

Dictionary-independent translation in CLIR between closely related languages

Anni Järvelin
+46-480-411662
anni.jarvelin@uta.fi

Sanna Kumpulainen
+358-505298901
sanna.kumpulainen@uta.fi

Ari Pirkola
+358-14-762278
pirkola@cc.jyu.fi

Eero Sormunen
+358-3-35516972
eero.sormunen@uta.fi

Department of Information Studies, FIN 33014, University of Tampere, Finland

ABSTRACT

This paper presents results from a study, where fuzzy string matching techniques were used as the sole query translation technique in Cross Language Information Retrieval (CLIR) between the closely related languages Swedish and Norwegian. It is a novel research idea to apply only fuzzy string matching techniques in query translation. Closely related languages share a number of words that are cross-lingual spelling variants of each other. These spelling variants can be translated by means of fuzzy matching. When cross-lingual spelling variants form a high enough share of the vocabulary of related languages, the fuzzy matching techniques can perform well enough to replace the conventional dictionary-based query translation. Different fuzzy matching techniques were tested in CLIR between Norwegian and Swedish and it was found that queries translated using skipgram matching and a combined technique of transformation rule based translation (TRT) and n-grams perform well. For the best fuzzy matching query types performance difference with respect to dictionary translation queries was not statistically significant.

Categories and Subject Descriptors

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

General Terms

Performance, Experimentation

Keywords

Cross-language retrieval, Fuzzy matching

1. INTRODUCTION

Information retrieval methods are based on comparing the words in requests with the words in documents. Cross Language Information Retrieval (CLIR) refers to the retrieval of documents in other languages than the language of the request. For an overview of different approaches to CLIR, see [6]. Fuzzy string matching methods are used for finding matches between words

that are similar but not identical. In CLIR fuzzy string matching has been used for handling proper names and technical terms, as well as other cross-lingual spelling variants not found in translation dictionaries [5, 12, 17]. McNamee and Mayfield [7] have used n-grams in corpus-based query translation.

Closely related languages have not been considered as a separate line of research in CLIR. The dominating approach, dictionary-based translation of queries, is a fairly effective technique, but has its problems in the limited coverage of dictionaries and the constant need for updating, which can make it an expensive technique. Closely related languages typically share a high number of spelling variants, i.e., equivalent words that share the same origin and are similar (but not identical). If the number of the shared cross-lingual variants is high enough, query translation can be handled by much cheaper and simpler fuzzy techniques.

Among fuzzy techniques n-gram and skipgram matching have been found to be effective in monolingual proper name [10] and cross-lingual spelling variant matching [5, 12] and transformation rule based translation technique (TRT) has been found to be an effective method for translating cross-lingual spelling variants [17]. N-grams and skipgrams are language independent techniques and the TRT technique can be easily adjusted for new language pairs. The methods are therefore easily applicable for new languages and thus ideal translation methods in CLIR. They are not dependent on expensive linguistic resources. In this study we used these dictionary-independent fuzzy string matching techniques as a query translation technique between closely related languages. The techniques were tested with the Scandinavian language pair Norwegian and Swedish, with Norwegian as the source language and Swedish as the target language.

Scandinavian languages have not been studied extensively from the information retrieval point of view. Hedlund et al. [3] is an exception. In their study characteristics of Swedish in information retrieval were analyzed. Swedish and Norwegian together with Danish form a language group where the speakers of one language can quite easily understand the other languages, especially in written form. Both the grammar and vocabulary of the languages are similar as they have developed in a close historical and cultural relation to one another. Some 50% of the Swedish and Norwegian (Bokmål) vocabulary is identical and around 40% similar, when inflected word forms and orthographical differences of using æ/ø instead of ä/ö are not considered [1]. There are also consistent and frequently occurring differences in the orthographies of Swedish and Norwegian. For

example, Norwegian avoids the use of letters c, z, and x (*center* (Swe) – *senter* (No)) and the letter d is often left out of words where Swedish has it (*kunde* (Swe) – *kunne* (No)), the Danish letters æ/ø are used in Norwegian instead of Swedish ä/ö and the Swedish word endings -sion, -ssion and -tion are written -sjon in Norwegian. These similar features suggest that the use of fuzzy string matching techniques and the statistical transformation rules might be efficient in query translation from Norwegian to Swedish.

The research problems investigated in this paper are as follows:

1. Are fuzzy string matching methods as effective as the dictionary-based translation techniques in CLIR between closely related languages like Norwegian and Swedish?
2. Which of the fuzzy string matching methods tested is the most suitable translation technique for CLIR between closely related languages?

To the best of our knowledge, attempting to solve the query translation problem in CLIR between closely related languages with fuzzy string matching techniques without dictionary translation is a novel research idea not tried before.

The rest of the paper is organized as follows. The fuzzy string matching techniques used in this study are introduced in Section 2. Section 3 presents the test environment, methods and data. The similarity between Norwegian and Swedish is discussed in Section 4. Section 5 presents the findings and Section 6 discussion and conclusions for the study.

2. TRANSLATION TECHNIQUES

2.1 N-grams and Skipgrams

N-gram matching is a language independent method for matching words whose character strings are similar [13, 14]. Query keys and words in documents are decomposed into n-grams, i.e. into substrings of length n . The degree of similarity between the query keys and index terms can then be computed by comparing their n-gram sets. For a description of the applications of the technique, see [14]. N-gram matching has been reported to be an effective technique among fuzzy string matching techniques in name searching [10] and in cross-lingual spelling variant matching [5]. McNamee and Mayfield [7] used a direct corpus-based n-gram query translation technique, where the source language n-grams were directly translated to the target language n-grams using aligned corpora. The translation technique using 4- and 5-grams was found feasible. They also found n-grams an effective technique in tokenization, as it outperformed the stemmer used. Also Adafre et al. [1] have used 4-grams combined to a parallel corpus in query translation.

N-grams can consist both of adjacent characters or non-adjacent characters of the original words. Pirkola et al. [12] devised a novel matching technique for n-grams formed of non-adjacent characters, called the classified skipgram matching technique. In this technique digrams are divided into categories (classes) on the basis of the number of the skipped characters and only the digrams belonging to the same class are compared with each other. *Gram class* indicates the number of skipped characters when digrams are formed from a string S . *Character combination index* (CCI) then indicates a set of gram classes enumerating all the digram sets to be produced from the string S . For example

$CCI = \{\{0\}, \{1,2\}\}$ means that two gram classes are formed from the string: one with conventional digrams formed of adjacent characters and one with skip-digrams formed both by skipping one and two characters [5]. The classified skipgrams have performed better than the traditional n-grams in the empirical tests examining the matching of cross-lingual spelling variants [5, 12].

It is common to use padding spaces in the beginning and in the end of the strings when forming n- and skipgrams. If the padding spaces are not used, the characters at the front and at the end of the strings will be under-represented in the gram set that is generated. Keskustalo et al. [5] tested different types of padding spaces for conventional digrams, trigrams and skipgrams, and found that using padding spaces both in the beginning and the end of the words gave the best results. However, the use of end padding spaces has been found unsuitable for inflectionally complex suffix languages, such as Finnish, where the use of the beginning padding only has been found beneficial [12]. This way of down-weighting the word ends – the inflectional suffixes – was assumed to be useful also when handling Swedish and Norwegian. For n-grams it is common to use a padding of n-1 characters [14]. For skipgrams a padding that varies according to the number of the skipped characters can be used.

The similarity values for n-grams are computed with a string similarity scheme [10]:

$$SIM(w_1, w_2) = \frac{|N_1 \cap N_2|}{|N_1 \cup N_2|}, \text{ where}$$

N_i is a digram set of a string w_i , $|N_1 \cap N_2|$ denotes the number of intersecting n-grams in strings w_1 and w_2 , i.e. n-grams that the strings have in common, and $|N_1 \cup N_2|$ denotes the number of unique n-grams in the union of N_1 and N_2 . The similarity measure for skipgrams is then defined between two strings S and T with respect to the given CCI as follows [5]:

$$SIM_{CCI}(S, T) = \frac{\sum_i i_{CCI} |DS_i(S) \cap DS_i(T)|}{\sum_i i_{CCI} |DS_i(S) \cup DS_i(T)|}, \text{ where}$$

DS_i is the digram set of a string, i denoting the gram class, $|DS_i(S) \cap DS_i(T)|$ denotes the number of intersecting n-grams and $|DS_i(S) \cup DS_i(T)|$ the number of unique n-grams in the union of the strings S and T .

2.2 The TRT Technique

Transformation rule based translation (TRT) is a fuzzy translation technique based on the use of statistically generated rules of regular character correspondences in cross-lingual spelling variants within a language pair. The technique resembles transliteration, phonetic translation across languages with different writing systems, but no phonetic elements are included and the technique is meant for processing languages sharing the same writing system. It is applied in two-steps: the transformation rules are combined to n-gram matching. The idea of the TRT and the generation of the transformation rules are described in more detail in [17].

A *transformation rule* contains source and target language characters and their context characters [17]. In addition the frequency and the confidence factor of the rule are recorded. *Frequency* refers to the number of the occurrences of the rule in the data used for generating the rules. *Confidence factor* is the frequency of a rule divided by the number of source words where the source substring of the rule occurs. They are important threshold factors that can be used for selecting the most reliable rules for the translation. An example of a Norwegian to Swedish rule is:

for för beginning 132 147 89.80

The rule can be read: the letter o, prior to r and after f, is transformed into the letter ö in the beginning of words, with the confidence factor being 89.80. The confidence factor is calculated from the frequency of the rule (132) and the number of source words where the string occurs (147).

In this study we used the thresholds of confidence factor = 50% and frequency = 2.

3. METHODS AND DATA

3.1 Test Topics and Collection

The performance of the fuzzy translation methods was tested by running CLIR tests with a set of 60 topics used in the CLEF evaluation forum in the year 2003 [9]. Norwegian and Swedish topics were used, of which Swedish topics were included in the collection of the CLEF topics. To get the Norwegian topics, English topics were translated by a native Norwegian speaker. Of the two official Norwegian languages the more common Bokmål was used. In ten of the topics, queries failed in preliminary test runs for technical reasons. These topics were removed from all of the queries and the final tests were run with the remaining 50 topics. The target document collection was the Swedish CLEF document collection containing 142819 newspaper articles obtained from the Swedish news agency TT (Tidningarnas Telegrambyrå) published in 1994-1995 [9]. The document collection was lemmatized using Swetwol morphological analyzer by Lingsoft Inc. Compounds were split into their constituents and both the original word and the constituents were lemmatized and indexed. Words not recognized by the morphological analyzer were indexed as such to a separate index of unrecognized words. We used the InQuery Retrieval System as the search engine. InQuery is a probabilistic information retrieval system based on the Bayesian inference net model, where queries can be presented as unstructured bag-of-words queries or they can be structured with a variety of operators [2].

3.2 Creating TRT Rules

To create the word-pair list used for generating the Norwegian to Swedish transformation rules a part of the Swedish document collection's index was translated to Norwegian with the Global Dix dictionary by Kielikone plc. Words not recognized by the morphological analyzer were removed and, as the index was too large to use as a whole, every sixth word of the index was chosen. This list contained 6714 word-pairs. Word-pairs with an edit distance value bigger than half of the length of the longer word in the word-pair or including a word shorter than four characters were removed. The final word-pair list included 3058 unique word-pairs. This list seemed to be insufficient for generating

enough high frequency transformation rules. This lack of high quality rules may have affected negatively the TRT technique's translation results.

3.3 N- and Skipgram Matching

The n- and skipgram translations were done by matching the n- or skipgrams of the topic words against the normalized index words of the Swedish test collection. The index was divided into two: the index of the words recognized by the morphological analyzer and the index of unrecognized words. Dividing the index is helpful when matching proper names [4]. For n-digram translation we used beginning weighted n-digrams with the padding of 1. Leaving out the padding at ends of words gives more weight to the beginnings of words, which can be useful when the words are inflected. For skipgram translation, a padding of the number of the skipped characters + 1 was used. For example for gram class 1, the skipgrams were formed with two padding spaces.

3.4 Queries

We used five sets of test queries, which were compared to three sets of baseline queries. The five translation methods tested were n-digrams, classified skipgrams with $CCI = \{\{0\}\{1\}\}$ (*Skip1*) and $CCI = \{\{0\}\{1,2\}\}$ (*Skip2*), plain TRT translation and the combined TRT and n-digram technique. The set of baseline queries consisted of a monolingual Swedish query set (*Swebase*), a monolingual Norwegian query set (*Nobase*) and a dictionary translated Norwegian to Swedish query set (*Dicbase*). The Global Dix dictionary was used for the translations. The Swebase and Dicbase gave high performing baselines, while the Nobase was used for testing how high performance is achieved without any translation and how much the fuzzy methods can improve this result.

The test query types were as follows. The query operators used in a query are presented in parentheses and examples of the queries are presented in Appendix 1.

- 1) Swedish monolingual baseline (#sum)
- 2) Norwegian monolingual baseline (#sum)
- 3) Dictionary baseline (#sum, #syn, #uw7)
- 4) N-digram query (#sum, #syn)
- 5) Skip1 query (#sum, #syn)
- 6) Skip2 query (#sum, #syn)
- 7) Plain TRT query (#sum, #syn)
- 8) Combined TRT and n-digram query (#sum, #syn)

The queries were formed from the title- and description fields of the CLEF topics. The topic words were lemmatized with the morphological analyzer Twol. For the dictionary translation, compound words were split into constituents that were then translated separately. This is because compound components are more often found in dictionaries than the whole compounds. For other query types, no compound splitting was done, as we assumed the compounds in Norwegian and Swedish to be similar. The lemmatized source words were translated and stop words were removed both before and after the translation.

The queries were formulated by grouping the query keys with InQuery's operators *sum*, *syn* and *uwn*. The sum-operator computes an average of query key weights for keys grouped by the operator. It is used for grouping the whole query and can include either the query keys without any structure or query key sets structured with the other operators. The syn-synonym operator treats its operand query keys as synonyms. The unordered proximity operator with a window size n (*uwn*) allows free word-order and combines the translations equivalents of the constituents of a source language compound [13].

The Swedish and Norwegian monolingual baseline queries were formed directly from the Swedish and Norwegian topic words as bag-of-words queries without any structure. The rest of the queries were structured with the syn-structure (*Pirkola's method*), which has been found effective in CLIR [11, 13, 16]. For the Dicbase queries all the translation equivalents of a source word were selected to the query and were grouped together with the syn-operator. When the translation was a noun phrase, its words were combined with a proximity operator of *uwn*, where we set the value of n to seven. Words not found in the dictionary were added to the query as such.

All the five test query types were structured queries, where the translation equivalents selected for a source word were grouped together with the syn-operator. For the n-gram and skipgram queries we selected for each source word the four highest ranked keys from the result list of n-gram matching. This selection was based on the findings by Hedlund et al. (2004), who showed that the best retrieval performance is achieved using just a few n-gram keys in queries [4]. These keys included two keys from the index of words recognized by the morphological analyzer and two from the index of unrecognized words.

For plain TRT-queries all the translated keys from the TRT result list were selected for each of the source word for the final queries. The combined TRT and n-digram queries were formed by selecting the first word form of each of the original source words from the TRT result list, which was then matched to the Swedish database index using n-digram matching. The word forms created with a rule combination with the highest confidence factor and frequency values get the highest position in the TRT result list. The four highest ranked keys from the result list of n-gram matching were then selected for the final queries like in other n-gram techniques.

3.5 Performance Measures

The effectiveness of the test queries was measured by Mean Average Precision (MAP) i.e., the average non-interpolated precision calculated over all relevant documents, and by interpolated recall precision averages at standard recall levels of 10 and 50, averaged over all queries. The test queries' precision-recall graphs were created using the eleven standard recall levels and the test queries' graphs were compared. The statistical significance of the results was tested using the Friedman two-way analysis of variance by ranks. The statistical significance levels are indicated in the tables.

4. SIMILARITY BETWEEN NORWEGIAN AND SWEDISH

To get an insight to how close two languages should be for the fuzzy matching to be practicable, the similarity of Swedish and Norwegian language was measured. A measure based on the Longest Common Subsequence (LCS) [8] was used, and German and English were used as a baseline language pair. They belong to the same language group but are not so closely related to make fuzzy matching alone a sufficient translation technique. The average similarity values measured for Swedish and Norwegian and for English and German were 0,815 and 0,556 respectively.

LCS is a measure that counts the maximum amount of letters that two words share and have in the same order, for example for an English - German word pair *motivation - motivierung* the longest common subsequence *motivin* has length 7. The data used for measuring the similarities between the languages included 167 word pairs for both language pairs. The vocabulary was selected from two sources: 71 words were chosen from the CLEF'03 topics and 96 words from a word list containing work environment vocabulary in all four languages (from the TNC-termbank by the Swedish national centre for terminology, TNC). The similarities were measured by first measuring the LCS values pair wise for all the words. Then each of these LCS values was divided by the length of the longer word of the word pair. Finally a mean value was calculated of these pair wise word similarity values for both language pairs. The similarity values range between 0-1. For example for the Swedish-Norwegian word pair *brevbomb - brevbombe* the LCS value is 8 and the similarity is counted by dividing it with the length of the longer of the words (here 9), with the similarity value being $8/9 \approx 0,889$.

Swedish, Norwegian and German are *compound languages* [4], i.e. languages where the components of multi-word expressions are written together, whereas English is a *non-compound language* where multi-word expressions are written as phrases (*fackförening, fagforening, gewerkschaft*, but *trade union*). The way the multi-word expressions are written is an important feature when measuring the orthographical similarity of languages. Therefore the test data included multi-word expressions. Phrases were written together by using a '_' to mark the space between the components (*trade_union*).

The similarity value of 0,815 measured for Swedish and Norwegian can be illustrated with examples: For a pair of short words such as *skola - skole* one character substitution results in a similarity value of 0,8. A longer word pair with a similarity value of 0,818 is *ioniserende - joniserande*, where two character substitutions happen. The orthographical differences in Swedish and Norwegian words are typically at this level. The mean similarity value of 0,556 measured for English and German corresponds to changes like *north_sea - nordsee*, which share five out of nine letters and get the similarity value of 0,556. The short word pairs *night - nacht* and *level - pegel*, where three letters out of five are common, get a similarity value of 0,6. The source of the vocabulary affected the similarity values slightly: the Swedish-Norwegian values for CLEF and TNC vocabularies were 0,829 and 0,805, respectively. English-German values were 0,582 for CLEF words and 0,536 for TNC words.

5. FINDINGS

5.1 The Performance of Fuzzy String Matching in Comparison to Baselines

Table 1 summarizes the Mean Average Precision values for all query types, and the performance differences between the test queries and the baseline queries. As the performances of the n-digram, skipgram and the combined TRT-n-gram queries were quite close to each other, they are referred together as the *n-gram queries* in the following. The performance differences between these queries are considered in Section 5.2.

The MAP is a measure that rewards techniques that retrieve relevant documents quickly [18]. When comparing the MAP values, the dictionary translation gives the best results, the monolingual Swedish baseline being second. The n-gram queries perform well: differences to the Dicbase and Swebase results are not statistically significant for any of the queries. The practical differences to the Dicbase are nevertheless noticeable (according to [15]) being over 5% for all fuzzy queries. All these techniques performed both statistically significantly and practically noticeably better than the Norwegian monolingual baseline. The plain TRT query's performance was better than the Nobase's, the difference not being statistically significant. The TRT query's performance was statistically significantly weaker than the Dicbase and Swebase baselines' performance.

Tables 2 and 3 present the recall precision averages at standard recall levels of 10 and 50. The Precision-Recall curves for all query types are shown in Figure 1. As can be seen from the P-R curves, the dictionary baseline and the Swedish monolingual baseline perform best on the high precision levels (0-20) and middle recall levels (20-80). For the high recall levels (80-100) the differences even up and the two skipgram queries perform as well as the Swebase baseline. Nobase and plain TRT queries still perform worse than the other queries.

At the recall level of 10 (Table 2), the dictionary baseline gets the highest precision average. The Swedish baseline is again the second best query type. The n-gram queries perform well, the differences to Dicbase and Swebase not being statistically significant. All the n-gram queries perform markedly better than the Nobase. Plain TRT query type is clearly worse than the Dicbase and Swebase baselines.

At the recall level of 50 (Table 3), the differences between different techniques diminish but the trend is still clear:

Dictionary translation gives the best result followed by the monolingual Swedish query. The n-gram queries perform also well, the difference to Dicbase and Swebase not having statistical significance, although the practical differences between n-gram queries and Dicbase are noticeable. The plain TRT queries and Nobase are clearly the two weakest query types; their differences to the other query types are statistically significant.

5.2 Best Fuzzy String Matching Technique

The fuzzy queries were also compared to each other to determine the most suitable technique for CLIR between closely related languages. As can be seen from Figure 1, the plain TRT queries' P-R curve is consistently clearly below the other curves. The difference to the other fuzzy queries is most of the time statistically significant or highly significant, and the practical difference is always noticeable. Therefore it can be concluded that, when used alone, it is not an adequate translation technique in CLIR between closely related languages. Earlier research results from [17] support this conclusion. In this research, the TRT queries' performance may have been negatively affected by the lack of high frequency transformation rules. This may also have affected the results of the combined TRT-n-gram queries.

The findings do not suggest one fuzzy string matching technique as being the best translation method in CLIR between closely related languages. The differences between the different n-gram queries were small and statistically insignificant. The combined TRT-n-gram queries performed best on the high precision levels and the practical difference to the plain n-gram queries was noticeable at the recall level of 10. On the middle recall levels all the n-gram queries were quite even and their differences had no statistical or practical significance at the recall level of 50. Here the skipgram queries gave the best results, the Skip2 -grams with CCI={{0}{1,2}} being the best query type. From the Figure 1 it can be seen that the P-R curves of skipgram queries are above the others fuzzy queries' curves at the high recall levels.

Even if the differences are small, the Skip2 queries and the combined TRT-n-gram queries performed slightly better than the other queries. At the same time, the combined TRT-n-gram queries outperformed the plain n-gram queries indicating that the transformation rules do improve n-gram results in CLIR between closely related languages.

Table 1. The MAP values (%) for the test queries and their difference to the baselines (%) (* statistically significant difference, ** statistically highly significant difference)

	Baseline queries			Test queries				
	Nobase	Swebase	Dicbase	Skip1	Skip2	N-gram	Plain TRT	TRT-n-gram
Precision	12,64	31,76	34,13	28,34	28,63	26,53	16,88	27,74
Difference to Nobase	0	19,12	21,49	15,7*	15,99*	13,89**	4,24	15,1**
Difference to Swebase		0	2,37	-3,42	-3,13	-5,23	-14,88**	-4,02
Difference to Dicbase			0	-5,79	-5,5	-7,6	-17,25**	-6,39

Table 2. The interpolated recall precision averages (%) at standard recall level 10 for the test queries, and their difference to the baselines. (* statistically significant difference, ** statistically highly significant difference)

	Baseline queries			Test queries				
	Nobase	Swebase	Dibase	Skip1	Skip2	N-gram	Plain TRT	TRT-n-gram
Precision	21,85	50,65	54,91	44,39	43,95	41,44	28,17	46,54
Difference to Nobase	0	28,8	33,06	22,54**	22,1*	19,59**	6,32	24,69**
Difference to Swebase		0	4,26	-6,26	-6,7	-9,21	-22,48**	-4,11
Difference to Dibase			0	-10,52	-10,96	-13,47	-26,74**	-8,37

Table 3. The interpolated recall precision averages (%) at standard recall level 50 for the test queries, and their difference to the baselines. (* statistically significant difference, ** statistically highly significant difference)

	Baseline queries			Test queries				
	Nobase	Swebase	Dibase	Skip1	Skip2	N-gram	Plain TRT	TRT-n-gram
Precision	13,1	31,03	35,64	28,81	29,58	27,02	15,78	28,77
Difference to Nobase	0	17,93	22,54	15,71	16,48	13,92*	2,68	15,67**
Difference to Swebase		0	4,61	-2,22	-1,45	-4,01	-15,25**	-2,26
Difference to Dibase			0	-6,83	-6,06	-8,62	-19,86**	-6,87

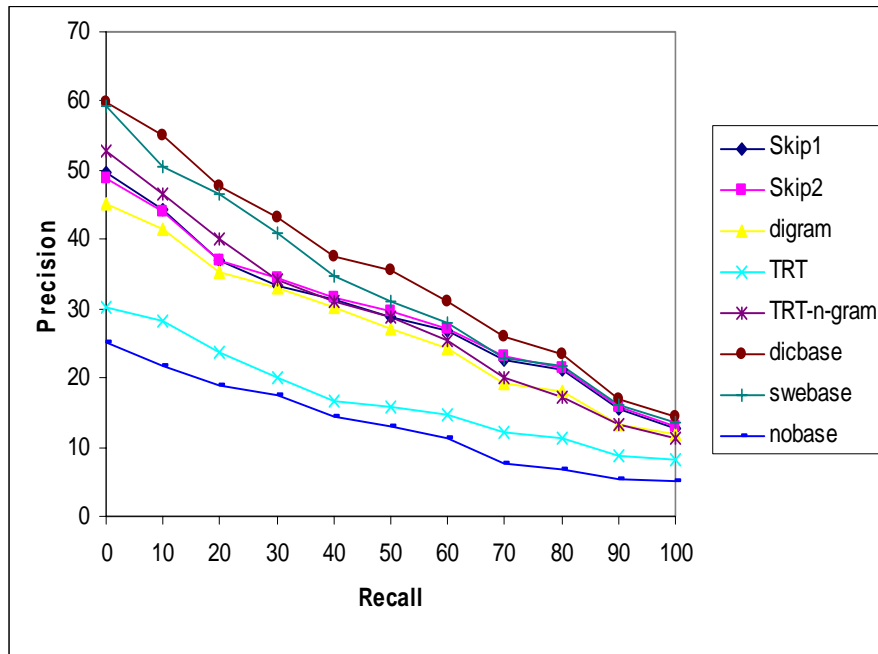


Figure 1. Recall-precision curves for all queries.

6. DISCUSSION AND CONCLUSIONS

The aim of this research was to find out (1) if fuzzy matching techniques are as effective as the dictionary-based translation techniques in CLIR between closely related languages like Norwegian and Swedish, and (2) the most suitable fuzzy string

matching technique for query translation in CLIR between closely related languages. The effectiveness of five fuzzy string matching techniques was tested for Norwegian to Swedish query translation with CLEF search topics from the year 2003. The fuzzy techniques were compared to three baseline techniques, which

were a dictionary translation baseline, a monolingual Swedish baseline and a monolingual Norwegian baseline.

Our main findings were:

- The fuzzy (n-gram) matching techniques are effective and applicable translation techniques in CLIR between closely related languages. For the best fuzzy matching query types performance difference with respect to dictionary translation queries was not statistically significant.
- The results do not suggest one best fuzzy matching technique for CLIR between closely related languages.
- The TRT technique alone is not a good approach (however, see below for the generation of transformation rules).

The results were encouraging giving support to our hypothesis that dictionary-based translation could be replaced by fuzzy string matching techniques in CLIR between closely related languages. The n-gram based techniques performed well, skipgrams being slightly better than conventional n-grams. This is in line with earlier research, where skipgrams have been found to be better than n-grams in matching cross-lingual spelling variants [5, 12]. Combining n-grams to the TRT techniques' statistical transformation rules improved results, the practical difference being of noticeable (5,1%) at the recall level 10. The TRT-n-grams also outperformed the best skipgrams at low recall levels. This suggests that the combined technique is useful in CLIR, as also found in earlier research [17]. The results also give reason to assume that combining the transformation rules to skipgram matching would be a good approach. This combination can be assumed to perform well, as the skipgrams have been shown to outperform the conventional n-grams in cross-lingual spelling variant matching [5, 12].

The results suggests that the transformation rules should be formed on a basis of a larger term pair list than was done in this study, or the list should be formed from technical terms instead of general vocabulary. The performance of the TRT queries might improve if the transformation rules were thereby improved. Better transformation rules might also further improve the performance of the combined TRT and n-gram queries.

In the present research, all the query words were lemmatized because the transformation rules in their current state can only handle base forms. Creating transformation rule collection capable of handling inflected word forms will be one of the next steps in our research. Our future research will also include testing the combination of TRT and skipgrams, and extending the research to concern Danish language.

7. REFERENCES

- [1] Adafre, S., van Hage, W., Kamps, J., de Melo, G. & de Rijke, M. 2004. The University of Amsterdam at CLEF 2004. CLEF 2004 Working Notes. Available at: <http://clef.iei.pi.cnr.it/>
- [2] Barðdal, J., Jörgensen, N., Larsen, G., & Martinussen B. 1997. Nordiska: Våra språk förr och nu. Lund, Studentlitteratur.
- [3] Broglio, J., Callan, J. & Croft B. 1993. Inquiry system overview. In Proceedings of the TIPSTER text program, 47-67. Available: <http://acl.ldc.upenn.edu/X/X93/X93-1008.pdf>
- [4] Hedlund, T., Pirkola, A. & Järvelin, K. 2001. Aspects of Swedish morphology and semantics from the perspective of mono- and cross-language information retrieval. *Information Processing & Management*, 37, 147-161.
- [5] Hedlund, T., Airio, E., Keskustalo, H., Lehtokangas, R., Pirkola, A. & Järvelin, K. 2004. Dictionary-based Cross-Language Information Retrieval: Learning Experiences from CLEF 2000-2002. *Information Retrieval – Special Issue on CLEF Cross-Language IR*, 7, 99-119.
- [6] Keskustalo, H. & Pirkola, A. & Visala, K. & Leppänen, Erkka & Järvelin, K. 2003. Non-adjacent Digrams Improve Matching of Cross-Lingual Spelling Variants. In: Nascimento, M.A., de Moura, E.S., Oliveira, A.L., (Eds.). *Proceedings of the 10th International Symposium, SPIRE 2003*. Manaus, Brazil, October 2003. Berlin: Springer, *Lecture Notes in Computer Science* 2857, pp. 252 - 265. ISSN 0302-9743, ISBN 3-540-20177-7.
- [7] Kraaij, W. 2004. Variations on language modeling for information retrieval. PhD thesis, University of Twente.
- [8] McNamee, P. & Mayfield, J. 2003. JHU/APL Experiments in Tokenization and Non-Words Translation. CLEF 2003 Working Notes. Available at: <http://clef.iei.pi.cnr.it/>
- [9] Navarro, G. 2001. A Guided tour to approximate string matching. *ACM Computing surveys (CSUR)* (33)1.
- [10] Peters, C. 2003. Introduction to the CLEF 2003 Working Notes. Available at: <http://clef.iei.pi.cnr.it/>
- [11] Pfeiffer, U., Poersch, T. & Fuhr, N. 1996. Retrieval effectiveness of proper name search methods. *Information Processing & Management*, 32(6), 667-679.
- [12] 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, August 24-28. New York: ACM, 55-63.
- [13] Pirkola, A., Keskustalo H., Leppänen, E., Käsälä, A.P. & Järvelin, K. 2002. Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants. *Information research*, 7(2) [<http://InformationR.net/ir/7-2/paper126.html>]
- [14] Pirkola, A., Puolamäki, D. & Järvelin, K. 2003. Applying query structuring in Cross-Language Retrieval. *Information Processing & Management* 39(3), 391-402.
- [15] Robertson, A.M. & Willet, P. 1998. Applications of n-grams in textual information systems. *Journal of Documentation*, 54(1), 48-69.
- [16] Spark Jones, K. 1974. Automatic indexing. *Journal of Documentation* (30) 4, 393-432.
- [17] Sperer, R. & Oard, D. W. 2000. Structured Translation for Cross-Language Information Retrieval. In Belkin, N. & Ingwersen, P. & Leong, M-K. (Eds.), *Proceedings of the 23rd Annual International SIGIR Conference on Research and Development in Information Retrieval*, 120-127. Athens, Greece.
- [18] Toivonen, J., Pirkola, A., Keskustalo, H., Visala, K., & Järvelin, K. 2005. Translating cross-lingual spelling variants

using transformation rules. Information Processing & Management, 41, 859-872.

[19] Voorhees, E. M. 2002. Overview of TREC 2002, Appendix 1. Common Evaluation Measures. The Proceedings of the

eleventh Text REtrieval Conference. Gaithersburg, Maryland. National Institute of Standards and Technology. [<http://trec.nist.gov/pubs.htm>]

Appendix 1. Examples for query types

Swebase

#sum(christo packeterar tyska riksdagshus konstnär christo inslagning tyska riksdagshus)

Nobase

#sum(christo pakke tysk riksdagsbygning innpakking tysk riksdag berlin kunstner christo)

Dicbase

#sum(christo #syn(paket packe bunt ask packa) #syn(tysk tyska) #syn(regerings stats stat statlig) dag #syn(byggnadsverk byggnad konstruktion hus) #syn(packning) #syn(tysk tyska) #syn(regerings stats stat statlig) dag berlin konstnär christo)

N-digram query

#sum(#syn(mchistori chefshistorik @christo @christos) #syn(paket pakets @paker @pak) #syn(tysk tysktysk @tyskl @tysklan) #syn(riksdagsbyggnad riksdagsbevakning @riksdagsoch @landsbyggsriksdagen) #syn(skinnpaj inpassning @pakkinen @iakkinen) tysk #syn(tysk tysktysk @tyskl @tysklan) #syn(riksdag riksdagsdag @riksdagsoch @riksdagsrupp) #syn(berliner berlinsk @berlin @berlins) #syn(kungstiger kungakonst @kunst @kunstler) #syn(mchistori chefshistorik @christo @christos))

Skip1 query (CCI = {{0},{1}})

#sum(#syn(chefjurist charterturistort @christo @christos) #syn(packe paket @takke @pakue) #syn(tysk tysktysk @tyskl @otysk) #syn(riksdagsbevakning riksdagsordning @riksdagsoch @riksdagsrupp) #syn(inpackning inpassning @ing @king) #syn(tysk tysktysk @tyskl @otysk) #syn(riksdag riksdagsdag @riksdagsoch @riksdagsrupp) #syn(berglin merlin @berlin @berlins) #syn(konstnär konstnummer @kunstler @köstner) #syn(chefjurist charterturistort @christo @christos))

Skip2 query (CCI = {{0},{1,2}})

#sum(#syn(tyristor mchistori @christo @christos) #syn(paket packe @pakue @takke) #syn(tysk tysktysk @tyskl @otysk) #syn(riksdagsbyggnad riksdagsbevakning @riksdagsebatten @riksdagsrupp) #syn(inpackning inpassning @king @parking) #syn(tysk tysktysk @tyskl @otysk) #syn(riksdag riksdagsdag @riksdagsoch @riksdagsrupp) #syn(berglin merlin @berlin @berlins) #syn(konstnär konstcenter @kunstler @kunstlers) #syn(tyristor mchistori @christo @christos))

TRT query

#sum(#syn(christo) #syn(packa pakka packe pakke) #syn(tysk) #syn(riksdagsbygning) #syn(innpacking innpakking) #syn(tysk) #syn(riksdag) #syn(berlin) #syn(kunstner) #syn(christo))

Combined TRT and n-digram

#sum(#syn(mchistori chefshistorik @christo @christos) #syn(packa packad @packard @packalén) #syn(tysk tysktysk @tyskl @tysklan) #syn(riksdagsbyggnad riksdagsbevakning @riksdagsoch @landsbyggsriksdagen) #syn(inpackning inpacka @inpac @racking) #syn(tysk tysktysk @tyskl @tysklan) #syn(riksdag riksdagsdag @riksdagsoch @riksdagsrupp) #syn(berliner berlinsk @berlin @berlins) #syn(kungstiger kungakonst @kunst @kunstler) #syn(mchistori chefshistorik @christo @christos))