

Data Driven Methods for Improving Mono- and Cross-lingual IR Performance in Noisy Environments

Antti Järvelin
Department of Computer
Sciences
University of Tampere
FIN-33014 University of
Tampere, Finland
antti.jarvelin@cs.uta.fi

Tuomas Talvensaari
Department of Information
Studies
University of Tampere
FIN-33014 University of
Tampere, Finland
tuomas.talvensaari@uta.fi

Anni Järvelin
Department of Information
Studies
University of Tampere
FIN-33014 University of
Tampere, Finland
anni.jarvelin@uta.fi

ABSTRACT

In cross-language information retrieval (CLIR), novel or non-standard expressions, technical terminology, or rare proper nouns can be seen as noise when they appear in queries or in the target collection. This kind of vocabulary is often out-of-vocabulary (OOV) for dictionaries that are used to translate queries. In historic document retrieval (HDR), OCR errors and historical spelling variants cause similar problems. In this paper, three data driven approaches to these problems are presented. The two first methods, the transformation rule based translation (TRT) method and the classified *s*-gram method, operate on string level. With them approximate matches of a query word can be recognized from the target document collection and included into the target query. In the third method, the corpus-based approach, parallel or comparable corpora are employed to derive translation knowledge that can be used to translate OOV words. Besides the overview of the methods, three case studies highlighting their practical applications in CLIR are also presented. The methods are shown to be effective in query translation without dictionaries between closely related languages (TRT and *s*-grams), OOV word translation (*s*-grams), and boosting dictionary-based CLIR performance by way of OOV word translation (corpus based methods).

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation

Keywords

Cross-language information retrieval, noise, OOV words, TRT, *s*-grams, corpus based methods

1. INTRODUCTION

Noisy data introduces problems to many information retrieval applications. Out-of-vocabulary (OOV) words in cross-language information retrieval (CLIR) and OCR errors and historical spelling variants in historic document retrieval are examples of this. OOV words introduce noise in query translation as they cannot be translated with the common dictionary-based query translation tools. Many typical OOV words, such as proper names and technical terms, are usually important query keys [19], which makes their translation an essential question in CLIR. Such OOV words are often cross-lingual spelling variants in different languages, i.e., they share a common root but are rendered with different spelling of the underlying sounds. The cross-lingual spelling variants are similar (e.g. a Swedish word *heksaklorid* and its English variant *hexachloride*), which provides a good basis for the use of, e.g., approximate string matching to translate OOV words. Other techniques suggested for the OOV word translation include the TRT technique (discussed later), corpus-based translation of OOV words, where translation equivalents are automatically mined from the web [5] and transliteration that has been used for phonetic translation of OOV words, e.g., between Arabic and English [1] and Japanese and English [7].

CLIR between closely related languages can be seen as a special case of OOV word translation: When a large part of two languages vocabularies are orthographically related words – cross-lingual spelling variants – all the query words can be handled as OOV words with relatively good query translation quality as a result [12]. Therefore CLIR between closely related languages can be seen as querying from noisy texts. This is also the case with historic document retrieval (HDR), where the spelling variation introduced by the language development and OCR errors in the scanned collections cause similar problems.

From the information retrieval system's point of view, both spelling variation (be it cross-lingual or historical spelling variation), and OCR errors in the digitalized collection, can be seen as problems of searching relevant documents from noisy target collection. The queries can also contain noise in the form of novel or non-standard expressions, typos, misspellings, etc. In this paper three data driven approaches for dealing with noisy target collections and queries are presented, and their performance in various CLIR-related tasks

are evaluated.

The first approach called *Transformation Rule based Translation* (TRT), is a fuzzy translation technique for cross-lingual spelling variants based on statistical rules that model typical character changes between cross-lingual spelling variants. The technique is used in two steps: First the statistical rules are used to create intermediate forms of the source words. Then the intermediate forms are matched with their target language equivalents through approximate string matching. Toivonen et al. [27] found that the two-step TRT technique performed clearly better than n -gram matching in translating technical terms and proper names. In Section 3.1 experiments where TRT and classified s -grams were used in CLIR between closely related languages, are presented. Norwegian queries were translated into Swedish with the TRT method, and a performance comparable to dictionary-based translation was achieved.

The second approach, the *classified s -gram matching*, is an approximate string matching technique, which generalizes the well known n -gram matching technique. Pirkola et al. [20] and Keskustalo et al. [14] showed that cross-lingual spelling variation can advantageously be modeled with the s -grams which thus can be effectively used for searching translation candidates for OOV words. Especially, they showed that in this task the classified s -gram matching outperformed other established approximate string matching techniques, such as edit distance, longest common subsequence, and n -grams. Section 3.2 presents extensive experiments where typical OOV words (proper nouns, technical terms) were translated with the classified s -gram method between 11 language pairs. The results indicate that s -grams outperform n -grams in OOV translation especially between remotely related languages.

The third approach presented in this paper is to mine parallel or comparable texts from the web, and align them so that passages in the source language are mapped to similar passages (i.e., translations or, in the case of comparable texts, topically related texts) in the target language. The alignments in turn can be used to derive translation knowledge which can be used in CLIR query translation [15, 26]. This kind of “real-life” training data can boost CLIR performance in situations when queries or the target collection are marred by noise such as novel or non-standard terminology or abbreviations. This kind of vocabulary is usually OOV for dictionaries. However, even in the vast WWW, it is hard to find clean parallel content for a given language pair and domain, and, consequently, it is often necessary to resort to noisier comparable texts. Thus, besides noisy queries or target collections, noise can also be induced into CLIR via noisy training data. In the experiments presented in Section 3.3, German queries belonging to a specific domain (genomics) were translated into English with various CLIR set-ups. A system based on combination of a dictionary and comparable corpora outperformed other approaches, indicating that noisier comparable corpora work well in OOV translation, especially in special domains.

The rest of this paper is organized as follows. Section 2 gives an introduction to the three methods presented in this paper. Section 3 discusses three case studies where the meth-

ods are applied to CLIR related problems. Section 3.1 gives an example how TRT, and s -gram methods can be used in CLIR to avoid dictionary translation of queries between closely related languages. In Section 3.2 the performance of classified s -grams in OOV word matching is illustrated. Corpus based methods for noisy translation are investigated in Section 3.3. Finally, Section 4 provides discussion and conclusions.

2. DATA DRIVEN METHODS FOR IMPROVING QUERY PERFORMANCE

2.1 TRT

Transformation rule based translation (TRT) is a fuzzy translation technique for translation of OOV words in CLIR [27]. It is based on the use of statistically generated rules that model typical character changes and correspondences between cross-lingual spelling variants within a language pair. The rules are created from a large set of cross-lingual spelling variants (words that have the same origin and only differ due to the orthographical differences between the languages), that are typically extracted from dictionaries. The TRT technique can be used both as a complement to dictionary-based query translation or as the sole translation technique for translation between closely related languages. The TRT technique has two steps: the transformation rules are combined to n -gram matching so that first all possible rules are used to create a set of reasonable (intermediate) translation candidates. Then n -grams are used to match these against the target document collection’s index to discard bad intermediate translations that are not real words. The idea of the TRT and the generation of the transformation rules are described in more detail in [27].

A *transformation rule* contains source and target language characters and their context characters. 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* shows how reliable a rule is, in practice it is the frequency of a rule divided by the number of source words where the source substring of the rule occurs. Frequency and confidence factor are 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 \Rightarrow för [beginning, 132, 147, 89.80]

The rule means that the letter o, between r and f, is transformed into the letter ö in the beginning of words, with the confidence factor being 89.80. The confidence factor is the frequency of the rule (132) divided by the number of source words where the string occurs (147).

The performance of the TRT technique can be enhanced by combining it to the FITE (frequency-based identification of translation equivalents) technique. The FITE-TRT [21] is a statistical technique for the identification of correct transformation equivalents of source words obtained by TRT. The core idea of FITE is that except for the correct 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 com-

mon in the language that the other word forms. Therefore document frequencies of words can be used to find the best translation alternatives given by TRT.

2.2 s -grams

The second approach, the *classified s -gram matching* [20], is an approximate string matching technique, which generalizes the well known n -gram matching technique. n -grams have earlier been applied successfully e.g. in proper name matching [18] and historic word variant matching [16]. In the classified s -gram matching the strings to be compared are split into substrings of length n (n -grams) and then the proximity of the strings is defined as the overlap of the strings' n -grams using some proximity measure. The technique differs from the n -gram matching in two important aspects. First of all, skipping characters is allowed when forming the substrings. Secondly, the substrings are produced using various skip lengths and are classified into sets called *gram classes* based on the number of characters skipped. Two or more gram classes may also be combined into more general gram classes. The *character combination index (CCI)* then indicates the set of all the gram classes to be formed from a string. For example, CCI $\{\{0\}, \{1, 2\}\}$ means that two gram classes are formed from a string: one formed of adjacent characters ($\{0\}$) and one formed both by skipping one and two characters ($\{1, 2\}$). Only the substrings belonging to the same gram class of the CCI are compared to each other, thus the name classified s -gram matching. Table 1 provides an example of forming s -grams of different CCIs for the string "abracadabra".

s -gram-based string proximity measures are based on strings' *s -gram profiles* which contain the information how many times each s -gram occurs in a given string (see [11] for details).

Definition 1. Let $w = a_1 a_2 \dots a_m$ be a string over a finite alphabet Σ , $n \in \mathbb{N}^+$ be a gram length, $k \in \mathbb{N}$ a skip length and let $x \in \Sigma^n$ be an s -gram. If $a_i a_{i+k+1} \dots a_{i+(k+1)(n-1)} = x$ for some i , then w has a $s_{n,k}$ -gram occurrence of x . Let $G_k(w)[x]$ denote the total number of $s_{n,k}$ -gram occurrences of x in w . The $s_{n,k}$ -gram profile of w is the vector $G_{n,k}(w) = (G_k(w)[x]), x \in \Sigma^n$.

s -gram profiles can easily be generalized for gram classes. The s -gram profiles for the gram classes are formed by summing up the s -gram profiles in a given gram class.

Definition 2. Let $w \in \Sigma^*$, $C \in \mathcal{P}(\mathbb{N})$ a gram class and $x \in \Sigma^n$. Let $G_C(w)[x] = \sum_{k \in C} G_k(w)[x]$. The gram class profile of w is the vector $G_{n,C}(w) = (G_C(w)[x]), x \in \Sigma^n$. In other words, $G_{n,C}(w) = \sum_{k \in C} G_{n,k}(w)$.

Sometimes the exact number of the occurrences of s -grams in the string is irrelevant, but merely the information if a specific s -gram occurs at all in the string is needed. In fact, Järvelin and Järvelin [10] show that in classified s -gram-based OOV word matching, proximity measures based on the binary gram class profile perform better or equally well as the proximity measures based on the general gram class

profile. The binary gram class profile is defined by binarizing the gram class profile of Definition 2.

Definition 3. Let $w \in \Sigma^*$, and $C \in \mathcal{P}(\mathbb{N})$ a gram class and $x \in \Sigma^n$. Let

$$B_C(w)[x] = \begin{cases} 1 & \text{if } G_C(w)[x] > 0 \\ 0 & \text{otherwise} \end{cases}.$$

The binary gram class profile of w is the binary vector $B_{n,C}(w) = (B_C(w)[x]), x \in \Sigma^n$.

Various proximity measures can be used to calculate string proximities based on the general and binary gram class profiles. For example, the Dice's coefficient, which was the best performing proximity measure in the tests of Järvelin and Järvelin [10], is defined for strings v and w by

$$D_{n,C}(v, w) = \frac{2B_C(v)^T B_C(w)}{\|B_C(v)\|^2 + B_C(v)^T B_C(w) + \|B_C(w)\|^2}, \quad (1)$$

where T denotes the transpose of a vector.

The classified s -gram matching is based on character combination indices, which contain at least one gram class. The CCI based string proximity measures are defined as the average proximity of strings' gram class proximities. That is, for a given CCI \mathcal{C} , gram length n , and a gram class proximity measure \mathcal{P} , the corresponding CCI based proximity measure $\mathcal{P}_{n,\mathcal{C}}(v, w)$ between the strings v and w is given by

$$\mathcal{P}_{n,\mathcal{C}}(v, w) = \frac{1}{|\mathcal{C}|} \sum_{C \in \mathcal{C}} \mathcal{P}_{n,C}(v, w). \quad (2)$$

Thus, for example, the CCI based Dice's coefficient is obtained by substituting $\mathcal{P}_{n,C}(v, w)$ for $D_{n,C}(v, w)$ of Eq. (1) in the Eq. (2).

2.3 Corpus Based Methods

Parallel or comparable corpora are often-used resources in CLIR query translation. A parallel corpus is a collection where texts in one language are aligned with their translations in another language. Comparable corpora, on the other hand, consist of texts that are not translations, but share similar topics. They can be, e.g., newspaper collections written in the same time period in different countries.

Parallel corpora are preferred to comparable corpora as translation resources because more dependable knowledge can be derived from them. However, a translation corpus has to fulfill two conditions before it can be of use: firstly, naturally, the source and target languages of the aligned collection have to match the source and target languages of the CLIR queries at hand. Secondly, the domain, or the topic of speech, of the translation corpus has to match the domain of the queries. For example, translating sports-related queries with a parallel corpus consisting of agricultural legislation would probably produce bad results.

The problem is that high quality parallel corpora do not exist for all language pairs and domains. Hence, it is sometimes necessary to resort to noisier comparable corpora. In

Table 1: The gram classes for forming the s -grams of length two with different CCIs for the string “abracadabra”.

CCI	Gram classes
$\{\{0\}\}$	$\{ab, br, ra, ac, ca, ad, da, ab, br, ra\}$
$\{\{1\}\}$	$\{ar, ba, rc, aa, cd, aa, db, ar, ba\}$
$\{\{2\}\}$	$\{aa, bc, ra, ad, ca, ab, dr, aa\}$
$\{\{1, 2\}\}$	$\{ar, ba, rc, aa, cd, aa, db, ar, ba, aa, bc, ra, ad, ca, ab, dr, aa\}$
$\{\{0\}, \{1, 2\}\}$	$\{ab, br, ra, ac, ca, ad, da, ab, br, ra\},$ $\{ar, ba, rc, aa, cd, aa, db, ar, ba, aa, bc, ra, ad, ca, ab, dr, aa\}$

CLIR query translation, comparable corpora can be utilized as a complementary resource to provide translations to words that are OOV for the more general resources. The use of noisier resources is acceptable in CLIR, because query translation is easier than machine translation (MT). Queries can be translated word-for-word, whereas in MT, syntactical knowledge is required in addition to lexical coverage.

There are many ways to utilize an aligned (i.e., parallel or comparable) corpus. For example, the alignments can be used in *cross-language query expansion* [3], or they can be used as training data for statistical translation models, which in turn can be employed in translating queries [15].

Another approach is to use the corpus as a *cross-language similarity thesaurus*, meaning a structure in which target language words are ranked based on their calculated similarity with a source language query word that is given as input. The more often two words co-occur in the aligned documents, the more similar they are. The highest ranking words are assumed to be either translations of the input word or related to it in some other manner. This approach was pioneered by Sherdan and Ballerini [22].

In similarity thesaurus calculation, the vector space model of information retrieval [2] can be used in an inverted way. In the classic vector model, documents are modelled as vectors whose features correspond to the words in the documents. The similarity of two documents (or the similarity of a query to a document, since queries can be seen as short documents) can then be calculated, e.g., with the cosine of the angle between the document vectors. In similarity thesaurus calculation, the source language word to be translated is thought of as the query, and target language words are retrieved as the answer. The aligned documents are thought of as the defining features of words, rather than the other way around, as in document retrieval. The distribution of a word across the documents determines its location in the semantic space defined by the documents. As in document retrieval, measures such as the cosine similarity can be used in defining the similarity between two words.

To use a comparable corpus as a similarity thesaurus, the documents of the two languages have to be aligned so that documents in the source language are mapped to documents in the target language that cover similar topics. Talvensaar et al. [26] proposed a method where source language documents are used as queries that are translated into the target language with an initial dictionary. The translated queries are then run against the target documents with a document

ranking algorithm, and a few highest ranking documents are aligned with the source document. Similarity score thresholds are used to filter out bad alignments. Consequently, only a subset of the original corpus is part of the aligned translation corpus.

3. CASE STUDIES

3.1 CLIR Between Closely Related Languages without Dictionary Translation

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 method. Closely related languages typically share a high number of spelling variants. If the number of the shared cross-lingual variants is high enough, query translation can be handled by cheaper and simpler fuzzy translation techniques. Norwegian and Swedish are closely related Scandinavian languages: Around 90 % of the vocabularies of the languages are similar having only some orthographical and inflectional differences [4]. The TRT technique and the s -gram matching technique were tested in query translation from Norwegian to Swedish. The goal was to reach translation quality that would enable CLIR effectiveness comparable to dictionary-based translation.

A typical CLIR test setting with a subset of 50 search topics and the test collection from Cross-Language Evaluation Forum (CLEF) 2003 (Cross-Language and More -track) was used. Norwegian search topics were not included in the CLEF test environment. Therefore English test topics were translated into Norwegian by a native speaker. The document collection and the search topics were lemmatized prior to the dictionary translation and the monolingual Swedish query. No morphological preprocessing was done prior to the fuzzy translation. The stop words were removed.

Test queries were formed from the title and description fields of the test topics. The s -gram translation was done by translating the Norwegian topics into Swedish by matching the s -grams of the topic words against the Swedish document collection’s index words’ s -grams. The four best matches were selected as translations for each word. The transformation rules were created by translating a part of the Swedish document collection’s index to Norwegian with the GlobalDix dictionary by Kielikone. The final Norwegian to Swedish word-pair list had 3058 unique word-pairs of non-identical words, with a maximum edit distance value of half of the length of the longer word and with minimum length of four characters. A confidence factor of 50 % and a low frequency threshold of 2 were used. The TRT was done

Table 2: AP values over recall levels 0-1 and the differences between the techniques.

Technique	Dibase	<i>s</i> -grams	TRT
Average precision	0.36	0.30	0.29
Difference to Dibase	-	-16.7 %	-18.1 %

as follows: First all the possible (fitting) TRT rules were applied to each source word to create intermediate translations. The intermediate translation with highest frequency and confidence factor was then selected and matched against the target document collection with *n*-grams. The four highest ranked keys from the result list of *n*-gram matching were selected for the final queries.

The GlobalDix dictionary by Kielikone was used for the dictionary translation. All the translations for each topic word were selected to the query. To tackle the ambiguity problem common in dictionary-based CLIR the queries were structured according to the Pirkola method [19] that is known to be an effective way to disambiguate CLIR queries [19, 23]. Also a “minimal effort” monolingual Swedish baseline query was run. The topic words were lemmatized but no compound splitting was done. This monolingual baseline performed slightly worse than the dictionary baseline and thus (and due to the lack of space) only the results for the dictionary baseline are reported here. The performance of the translation techniques was measured as interpolated average precision (AP) over the eleven standard recall levels averaged over all queries. The statistical significance of the results was tested using Friedman test [6] on inversed ranks.

The results are presented in Table 2 and in Fig. 1. Both the TRT and the *s*-grams achieved on average over 80 % of the dictionary baseline. The differences in the techniques’ average precisions were not statistically significant. This is a result that shows that fuzzy translation is a promising and interesting approach to query translation between closely related languages. The results are the same when comparing the precision at the higher ranks, i.e., at the 0-0.1 recall level (Fig. 1). The TRT and the *s*-grams achieved over 80 % of the dictionary baseline’s performance. The differences between the translation techniques were not statistically significant here either. The documents placed at the top of the result list are the most important ones from the practical user perspective. The *s*-grams and the TRT performed quite equally. TRT performs better at low recall levels (0-0.2) but has slightly lower AP than the *s*-grams. However, the transformation rules were generated from a relatively small word-pair list, with general vocabulary from newspaper texts. This list seemed to be insufficient for generating enough high frequency transformation rules. The lack of high quality rules probably affected negatively the TRT technique’s translation results.

3.2 Out-of-Vocabulary Word Translation with *s*-grams

To demonstrate the value of the classified *s*-gram matching technique in the OOV word translation, a set of 271 typical OOV words was collected and translated between 11 language pairs using classified *s*-gram matching. These

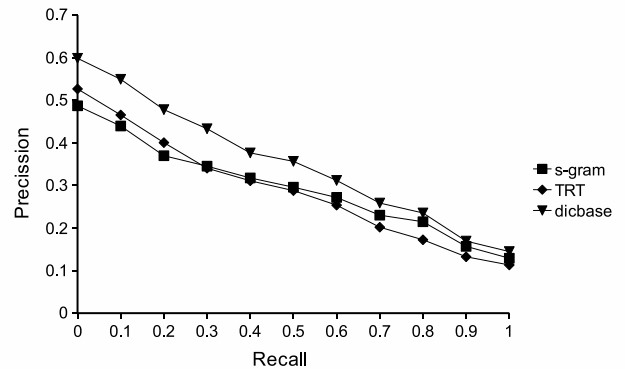


Figure 1: Precision-recall graph over the eleven standard recall levels.

OOV words were mostly technical terms from the domains of biology, medicine, economics and technology, but also a list of geographical names obtained from [14] was included. The search keys were expressed in seven languages (English, Finnish, French, German, Italian, Spanish, and Swedish) and were translated into four target languages (English, German, Finnish, and Swedish). English was combined to all of the other languages as a target language and was also used as a source language with Finnish, German and Swedish. Translation was also done both ways between Swedish and German. Examples from the English search key collection include *adrenalin*, *Pyongyang*, and *zygote*.

Target word lists (TWLs), from where the translations for the OOV words were searched from, consisted of CLEF 2003 [17] document collections’ indices for the target languages. The size of the collections, and thus the TWLs, varies between languages. The English TWL consisted of ca 257,000, the Swedish TWL of ca 388,000, and the Finnish TWL of ca 535,000 unique word forms. The German CLEF03 collection was considerably larger and thus only a part of it was used for creating a TWL including ca 391,000 unique word forms. All the TWLs were lemmatized with the TWOL morphological analyzer by Lingsoft Ltd. The words not recognized by the morphological analyzer were indexed as they appeared in the text. Compounds were split and both the original compounds and their constituents were indexed. The missing translation equivalents of the search keys were added to the TWLs, and there was exactly one correct translation for each search key in the TWLs.

The gram length was set to two in this experiment, as it has been found suitable with the classified *s*-grams [14, 20]. The Dice’s coefficient was used as the proximity measure between the strings, because it seems to be the best performing proximity measure with the classified *s*-grams [10]. Totally six CCIs were tested (see Table 3). The tested CCIs contain also *n*-grams which correspond to CCI₀.

For each search key 100 best translations were produced, with exception of ties at the last place when all translations within the cohort of equal proximity values were included into the result set. Translations ranked lower were

Table 3: The six tested CCIs. CCI₀ corresponds to the n -grams.

CCI ₀	{{0}}	CCI ₃	{{0}, {0, 1}}
CCI ₁	{{0, 1}}	CCI ₄	{{0}, {1}, {1, 2}}
CCI ₂	{{0, 1, 2}}	CCI ₅	{{0}, {1, 2}}

not taken into consideration. This is well motivated as taking more than 2-4 translation candidates into a query tends to deteriorate the query performance [8]. To compare the CCIs, the mean average precision (MAP) was calculated for each language pair and CCI at three different levels: among top 2, top 5 and top 100 highest ranked translation candidates. The top 2 and top 5 levels were the most interesting ones, as more translation candidates would deteriorate the query performance. If the correct translation was in a cohort of words sharing the same proximity value with the target word, the average rank of the cohort was used. The statistical significance of the differences between the n -grams and different CCIs inside each language pair was tested with Friedman test.

The results are presented in Table 4 when the top 5 translation candidates are considered. The results for top 2 and top 100 translation candidates gave the same order for the CCIs the overall MAPs being slightly lower in top 2, and slightly higher in top 100. The differences between the CCIs were biggest with language pairs that were linguistically remotely related. Thus the classified s -gram matching technique performed better than the n -grams in noisy environments, i.e., when the languages were not closely related. When the environment was less noisy (the languages are more closely related), the differences between the classified s -gram matching and n -gram matching diminished. CCIs 4 and 5 are examples of how the classification of skip indices into suitable gram classes benefits the technique as the differences between these CCIs and n -grams were almost always statistically highly significant.

3.3 CLIR Based on Noisy Comparable Corpora

In [26], Talvensaari et al. created a German-English comparable corpus in the genomics domain. The translation corpus was created by first retrieving genomics-related text in German and English from the web. This was done by means of language-aware focused web crawling. The texts were then aligned with the procedure described in Section 2.3. A total of 39,143 German paragraphs extracted from the web pages were aligned with 39,190 unique English paragraphs (some target documents were part of more than one alignments). The alignments were made in 1-to- n manner, meaning that one source document was aligned with one or more target documents.

The resulting aligned comparable corpus was used as a cross-language similarity thesaurus (see Section 2.3) in order to translate German queries into English. The queries were created from the test topics of the TREC genomics track [9], and the MEDLINE collection of 4.6M English medical documents were used as the target test collection. Prior to the experiments, the English topics were first translated

into German by a native German speaker, who also had a background in molecular biology. Of the 50 topics, 20 were used to train the parameters of the similarity thesaurus, and the remaining 30 were used in the actual experiments.

Some of the TREC topics contain key vocabulary that cannot, or should not, be translated (e.g., “Find articles about function of FancD2.”), which, it could be argued, make the topics too “easy” for CLIR experiments. However, most topics also contain lots of important query words that should be translated to achieve acceptable CLIR performance (e.g., “Find protocols for generating transgenic mice.”). It should be noted that such variation in the difficulty of CLIR queries also occurs in more general domains. In the news domain, for example, queries often have proper names that need not be translated.

The CLIR performance of the similarity thesaurus system was compared to, and combined with, two other resources: the Utaclir dictionary-based query translation program [13], and the JRC-Acquis parallel corpus consisting of legislative documents of the European Union [24]. Naturally, the German-English alignments of the JRC corpus were used in the experiments. A total of five different CLIR approaches were applied (see [25] for more detailed account of the experiments):

GenWeb The German-English genomics web corpus used as a similarity thesaurus.

JRC The JRC-Acquis parallel corpus used as a similarity thesaurus.

UC The Utaclir dictionary-based query translator.

GenWeb+UC The combined output of GenWeb and UC used as the target query.

JRC+UC The combined output of JRC and UC used as the target query.

Table 5 depicts the performance of the mentioned CLIR approaches in mean average precision (MAP), precision after 10 retrieved documents (P@10), and the OOV rate of the approaches (i.e., the percentage of query words that could not be translated). The JRC approach performs very poorly by all measures, which is caused by the fact that its domain (legislation) does not match that of the queries (genomics). GenWeb performs better than JRC, even though it is a noisy comparable corpus with only topically matching alignments. The best approach, by a clear margin, is the combination of dictionary-based translation and genomics comparable corpus. In this approach, the dictionary covers the general vocabulary, and the comparable corpus provides translations for the domain-specific vocabulary. The results show that, especially in special domains, noisy translation resources can be useful.

4. DISCUSSION

Many “real-life” document collections contain noise which is problematic for the traditional IR systems. The noise might be introduced to the target document collection due the spelling errors in the original collection or from other sources

Table 4: The MAPs for among top 5 translation candidates for all CCIs and language pairs. The best performing CCI for each language pair is on bold. Note that CCI₀ corresponds to *n*-grams. The performance gain achieved by using of CCI₁-CCI₅ compared to *n*-grams of CCI₀ is marked into parenthesis. Statistically highly significant differences ($p < 0.001$) between CCI₀ (*n*-grams) and the other CCIs are marked with ** and statistically significant differences ($p < 0.01$) with *.

Language	CCI ₀	CCI ₁	CCI ₂	CCI ₃	CCI ₄	CCI ₅
EN-FI	0.37	0.43 (+0.07)**	0.42 (+0.05)**	0.43 (+0.06)	0.45 (+0.08)**	0.45 (+0.08)**
FI-EN	0.40	0.44 (+0.04)	0.42 (+0.02)	0.45 (+0.05)**	0.45 (+0.05)**	0.46 (+0.06)**
IT-EN	0.45	0.50 (+0.05)**	0.48 (+0.03)**	0.50 (+0.04)**	0.53 (+0.07)**	0.52 (+0.06)**
SP-EN	0.53	0.54 (+0.01)	0.53 (+0.00)	0.55 (+0.02)	0.56 (+0.03)**	0.57 (+0.04)**
EN-SW	0.55	0.56 (+0.01)	0.55 (+0.00)	0.56 (+0.01)	0.57 (+0.02)**	0.56 (+0.02)*
SW-EN	0.54	0.56 (+0.02)	0.55 (+0.01)	0.57 (+0.03)**	0.59 (+0.05)**	0.58 (+0.04)**
GE-EN	0.57	0.60 (+0.03)*	0.60 (+0.03)*	0.61 (+0.04)**	0.63 (+0.06)**	0.63 (+0.06)**
EN-GE	0.59	0.61 (+0.03)	0.59 (+0.00)	0.60 (+0.01)	0.61 (+0.03)**	0.63 (+0.05)**
FR-EN	0.68	0.71 (+0.03)**	0.69 (+0.01)	0.71 (+0.03)**	0.72 (+0.04)**	0.71 (+0.03)**
GE-SW	0.74	0.75 (+0.00)	0.73 (-0.01)	0.75 (+0.01)	0.77 (+0.02)**	0.76 (+0.02)*
SW-GE	0.75	0.75 (+0.00)	0.73 (-0.02)*	0.76 (+0.01)	0.76 (+0.02)	0.77 (+0.02)*
MEDIAN	0.55	0.56 (+0.01)	0.55 (+0.00)	0.57 (+0.02)	0.59 (+0.04)	0.58 (+0.04)

Table 5: Results for the German-English genomics CLIR experiments

Approach	MAP	P@10 docs	OOV %	Resource ¹
JRC	0.087	0.210	35.0	PC
GenWeb	0.137	0.297	29.1	CC
UC	0.170	0.270	56.6	D
JRC+UC	0.136	0.303	29.3	PC+D
GenWeb+UC	0.225 ²	0.407	17.3	CC+D

¹ PC = parallel corpus, CC= comparable corpus, D = dictionary

² Significantly better than JRC and JRC-UC according to Friedman test ($p < 0.05$)

such as OCR errors in scanned collections. In CLIR, besides these “normal” sources of noise, also OOV words can be seen as a source of noise, because typical examples of OOV words include technical terminology and proper names, which, in turn, are often cross-lingual spelling variants between languages. Likewise, queries can contain noise in the form of novel or non-standard expressions, typos, misspellings etc.

In this paper, three different data driven methods, TRT, classified *s*-grams, and corpus-based methods, were presented to overcome the problematic noise in queries and the target document collection. The presented methods are language independent and economical in production use.

Three CLIR case studies utilizing the introduced methods were also reported. First the performance of the TRT and the classified *s*-gram technique was compared to the dictionary translation between closely related languages (Norwegian to Swedish). The results showed that both the TRT and the classified *s*-grams achieved on average over 80 % of the dictionary baseline’s performance. The differences in the techniques’ average precisions were not statistically significant. This proves the techniques a viable alternative in CLIR between closely related languages.

To be able to create transformation rules, a training set con-

taining translation equivalents between the language pairs, from which the TRT-rules can be mined, is needed. Without a training set, transformation rules cannot be generated. In this situation classified *s*-gram matching can be used instead, as it does not require extra information about the language pair at hand and is thus more language independent method than the TRT method.

The second case study illustrated how the classified *s*-gram technique can be utilized in OOV word translation in the CLIR. The performance of six *s*-gram types were compared using eleven language pairs. Generally the classified *s*-grams performed better than the *n*-grams (CCI₀), the differences being bigger the less the languages were related. Therefore, classified *s*-grams seem to suit better than *n*-grams into noisy environments, e.g., for OOV word translation between remotely or non related languages.

An application of corpus-based methods in OOV word translation was presented in the third case study. The experiments showed that noisy comparable corpora can provide translations for technical vocabulary that is OOV for dictionaries and parallel corpora that cover more general vocabulary. This is beneficial, because comparable corpora are far easier to obtain than parallel corpora. Moreover, as opposed to the string-level methods, corpus-based methods produce not only translations, but also topically related query expansion keys that can enhance query recall. However, the comparison of the performance of the corpus-based approach to that of other similar approaches is difficult, because the TREC genomics track topics have not been used in CLIR experiments by other research groups to our knowledge.

The case studies presented in this paper concentrated on the CLIR applications of the presented techniques. However, the techniques are data independent, and can thus be applied also to other domains. Especially, the TRT technique and the classified *s*-gram technique could be utilized in the mono-lingual IR to overcome the noise introduced by spelling or OCR errors. The methods could also be applied to historic document retrieval where the OCR errors and

spelling variation caused by the language evolution introduce similar problems.

5. ACKNOWLEDGMENTS

The authors wish to thank Academy Professor Kalervo Järvelin from the University of Tampere for his constructive comments on the manuscript. The second author was funded by the Academy of Finland (project name “Focused Web Crawling”), and the third author by the Tampere Graduate School of Information Science and Engineering (TISE). The TWOL was provided by the Lingsoft Ltd. The GlobalDix was provided by Kielikone Ltd.

6. REFERENCES

- [1] N. AbdulJaleel and L. S. Larkey. Statistical transliteration for English-Arabic cross language information retrieval. In *Proceedings of the 12th International Conference on Information and Knowledge Management*, pages 139–146, 2003.
- [2] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.
- [3] L. Ballesteros and W. B. Croft. Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st ACM SIGIR Conference*, pages 64–71, 1998.
- [4] J. Barðdal, N. Jörgensen, G. Larsen, and B. Martinussen. *Nordiska: Våra språk förr och nu*. Studentlitteratur, 1997.
- [5] P.-J. Cheng, J.-W. Teng, R.-C. Chen, J.-H. Wang, W.-H. Lu, and L.-F. Chien. Translating unknown queries with web corpora for cross-language information retrieval. In *Proceedings of the 27th ACM SIGIR Conference*, pages 146–153, 2004.
- [6] W. J. Conover. *Practical Nonparametric Statistics*. Wiley, New York, NY, USA, 3rd edition, 1999.
- [7] A. Fujii and T. Ishikawa. Japanese/English cross-language information retrieval: Exploration of query translation and transliteration. *Computers and the Humanities*, 35(4):389–420, 2001.
- [8] T. Hedlund, E. Airio, H. Keskustalo, R. Lehtokangas, A. Pirkola, and K. Järvelin. Dictionary-based cross-language information retrieval: Learning experiences from CLEF 2000 – 2002. *Information Retrieval - Special Issue on CLEF Cross-Language IR*, 7(1–2):99–119, 2004.
- [9] W. R. Hersh. Report on the TREC 2004 genomics track. *SIGIR Forum*, 39(1):21–24, 2005.
- [10] A. Järvelin and A. Järvelin. Comparison of s-gram proximity measures in out-of-vocabulary word translation. To appear in the *Proceedings of the 15th International Symposium on String Processing and Information Retrieval (SPIRE 2008)*, 2008.
- [11] A. Järvelin, A. Järvelin, and K. Järvelin. s-grams: defining generalized n-grams for information retrieval. *Inf. Process. Manage.*, 43(4):1005–1019, 2007.
- [12] A. Järvelin, S. Kumpulainen, A. Pirkola, and E. Sormunen. Dictionary-independent translation in CLIR between closely related languages. In *Proceedings of the 6th Dutch-Belgian Information Retrieval Workshop (DIR 2006)*, pages 25–32, 2006.
- [13] H. Keskustalo, T. Hedlund, and E. Airio. Utaclir : general query translation framework for several language pairs. In *Proceedings of the 25th ACM SIGIR Conference*, pages 448–448, 2002.
- [14] H. Keskustalo, A. Pirkola, K. Visala, E. Leppänen, and K. Järvelin. Non-adjacent digrams improve matching of cross-lingual spelling variants. In *Proceedings of the 10th International Symposium on String Processing and Information Retrieval (SPIRE)*, volume 2857 of *LNCS*, pages 252–265, Berlin, 2003. Springer.
- [15] W. Kraaij, J.-Y. Nie, and M. Simard. Embedding web-based statistical translation models in cross-language information retrieval. *Comput. Linguist.*, 29(3):381–419, 2003.
- [16] A. O’Rourke, A. Robertson, and P. Willett. Word variant identification in old french. *Information Research*, 2(4), 1997.
- [17] C. Peters. Introduction to the CLEF 2003 working notes, 2003. Available at: <http://clef.iei.pi.cnr.it/>.
- [18] U. Pfeiffer, T. Poersch, and N. Fuhr. Retrieval effectiveness of proper name search methods. *Inf. Process. Manage.*, 32(6):667–679, 1996.
- [19] A. Pirkola. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st ACM SIGIR Conference*, pages 55–63, 1998.
- [20] A. Pirkola, H. Keskustalo, E. Leppänen, A.-P. Käsälä, and K. Järvelin. Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants. *Information Research*, 7(2), 2002. Available at <http://InformationR.net/ir/7-2/paper126.html>.
- [21] A. Pirkola, J. Toivonen, H. Keskustalo, and K. Järvelin. FITE-TRT: A high quality translation technique for OOV words. In *Proceedings of the 2006 ACM Symposium on Applied Computing*, pages 1043–1049, 2006.
- [22] P. Sheridan and J. P. Ballerini. Experiments in multilingual information retrieval using the SPIDER system. In *Proceedings of the 19th ACM SIGIR Conference*, pages 58–65, 1996.
- [23] R. Sperer and D. W. Oard. Structured translation for cross-language information retrieval. In *Proceedings of the 23rd ACM SIGIR Conference*, pages 120–127, 2000.
- [24] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, and D. Varga. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 2006.
- [25] T. Talvensaari. Effects of aligned corpus quality and size in corpus-based CLIR. In *The Proceedings of the 30th European Conference on IR Research (ECIR 2008)*, volume 4956 of *LNCS*, pages 114–125. Springer, 2008.
- [26] T. Talvensaari, A. Pirkola, K. Järvelin, M. Juhola, and J. Laurikkala. Focused web crawling in acquisition of comparable corpora. *Information Retrieval*, 11, 2008.
- [27] J. Toivonen, A. Pirkola, H. Keskustalo, K. Visala, and K. Järvelin. Translating cross-lingual spelling variants using transformation rules. *Inf. Process. Manage.*, 41:859–872, 2005.