants. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Toronto, Canada). ACM, New York, 345 – 352.

PIRKOLA, A., TOIVONEN, J., KESKUSTALO, H., AND JÄRVELIN, K. 2006. FITE-TRT: A high quality translation technique for OOV words. In *Proceedings of the 21st Annual ACM Symposium on Applied Computing* (Dijon, France), April 23 - 27, 2006.

SPERER, R., AND OARD, D. 2000. Structured translation for cross-language IR. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Athens, Greece). ACM, New York, 120-127.

TOIVONEN, J., PIRKOLA, A., KESKUSTALO, H., VISALA, K., AND JÄRVELIN, K. 2005. Translating cross-lingual spelling variants using transformation rules. *Inf. Proc. Manage. 41*, 4, 859-872.

ZHANG, Y AND VINES P. 2004. Using the Web for automated translation extraction in cross-language information retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Sheffield, UK). ACM, New York, 162-169.

APPENDIX

FITE identification of the target language equivalent for a source language word by the algorithm *find-equivalent*:

input

   array $S = \{<w_1, f_1>, \ldots, <w_n, f_n>\}$ /* the source language *SL* word frequency list

                                              /* – see Section 4.1

   array $T = \{<v_1, g_1>, \ldots, <v_m, g_m>\}$ /* the target language *TL* word frequency list

   rulebase *SL->TL*   /* the source-target TRT rule base – Section 2

   rulebase *SL->PL*   /* the 1st source-pivot TRT rule base

   rulebase *SL->QL*   /* the 2nd source-pivot TRT rule base

   rulebase *PL->TL*   /* the 1st pivot-target TRT rule base

   rulebase *QL->TL*   /* the 2nd pivot-target TRT rule base

   integer α, β /* the corpus coefficients – Sections 4.2-4.3

output

   string *equivalent*   /* the *TL* equivalent or *nil*

procedure find-equivalent(*sw*): *String*   /* finds the *equivalent* for the source

                                              /* language word *sw*

   $i \leftarrow 1$;   /* try the candidate word position 1 in *R*

A: *equivalent* $\leftarrow$ direct-trans(*sw, i, SL, TL*);   /* Step A: direct translation at *i*th position

   if *equivalent* = *nil*   /* not found

      then $XL \leftarrow PL$;   /* Step B: pivot language is *PL*

         equivalent $\leftarrow$ pivot-trans(*sw, i, SL, XL, TL*);   /* try first pivot translation

        if *equivalent* = *nil*   /* not found

            then $XL \leftarrow QL$;   /* Step C: pivot language is *QL*

          equivalent $\leftarrow$ pivot-trans(*sw, i, SL, XL, TL*);

                                       /* try 2nd pivot translation

   if *equivalent* = *nil* and $i = 1$ /* still not found for the 1st position

     then $i \leftarrow 2$;   /* try the candidate word position 2 in *R*

        goto A   /* go through steps A – C

   else output *equivalent*.   /* output the translation or "nil"

procedure test-cand(*tw1*, *tw2*, *sw*, *tw*): *Boolean*   /* test criteria 1-3 of *tw1* against *tw2* and *sw*

  if *freq-pattern-ok*(*tw1*, *tw2*, β, *T*)   /* criterion 1 : Definition 1

    and *rel-df*(*tw1*, *sw*, α, *T*, *S*) > 1   /* criterion 2 : Definitions 2-3

    and *tw-len-ok*(*tw*, *sw*)   /* criterion 3 : Definition 4

    then true

else false.

procedure direct-trans(*sw, i, SL, TL*): *String*   /* translate directly into *TL*

   $R \leftarrow$ *trt-frank*($TRT_{SL->TL}(sw), T$)   /* generate frequency ranked *TL* candidates

   if test-cand(*R*[i], *R*[i+1], *sw*, *R*[1])   /* test criteria 1-3 for the *i*th candidate in *R*

      then *equivalent* $\leftarrow$ *R*[1]   /* the equivalent was found – the 1st component

   else if test-cand(*R*[i], *R*[i+1], *sw*, *R*[*i*])   /* test criteria 1-3 for the *i*th candidate in *R*

      then *equivalent* $\leftarrow$ *R*[*i*]   /* the equivalent was found – the 2nd component

   else *equivalent* $\leftarrow$ *nil*.   /* the equivalent was not satisfactory

procedure pivot-trans(*sw, i, SL, XL, TL*): *String*   /* translate via pivot language

   if $|TRT_{SL->XL}(sw)| \leq 40$ /* check if too many candidates generated

      then $PWS \leftarrow TRT_{SL->XL}(sw)$;   /* produce pivot language candidate set for *sw*

         $TWS \leftarrow \cup_{pw \in PWS}TRT_{XL->TL}(pw)$;   /* produce *TL* candidate set for strings in *PWS*

   else $PWS \leftarrow TRT_{SL->XL|strict}(sw)$;   /* produce strict pivot language candidate

                                                /* set for *sw*

        $TWS \leftarrow \cup_{pw \in PWS}TRT_{XL->TL|strict}(pw)$;   /* produce strict *TL* candidate set for

                                              /* strings in *PWS*

   $R \leftarrow$ *trt-frank*(*TWS, T*)   /* produce ranked *TL* translations for

                                              /* candidates in *TWS*

   if test-cand(*R*[i], *R*[i+1], *sw*, *R*[1])   /* test criteria 1-3 for the *i*th candidate in *R*

      then *equivalent* $\leftarrow$ *R*[1]   /* the equivalent was found – the 1st component

   else if test-cand(*R*[i], *R*[i+1], *sw*, *R*[*i*])   /* test criteria 1-3 for the *i*th candidate in *R*

      then *equivalent* $\leftarrow$ *R*[*i*]   /* the equivalent was found – the 2nd component

   else *equivalent* $\leftarrow$ *nil*.   /* the equivalent is "nil"