Information Research, Vol. 7 No. 2, January 2002,

Targeted s-gram matching: a novel n-gram matching technique for cross- and mono-lingual word form variants

Ari Pirkola, Heikki Keskustalo, Erkka Leppänen,
Antti-Pekka Känsälä and Kalervo Järvelin
Department of Information Studies
University of Tampere
Finland

Abstract

We present a novel n-gram based string matching technique, which we call the targeted s-gram matching technique. In the technique, n-grams are classified into categories on the basis of character contiguity in words. The categories are then utilized in matching. The technique was compared with the conventional n-gram technique using adjacent characters as n-grams. Several types of words and word pairs were studied. English, German, and Swedish query keys were matched against their Finnish spelling variants and Finnish morphological variants using a target word list of 119 000 Finnish words. In all cross-lingual tests done, the targeted s-gram matching technique outperformed the conventional n-gram matching technique. The technique was highly effective also for monolingual word form variants. The effects of query key length and the length of the longest common subsequence (LCS) of the variants on the performance of s-grams were analyzed.

Introduction

Word form variation, which involves cross-lingual spelling variation and monolingual morphological variation, is an important and challenging issue in mono- and cross-language information retrieval. Different forms of the same word represent the same concept, thus being equal from the standpoint of users' requests. However, in traditional retrieval systems based on exact string matching, a query key contributes to retrieval success only if it is identical to the corresponding index term. *Approximate string matching* techniques, however, are capable of finding word form variants. In this paper, we will explore one such technique - n-gram based string matching. We will explore cross-lingual *spelling variants*, i.e., equivalent words in different languages which differ slightly in spelling, as well as monolingual *morphological variants*.

This study was motivated by our concern of untranslatable proper names and technical terms in *dictionary-based cross-language retrieval* (CLIR). For an overview of the approaches to cross-language retrieval, see (Oard and Diekema, 1998). For an overview of the methods used in *dictionary-based CLIR*, see (Pirkola *et al.*, 2001). Our experience on CLIR system development and evaluation at the University of Tampere, Information Retrieval Laboratory (Hedlund *et al.*, 2001; Pirkola, 1998), and the analysis of request key properties (Pirkola and Järvelin, 2001) have shown that proper names and technical terms often are prime keys in requests, and if not translated by dictionaries, query performance may deteriorate. Proper names (technical terms) in different languages often have the same origin, being thus spelling variants of each other. General translation dictionaries may include some proper names, such as the names of countries and capital cities, as well as common technical terms, but generally these kinds of words are untranslatable. In dictionary-based cross-language retrieval untranslatable query keys are typically used in target language queries in their original source language forms. Unless they are identical to the corresponding database index terms, they do not contribute to retrieving relevant documents. However, the fact that proper names (technical terms) often are form variants of each other allows the use of approximate string matching

techniques to find the target language correspondents of source language keys.

A common method to handle morphological variation in retrieval systems is to conflate different morphological forms of the same word - or related words in the case of derived words - into the same form using stemming algorithms based on suffix lists and desuffixing rules (Pirkola, 2001). In morphologically complex languages, retrieval effectiveness can be improved, and users can be freed from taking into account word morphology, by using dictionary-based morphological analyzers, which normalize inflected word forms into their base forms. As translation dictionaries, the dictionaries of morphological analyzers are incomplete, just in part covering the (theoretical) lexicons of languages. Therefore a given query key may be retained, say, in a genitive form in the morphological normalizing of a query, while the corresponding index term may be in several different forms. This problem concerns both mono- and cross-lingual retrieval systems in which morphological analyzers are employed, and could be addressed using approximate string matching techniques. In addition to spelling and morphological variation, there are other situations in which the use of approximate string matching techniques would be useful, particularly spelling errors (Zobel and Dart, 1995) and difficulty in knowing exact spellings of names.

Approximate matching techniques involve *Soundex* and *Phonix*, which compare words on the basis of their phonetic similarity (Gadd, 1988; Gadd, 1990). In the techniques, phonetic codes are computed for the strings that are compared, and the strings with similar codes are counted similar. *Damerau-Levenstein metric* (Damerau, 1964) was developed specifically for spelling errors. In *n-gram based matching* (Angell *et al.*, 1983; Hall and Dowling, 1980; Pfeifer *et al.*, 1996; Robertson and Willett, 1998; Salton, 1989; Zobel and Dart, 1995), text strings are decomposed into n-grams, i.e., substrings of length *n*, which usually consist of the adjacent characters of the text strings. *Digrams* contain two and *trigrams* three characters. The degree of similarity between the strings is computed on the basis of the number of similar n-grams and the total number of unique n-grams in the strings.

In this paper, we will investigate n-gram based word matching. The aim is to develop an effective n-gram matching technique particularly for cross-lingual spelling variants, as well as monolingual morphological variants. N-gram matching is a language independent matching technique. It thus seems to be an ideal approximate matching technique for CLIR systems processing different languages. Moreover, n-gram matching has been reported to be an effective technique among various approximate matching techniques (Pfeifer *et al.*, 1996; Zobel and Dart, 1995).

We will investigate the effectiveness of *digrams* combined both of adjacent and *non-adjacent characters* of words. These kinds of n-grams we call *s-grams* (where *s* refers to the term *skip*). We assumed that for cross-lingual spelling variants s-grams may be more effective than n-grams combined of adjacent characters only, because variant forms may share just a few (1-2) digrams formed of adjacent characters, or no digrams at all if the words are very short. If also non-adjacent characters are used, however, even short words may have many similar digrams.

Cross-lingual spelling variation involves substitution and addition/deletion of letters in words. For instance, in the Finnish word *kalsitoniini* (*calcitonin*) there are many transformations typical of Finnish spelling variants: *c* → *k* and *c*→*s* substitutions, the lengthening of a vowel (one type of addition) and addition of a single vowel. The second *i*-vowel in *calcitonin* is transformed into a double vowel (*ii*) in *kalsitoniini*. The *i*-vowel at the end of *kalsitoniini* represents the case of single vowel addition.

We will test several types of character combinations regarding the number of skipped characters, as well as a novel technique to compare the s-grams of query keys with those of target words. We call the technique *the targeted s-gram matching technique*, or in more specific contexts, *the classified s-gram matching technique*. In the technique, s-grams are classified into categories on the basis of the number of skipped characters, and only the s-grams belonging to the same category are compared with one another. We will demonstrate that the technique is effective for many types of word form variants. In all cross-lingual tests done in this study, it outperformed the commonly used n-gram matching technique where n-grams are composed of adjacent characters. Also in the case of monolingual word form variants the technique was very effective.

The rest of this paper is organized as follows. The next section introduces to the research problems investigated in this study and presents the problems. This is followed by sections on methods and data, findings, discussion and conclusions.

Research problems

Generally, the effectiveness of n-gram matching depends on the number of similar n-grams and the total number of n-grams generated from the words that are compared, as well as the number of words in *a target word list* (TWL), i.e., a list of words against which a query key is matched, and the frequencies of different n-grams generated from the target words. For example, for a spelling variant pair which shares many high frequency n-grams, matching can be expected to be less effective than for a spelling variant pair only sharing low frequency n-grams.

The relative effectiveness of different kinds of n-gram matching techniques can be evaluated empirically by selecting query keys, and defining for each key all the relevant word forms in the TWL (i.e., the recall bases of query keys). The n-grams of a key are then matched against the n-grams of TWL words. The effectiveness of an n-gram matching technique can be calculated using the measure of *precision*, i.e., the proportion of relevant words among all the words retrieved. Our test data consisted of English-Finnish, German-Finnish, and Swedish-Finnish cross-lingual spelling variants, and Finnish morphological variants (Section 3). The TWL contained 119 000 words. By using this test data we investigated empirically several research problems described below.

When s-grams are formed from words, it is possible to operate on different character combinations with respect to the number of skipped characters (0, 1, 2, ..., m-2 skipped characters), where *m* refers to the number of characters in a word w_i. For *digrams*, we use *a character combination index (CCI*) to indicate the number of skipped characters as s-digrams are formed. In the notation, each number refers to the number of characters between the constituent characters of s-digrams. For example, CCI=(1, 2) refers to s-digrams composed of characters separated by one character and two characters in the words. Conventional digrams composed of adjacent characters are marked as CCI=(0). CCI = (0, 1, ..., (m-2)) refers to a character combination operation where all possible digrams are formed. The s-digrams belonging to the same category in the case of classified s-grams are marked using the parentheses '[' and '] '.

Table 1 shows the s-digrams with CCI=(0, 1, 2) for the spelling variant pair *pharmacology* and *farmakologian* (the Finnish correspondent for *pharmacology* in a genitive form).

Word	CCI	S-digram set
pharmacology	(0)	{ph,ha,ar,rm,ma,ac,co,ol,lo,og,gy}
	(1)	{pa,hr,am,ra,mc,ao,cl,oo,lg,oy}
	(2)	{pr,hm,aa,rc,mo,al,co,og,ly}
farmakologian	(0)	{fa,ar,rm,ma,ak,ko,ol,lo,og,gi,ia,an}
	(1)	{fr,am,ra,mk,ao,kl,oo,lg,oi,ga,in}
	(2)	{fm,aa,rk,mo,al,ko,og,li,oa,gn}

Table 1: Examples of s-digrams with CCI=(0,1,2).

We assumed that the effectiveness of s-gram matching depends on CCI. The use of extensive skipping, e.g., all possible digrams are formed, often gives many high frequency digrams, which occur in many words. This may depress the effectiveness. On the other hand, restricted skipping provides less similar digrams for related words than an extensive skipping. This also may depress the effectiveness. Therefore determining a balanced CCI seems crucial for s-gram matching to be effective.

The use of non-adjacent characters as n-grams is not a new idea (Robertson and Willett, 1998; Ullmann, 1977). However, the classification of s-grams on the basis of character contiguity, and matching based on the s-gram categories is a novel idea. Also the perspectives adopted in this study on the issue, and the research problems we will explore are novel. Ullmann (1977) studied spelling errors in terms of efficiency (speed) and used n-grams combined of non-adjacent characters, such as quadgrams 1356 and 1245, where the numbers stand for letter positions in the words containing six letters (the test dictionary only contained six-letter words). The method was applied for finding from a dictionary all the words that differed from a given input word by 1-2 letters. Ullmann concluded that parallel processing of sets of n-grams probably is faster than scanning through the dictionary in the case of large dictionaries.

The research problems investigated in this paper are as follows:

1. **The effects of CCI on the effectiveness of s-gram matching**: Several character combination types were tested, i.e., CCIs were varied. The effects of different combination types were evaluated using English - Finnish spelling variants.
2. **The effectiveness of the targeted s-gram matching**: The targeted s-gram matching technique was compared with the conventional n-gram technique using adjacent characters as n-grams (digrams and trigrams). Also unclassified s-grams were tested. Regarding the use of a blank space as a consituent character of s-grams, three types of tests were done: (1) no space, (2) the start space, and (3) both start and end spaces allowed as consituents.
3. **The effectiveness of s-gram matching for various types of words and word pairs**: English, German, and Swedish query keys against the Finnish words in the TWL were matched. Query key lists contained medical and pharmacological terms (the first key list) and geographical names (the second list). Finnish query keys in base and genitive forms were matched against the Finnish words in the TWL. In the TWL, Finnish words were in many different forms owing to word inflection.
4. **Factors affecting the effectiveness of the targeted s-gram matching technique**: To explain the results we analyzed the effectiveness of classified s-gram matching with respect to that of conventional n-gram matching in terms of query key length and the number of characters in the longest common subsequence (LCS) of the variants that were compared. For two words, their LCS is the longest character sequence of the sequences that occur in both words. For example, for the words **r e** **t** **r i** **e** **v a l** and **r e** **v i v a l** the LCS is **r e i v a l**.

Methods and data

The target word list

In the Information Retrieval Laboratory at the University of Tampere there are many full text research databases. Laboratory's Finnish database contains 55 000 documents (Finnish newspaper articles). A set of 35 test topics has been created on the basis of the articles. Around 17 000 human relevance assessments have been made to judge the relevance of the documents against the test topics. The database has been used in many IR studies (e.g., Järvelin and Kekäläinen, 2000; Sormunen, 2000). The words were normalized using the Twol morphological analyzer of Lingsoft Corp. Those words that the Twol did not recognize were indexed in a separate index. In this study we used that separate index as a target word list.

Thus, we used as test words such words that actually are problematic in IR. This method of isolating the difficult cases in a separate file is reasonable from the n-gram matching perspective, otherwise the effectiveness of n-gram matching would be lower due to the higher number of TWL words. There is no need to apply n-gram matching for the words which can be handled using a morphological analyzer.

The TWL contains some 119 000 words. It includes Finnish proper names (e.g., personal names, company names, and geographical names) and Finnish common nouns, Finnish words borrowed from other languages, i.e., spelling variants, English and other foreign language words, and Finnish spelling error forms. Finnish is a morphologically complex language, and many of the words in the list occur in several inflected forms. The most common word forms in Finnish are the base (nominative) and genitive forms (Karlsson, 1983).

Gathering query keys

Altogether 8 query key lists were used in the experiments of this study. For different lists we used different query key gathering methods, as described in this section.

Eng1, Ger, and Swe lists

The target word list was browsed from the start to end. A list of pharmacological and medical terms was gathered. Each term was looked up in a medical dictionary to find its English equivalent. If the dictionary translated the term into English, the English word was selected as a query key. Sometimes the dictionary gave more than one translation equivalent. In these cases, the orthographically closest equivalent was chosen for the test (this also holds for the Eng2 list below). The selected English keys were translated into German and Swedish by means of medical dictionaries. In some cases translations were searched in the Web.

Eng2 list

A list of English place names and their Finnish equivalents was collected from a place name dictionary, which contains world's place names in both languages. Each Finnish name was searched in the target word list. If the Finnish name was found in the list, the English name was selected as a query key.

Fin lists

A set of Finnish words in different morphological forms beginning with the letters *a* and *k* was gathered from the TWL. In both cases this was done systematically by selecting from the list the first 50 original Finnish words. A native Finnish speaker can readily recognize the original Finnish words. The *k*-words words represent long words while the average length of *a*-words is much lower (Table 2). Thus, the use of these two samples allows studying the effects of target word length on the effectiveness of n-gram matching. The base and genitive forms of these words were used as query keys, while the words that were gathered from the TWL formed the recall bases of the keys.

Query key lists and recall bases

The query key lists are described in Table 2.

For each query key, the corresponding Finnish word in different forms in the TWL formed the recall base of the key (i.e., the set of relevant words (word forms)). The last column in Table 2 shows the average number of relevant words in the TWL for different lists. As can be seen in Table 2, on the average one key in the Eng2 list had 3.9 relevant target words. For the other lists, the number of relevant words varied between 1.6-2.0.

Compound words containing relevant words as their components were judged as relevant target words. *Adjective derivatives* of noun keys were judged relevant. Both compound and derivative correspondents were more common in the Eng2 list than in the other lists.

Below is an example of a result list for the query key *calcitonin* (the top ranked words ordered by decreasing SIM value) The correct correspondent (*kalsitoniini*) is at the sixth position in the list.

1. 0.472222 calcitonin halcionin
2. 0.459459 calcitonin billitonin
3. 0.388889 calcitonin kalitinin
4. 0.388889 calcitonin calvinon
5. 0.384615 calcitonin halcioniin
6. *0.380952 calcitonin kalsitoniini*
7. 0.371429 calcitonin lintonin
8. 0.365854 calcitonin calutronin
9. 0.361111 calcitonin calvinin

   ...

An example of a compound word containing a relevant word is the compound *sambesijoki* (*zambezi river*) for the key *zambezi* (*sambesi*). An example of an adjective derivative is the word *katatoninen* (*catatonic*), which is derived from the noun *katatonia* (*catatonia*). (Both words, *katatoninen* and *katatonia*, occurred in the TWL.)

List Name Number of Query Keys	Avg. Word Length	Query Key Types	Avg. No TWL Words
ENG1, N=52	9,4	English medical and pharmacological spelling variants	2,0
GER, N=52	9,3	German medical and pharmacological spelling variants	

SWE, N=52	8,8	Swedish medical and pharmacological spelling variants	
ENG2, N=41	7,4	English geographical spelling variants	3,9
FIN-base/a-words, N=50	7,6	Finnish words beginning with the letter 'a'in a base form	2,0
FIN-gen/a-words, N=50	8,5	Finnish words beginning with the letter 'a' in a genitive form	
FIN-base/k-words, N=50	11,9	Finnish words beginning with the letter'k' in a base form	1,6
FIN-gen/k-words, N=50	12,8	Finnish words beginning with the letter'k' in a genitive form	

Table 2: Query key and TWL word statistics.

S-digram types

Classified and unclassified s-digrams, and the following types of character combinations were tested in the study:

- Unclassified s-digrams: CCI=(0, 1), CCI=(0, 1, 2), CCI = (0, 1, ..., (m-2))
- Classified s-digrams: CCI=([0], [1]), CCI=([0], [1, 2]), CCI=([0], [1], ..., [9])

In the case of unclassified digrams no restrictions were set, but each digram of a query key was compared with each digram of TWL's words. In the case of classified digrams, the digrams of a query key and those generated from TWL's words belonging to the same category were compared to each other. For example, in the case of CCI=([0], [1]) the digrams with CCI=(0) of a key were compared with the digrams with CCI=(0) of TWL's words, and the digrams with CCI=(1) of a key were compared with the digrams with CCI=(1) of TWL's words. In the case of CCI=([0], [1, 2]), digrams with CCI= (1) and digrams with CCI=(2) were put into the same category. Digrams belonging to this category of 1-2 skipped characters were compared to each other but not to digrams with CCI=(0) (and vice versa).

Table 3 presents examples of unclassified and classified s-grams with different CCIs. Note that in the case of CCI=([0], [1, 2]) the words *abcde* and *axxc* - perhaps surprisingly - have a similar digram (*ac*).

Word	CCI	Digram set(s)
abcde	(0)	{ab,bc,cd,de}
abcde	(0, 1)	{ab,ac,bc,bd,cd,ce,de}
abcde	(0, 1, 2)	{ab,ac,ad,bc,bd,be,cd,ce,de}
abcde	([0], [1])	{ab,bc,cd,de} and {ac,bd,ce}
abcde	([0], [1, 2])	{ab,bc,cd,de} and {ac,ad,bd,be,ce}
axxc	([0], [1, 2])	{ax,xx,xc} and {ax,ac,xc}

Table 3: Examples of unclassified and classified s-grams.

Computing similarity values

Similarity values were computed using the following string similarity scheme (Pfeifer *et al.*, 1996):

$SIM(N_1, N_2) = | N_1 \cap N_2| / | N_1 \cup N_2|,$

where N_1 and N_2 are digram sets of two words. $| N_1 Ç N_2|$ denotes the number of intersecting (similar) digrams, and $| N_1 È N_2|$ the number of unique digrams in the union of N_1 and N_2. For example, the degree of similarity for the words *rwanda*

and *ruanda* is calculated as follows (for n-grams with CCI = (0)):

SIM({rw,wa,an,nd,da},{ru,ua,an,nd,da}) = |{an,nd,da}| / |{rw,wa,an,nd,da,ru,ua}| = 3/7 (0.428).

For the word form compared to itself the similarity value is 1.0.

Findings

The performance of s-grams

The results were evaluated as average precision at 100% recall. In other words, we computed the proportion of relevant words to all words at the last relevant word in the result list. We did not use any other evaluation measure (such as precision at different recall levels), because the average number of relevant TWL words was low (Table 2).

The results were analyzed manually. The result lists were cut at the SIM-value of 0.2. This means that for each query key the result list contained several hundreds words. In some cases the last relevant word did not occur in the list, in which case the default precision value of 0% was used.

Statistical significance of the difference between the performance of s-grams and that of baseline n-grams was tested using *Wilcoxon signed ranks test.* The test uses both the direction and the relative magnitude of the difference of comparable samples. The statistical program that was used is based on Conover (1980). The statistical significance levels of 0.01, and 0.001 are indicated in the tables.

In all cases n-grams with CCI=(0), i.e. the conventional n-grams combined of adjacent characters were used as baseline. We used both digram and trigram baselines. Digrams were run for all 8 lists. Trigrams were run for the following lists: Eng1, Eng2, Finnish a-words/base forms and Finnish a-words/genitive forms. For cross-lingual spelling variants, the digram baseline always performed better than the trigram baseline (Tables 4-5). For Finnish morphological variants, the trigram baseline sometimes performed better than the digram baseline (Tables 8a and 8b). The effectiveness of test digrams was compared with that of the better baseline (digrams/trigrams).

All the s-gram types were first tested on the Eng1 list. In the first experiment the effects of CCI were tested. The findings of the Eng1 tests are presented in Table 4. As can be seen, classified s-grams with CCI=([0], [1, 2]) perform markedly better (avg. precision 64.1%) than the baseline n-grams with CCI=(0) (avg. precision 55.2%). Also unclassified s-grams with CCI=(0, 1) and CCI=(0, 1, 2) perform well in relation to the baseline. The unrestricted s-gram technique, in which all the possible digrams are formed is the worst method, giving much lower precision (38.2%) than the baseline (55.2%).

The best s-gram techniques of the Eng1 tests, i.e., classified s-grams with CCI=([0], [1, 2]) and unclassified s-grams with CCI=(0, 1) were tested using the other lists as test data. The results are presented in Tables 5-8. As shown, in all tests (i.e., in all lists and experiments regarding the use of a blank space as a digram character) classified s-grams with CCI=([0], [1, 2]) perform better than the baseline n-grams with CCI=(0). Unclassified s-grams with CCI=(0, 1) perform better or as well as the baseline n-grams.

As shown in Tables 4-7, in the Eng1, Eng2, Swe, and Ger tests the highest performance improvements are achieved in the case of start and end spaces. For classified s-grams with CCI=([0], [1, 2]), the relative improvement percentages with respect to baseline are 18.2% (Eng1, Table 4), 49.7% (Eng2, Table 5), 20.7% (Ger, Table 6), and 17.1% (Swe, Table 7). The results are statistically significant at the levels of 0.01-0.001.

In the Fin tests performance improvements are smaller (Tables 8a-8d). In one case (Finnish a-words/genitive forms, with start + end spaces) the baseline n-grams perform better than the classified s-grams (Table 8b). In Finnish, the application of the classified s-gram technique seems to be useful particularly for words possessing the inflectional pattern of *wordstem inflection*. The term refers to words whose word stems are changed in inflection, e.g., *Asonen* (personal name in a base form) and *Asosen*(the genitive form of the name *Asonen*). For example, in Fin a-word/base

form tests, the application of the classified s-gram technique gave performance improvements for 10 matching cases (on the average precision was improved from 79,3% to 82,7% for the 52 matching cases; Table 8a). Six of the ten (i.e., 60%) positive cases concerned the matching of inflectional stem words. For all the 52 matching cases in the a-word/base form test the frequency of such inflectional stem words that contributed to precision was much lower, that is, 19,2% (10/52).

In the Eng1, Eng2, Swe, and Ger tests, the use of the start space yields lower relative improvements than the other two cases. However, in three of four tests, with the exception Eng2, classified s-grams with the start space is the best matching technique. In the Eng2 test classified s-grams with start and end spaces perform slightly better.

For Finnish a-words/genitive forms (Table 8b) the case of classified s-grams with the start space yields substantial improvements with respect to the case of classified s-grams with no space. In the former case precision is 90.6% and in the latter case 74.2%. This is a remarkable in the sense that the comparison precision of 74.2% is high.

Analyzing the factors affecting the performance of classified s-grams

To explain the superior performance of classified s-grams, we analyzed the results as follows.

(1) The effects of key length on the performance of classified digrams with CCI=([0], [1, 2]) and digrams with CCI=(0) was evaluated using the Eng1 and Eng2 lists. As can be seen in Table 9, the same trends hold for Eng1 and Eng2 lists: the shorter the word, the higher the relative performance of s-grams. In the word group of >= 9 letters performance improvement is small. In fact, for Eng2 words, the precision of baseline n-grams is slightly better than that of s-grams. For the medium length words performance improvements are substantial for both lists. The low performance figures in the word group of =< 6 letters suggest that in the case of short words it is often impossible to find the correspondents whatever n-gram method is used (see Discussion section). In a few cases, however, the use of s-grams yields substantial performance improvements. For instance, in the Eng2 list precision for the name *Ithaca* is improved from 2,2% to 25,0% owing to applying the classified s-gram technique.

(2) For query keys in the Eng1, Eng2, Ger, and Swe lists and the corresponding relevant words at the last positions in the results lists (at which point precision was computed) the number of characters in the LCS of a query key and the corresponding TWL word was calculated (the analysis was done if the last relevant word occurred in the result lists of both matching techniques tested; see Section 4.1). In other words, the length of LCS was determined. Each query key/TWL word pair was put into the category of short (=< 8 characters) or long (> 8 characters) LCS. For each category, the performance of classified digrams with CCI=([0], [1, 2]) and digrams with CCI=(0) was computed.In the case of short LCSs the performance of both classified s-grams and baseline n-grams is much worse than in the case of long LCSs (Table 10). This holds for all the four lists. However, for baseline n-grams the performance drop is more striking. This is shown in the last column of Table 10, which presents the performance of baseline n-grams with respect to classified s-grams; for all lists, the relative performance of baseline n-grams is markedly worse for short than long LCSs.

Discussion

In CLIR, proper names often are untranslatable due to limited coverages of translation dictionaries. Similarly, some words cannot be normalized, because the dictonaries of morphological analyzers are incomplete. In such cases, approximate matching techniques can be applied in searching for cross-lingual spelling variants and morphological variants. N-gram matching is a language independent means to recognize word from variants. It has been reported to be an effective technique among different approximate matching techniques in indexed systems, such as text retrieval systems (Pfeifer *et al*., 1996; Zobel and Dart, 1995). Pfeifer et al. (1996) studied name searching and tested the following approximate matching techniques: Soundex, Phonix, Damerau-Levenstein metric, Skeleton-key, and Omission-key. The most effective single technique was n-gram matching. Digrams were more effective than trigrams. Digrams with a space as their constituent character performed better than digrams in which only alphabetic characters were used. N-gram matching also could be utilized in resolving spelling errors which may be common in some databases (Zobel and Dart, 1995), in searching for historical word variants (O'Rourke *et al*., 1997), as well as an alternative method for stemming algorithms (Kosinov, 2001; Xu and Croft, 1998).

In this study, the effectiveness of various types of s-digram matching techniques with respect to that of the conventional n-gram matching technique where only adjacent characters are used as n-grams was tested empirically.

In summary, our main findings are as follows:

1. The effects of different character combinations types (i.e., CCIs were varied) were evaluated using English - Finnish spelling variants. We found that s-grams perform well if a relatively low CCI is chosen for matching.
2. We discovered an effective n-gram matching technique which we call *the classified (targeted) s-gram matching technique*. In all the cross-lingual experiments we did, the technique outperformed the conventional n-gram matching technique.
3. Several types of words and word pairs were studied. The types were English - Finnish medical (pharmacological) and geographical spelling variants, German - Finnish and Swedish - Finnish medical spelling variants, and Finnish morphological variants. Both unclassified s-grams, and particularly the classified s-grams, were effective for all these word types.
4. The effectiveness of s-gram matching with respect to that of the conventional n-gram matching depends on query key length and the number of characters in the longest common subsequence (LCS) of the variants. The s-gram technique is more effective than the n-gram technique particularly for short words and short LCSs.
5. The use of the end space (together with the start space) gave the worst matching performance both for conventional n-grams and classified s-grams. This reflects the complex suffix-based inflectional system of Finnish; many of the target words were in inflected forms. The use of the end space in n-gram matching is not suited for inflectionally complex suffix languages.

In all tests of this study the target language was Finnish. Whether the finding of the effectiveness of classified s-gram matching can be generalized for other (target) languages is a research problem of future research. However, it is likely that the technique is also suited for other languages, because spelling variation is the same type of phenomenon in most languages (deletion, addition, and substitution of letters in words). Nevertheless, the degree of spelling variation depends on the language pair. The analysis of the factors affecting the performance of classified s-grams showed that the effectiveness of the technique depends on query key and LCS lengths. Word length is a language and domain dependent property. LCS length is dependent on the specific language pair considered. If the degree of spelling variation is small for two languages and variant forms often have long LCSs, the effectiveness of conventional n-gram matching may be good, while in the case of more extensive variation the classified s-gram matching technique may be much more effective.

Finnish is a highly complex suffix language (Pirkola, 2001). It has been estimated that theoretically a Finnish noun may have over 2000 inflectional forms. In practice, most words occur in several inflectional forms in databases. The fact that the use of the start space as a s-gram character yields the best performance reflects the morphological features of Finnish. However, it may also be true that especially the end parts of English (German, Swedish) - Finnish spelling variants are different. For some language pairs, spelling variation concerns particularly the initial parts of words. For example, Spanish words often begin with the letter *e*, while the corresponding English words do start with other letters. Therefore it does not seem reasonable to use the start space for spelling variant matching for language pairs having that kind of variation.

In this study, we classified s-grams on the basis of character contiguity. It may be possible to improve the technique by utilizing information on s-gram locations in words. The method could be further improved by taking into account s-gram frequencies. The capability of high frequency s-grams to discriminate between words is low. Therefore, the down-weighting of high frequency s-grams seems a method worth testing.

In some cases the extent of cross-lingual spelling variation is so high that no n-gram technique is able to fing right target language correspondents. For example, the Finnish correspondent for the name *Chechnya* is *Tsetshenia*. It seems that the only means to find right correspondents in cases like this is to use transliteration rules. For transliteration in CLIR (Japanese-English word transliteration), see (Fujii and Ishikawa, 2001). At the University of Tampere our objective is to develop language independent methods for CLIR. In agreement with this objective we are developing a method which automatically generates transliteration rules for different language pairs based on the information included in translation dictionaries. Transliteration may be used in combination with s-gram matching for better matching performance.

Conclusions

In this study we discovered an effective n-gram technique which we call the targeted s-gram matching technique. We demonstrated that the technique is effective for many types of word form variants when a proper character

combination operation is used. The results showed that with respect to conventional n-gram matching s-gram matching is effective particularly for short words and short LCSs.

This study was the first in our n-gram research project at the University of Tampere Information Retrieval Laboratory. In the project we are studying n-gram based translation of proper names and other spelling variants. In the next phrase, we will set up a new research environment (fully automated analysis methods and English language as a target language). Our future plans involve exploring positional and frequency statictics of s-grams to improve the effectiveness of s-gram matching, and developing a method that automatically generates transliteration rules for various language pairs.

ENG medical and pharmacological words	Average Precision	% Change	Stat. Sign. Level
No space			
Digram baseline, CCI=(0)	55,2	—	—
Trigram baseline, CCI=(0)	52,6	—	—
Unclassified, CCI=(0, 1, ., (m-2))	38,2	-30,8	0,001
Unclassified, CCI=(0, 1)	61,2	+10,9	0,01
Unclassified, CCI=(0, 1, 2)	60,5	+9,6	—
Classified, CCI=([0], [1], ..., [9])	43,8	-20,7	0,001
Classified, CCI= ([0], [1])	56,9	+3,1	—
Classified, CCI= ([0], [1, 2])	64,1	+16,1	0,001
Start + end spaces			
Digram baseline, CCI=(0)	54,3	—	—
Trigram baseline, CCI=(0)	53,6	—	—
Classified, CCI= ([0], [1, 2])	64,2	+18,2	0,001
Start space			
Digram baseline, CCI=(0)	62,7	—	—
Trigram baseline, CCI=(0)	55,6	—	—
Classified, CCI= ([0], [1, 2])	67,0	+6,9	—

Table 4: The performance of s-grams. ENG1 list.

ENG geographical names	Average Precision	% Change	Stat. Sign. Level
No space			
Digram baseline, CCI=(0)	18,4	—	—
Trigram baseline, CCI=(0)	15,5	—	—
Unclassified, CCI=(0, 1)	22,2	+20,7	0,01

Classified, CCI= ([0], [1, 2])	25,1	+36,4	0,001
Start + end spaces			
Digram baseline, CCI=(0)	19,9	—	—
Trigram baseline, CCI=(0)	19,5	—	—
Classified, CCI= ([0], [1, 2])	29,8	+49,7	0,01
Start space			
Digram baseline, CCI=(0)	21,6	—	—
Trigram baseline, CCI=(0)	20,4	—	—
Classified, CCI= ([0], [1, 2])	29,0	+34,3	0,01

Table 5: The performance of s-grams. ENG2 list.

GER medical and pharmacological words	Average Precision	% Change	Stat. Sign. Level
Nospace			
Baseline, CCI=(0)	62,0	—	—
Unclassified, CCI=(0, 1)	69,6	+12,3	0,01
Classified, CCI= ([0], [1, 2])	70,7	+14,0	0,01
Start + end spaces			
Baseline, CCI=(0)	56,9	—	—
Classified, CCI= ([0], [1, 2])	68,7	+20,7	0,001
Start space			
Baseline, CCI=(0)	69,2	—	—
Classified, CCI= ([0], [1, 2])	73,3	+4,1	—

Table 6: The performance of s-grams. GER list.

SWE medical and pharmacological words	Average Precision	% Change	Stat. Sign. Level
No space			
Baseline, CCI=(0)	68,9	—	—
Unclassified, CCI=(0, 1)	75,8	+10,0	—
Classified, CCI= ([0], [1, 2])	75,6	+9,7	0,01
Start + end spaces			
Baseline, CCI=(0)	63,0	—	—

Classified, CCI= ([0], [1, 2])	73,8	+17,1	0,01
Start space			
Baseline, CCI=(0)	74,5	—	—
Classified, CCI= ([0], [1, 2])	77,7	+4,3	—

Table 7: The performance of s-grams. SWE list.

FIN Words	Average Precision	% Change	Stat. Sign. Level
No space			
Digram baseline, CCI=(0)	79,3	—	—
Trigram baseline, CCI=(0)	78,9	—	—
Unclassified, CCI=(0, 1)	79,3	0,0	—
Classified, CCI= ([0], [1, 2])	82,7	+4,3	0,01
Start + end spaces			
Digram baseline, CCI=(0)	79,7	—	—
Trigram baseline, CCI=(0)	82,4	—	—
Classified, CCI= ([0], [1, 2])	84,8	+6,4	—
Start space			
Digram baseline, CCI=(0)	85,6	—	—
Trigram baseline, CCI=(0)	86,6	—	—
Classified, CCI= ([0], [1, 2])	88,8	+2,5	—

Table 8a: The performance of s-grams. FIN list, a-words/base forms.

FIN Words	Average Precision	% Change	Stat. Sign. Level
No space			
Digram baseline, CCI=(0)	68,1	—	—
Trigram baseline, CCI=(0)	65,9	—	—
Unclassified, CCI=(0, 1)	71,6	+5,1	—
Classified, CCI= ([0], [1, 2])	74,2	+9,0	0,01
Start + end spaces			
Digram baseline, CCI=(0)	65,6	—	—
Trigram baseline, CCI=(0)	77,4	—	—

Classified, CCI= ([0], [1, 2])	71,4	-7,8	—
Start space			
Digram baseline, CCI=(0)	89,9	—	—
Trigram baseline, CCI=(0)	85,7	—	—
Classified, CCI= ([0], [1, 2])	90,6	+0,8	—

Table 8b: The performance of s-grams. FIN list, a-words/genitive forms.

FIN Words	Average Precision	% Change	Stat. Sign. Level
No space			
Baseline, CCI=(0)	91,1	—	—
Unclassified, CCI=(0, 1)	91,1	0,0	—
Classified, CCI= ([0], [1, 2])	94,3	+3,5	—
Start + end spaces			
Baseline, CCI=(0)	93,0	—	—
Classified, CCI= ([0], [1, 2])	93,3	+0,3	—
Start space			
Baseline, CCI=(0)	96,2	—	—
Classified, CCI= ([0], [1, 2])	98,3	+2,2	—

Table 8c: The performance of s-grams. FIN list, k-words/base forms

FIN Words	Average Precision	% Change	Stat. Sign. Level
No space			
Baseline, CCI=(0)	91,2	—	—
Unclassified, CCI=(0, 1)	90,8	-0,4	—
Classified, CCI= ([0], [1, 2])	96,7	+6,0	—
Start + end spaces			
Baseline, CCI=(0)	89,4	—	—
Classified, CCI= ([0], [1, 2])	95,0	+6,3	—
Start space			
Baseline, CCI=(0)	96,2	—	—

Classified, CCI= ([0], [1, 2])	98,2	+2,0	—

Table 8d: The performance of s-grams. FIN list, k-words/genitive forms

List Type, Key Length	Average Precision	% Change	Stat. Sign. Level
ENG1, N=52; *No of characters >= 9, N=34*			
Baseline, CCI=(0)	68,0	+10,3%	—
Classified, CCI= ([0], [1, 2])	75,0		
No of characters 7-8, N=12			
Baseline, CCI=(0)	36,9	+43,0%	—
Classified, CCI= ([0], [1, 2])	52,9		
No of characters =< 6, N=6			
Baseline, CCI=(0)	19,3	+26,9%	—
Classified, CCI= ([0], [1, 2])	24,5		
ENG2, N=41; No of characters >= 9, N=9			
Baseline, CCI=(0)	47,3	-8,2%	—
Classified, CCI= ([0], [1, 2])	43,4		
No of characters 7-8, N=20			
Baseline, CCI=(0)	14,7	+78,9%	0,001
Classified, CCI= ([0], [1, 2])	26,3		
No of characters =< 6, N=12			
Baseline, CCI=(0)	2,5	+240,0%	—
Classified, CCI= ([0], [1, 2])	8,5		

Table 9: The effects of query key length on the performance of classified s-grams

List Type, LCS Length	Average Precision	Baseline/Classified
ENG1, N=48 - No of shared characters =< 8, N=26		
Baseline, CCI=(0)	34,3	70,7
Classified, CCI= ([0], [1, 2])	48,5	
No of shared characters > 8, N=22		
Baseline, CCI=(0)	84,6	95,5

Classified, CCI= ([0], [1, 2])	88,6	
ENG2, N=27 - No of shared characters =< 8, N=24		
Baseline, CCI=(0)	18,9	70,5
Classified, CCI= ([0], [1, 2])	26,8	
No of shared characters > 8, N=3		
Baseline, CCI=(0)	70,8	96,6
Classified, CCI= ([0], [1, 2])	73,3	
GER, N=48 - No of shared characters =< 8, N=19		
Baseline, CCI=(0)	38,7	78,2
Classified, CCI= ([0], [1, 2])	49,5	
No of shared characters > 8, N=29		
Baseline, CCI=(0)	84,0	92,8
Classified, CCI= ([0], [1, 2])	90,5	
SWE, N=48 - No of shared characters =< 8, N=19		
Baseline, CCI=(0)	53,6	82,5
Classified, CCI= ([0], [1, 2])	65,0	
No of shared characters > 8, N=29		
Baseline, CCI=(0)	86,0	96,1
Classified, CCI= ([0], [1, 2])	89,5	

Table 10: The effects of LCS length on the performance of classified s-grams.

Acknowledgements

This research is part of the research project *Query structures and dictionaries as tools in concept-based and cross-lingual information retrieval* funded by the Academy of Finland (Research Projects 44703; 49157).

References

- Angell, R., Freund, G., and Willet, P. (1983) "Automatic spelling correction system using a trigram similarity measure ". *Information Processing & Management*, **19**(4), 255-261.
- Conover, W.J. (1980) *Practical non-parametric statistics*. New York, NY: John Wiley & Sons.
- Damerau, F. (1964) "A technique for computer detection and correction of spelling errors ". *Communications of the ACM*, **7**, 171-176.
- Fujii, A. and Ishikawa, T. (2001) "Japanese/English cross-language information retrieval: exploration of query translation and transliteration ". *Computers and the Humanities*, **35**(4), 389-420.
- Gadd, T. (1988) "Fisching for werds: phonetic retrieval of written text in information retrieval systems ".

Program, **22**(3), 222-237.

- Gadd, T. (1990) "Phonix: the algorithm ". *Program*, **24**(4), 363-369.
- Hall, P. and Dowling, G. (1980) "Approximate string matching. " *Computing Surveys*, **12**(4), 381-402.
- Hedlund, T., Keskustalo, H., Pirkola, A., Sepponen, M., Järvelin, K. (2001) "Bilingual tests with Swedish, Finnish and German queries: dealing with morphology, compound words and query structure. " *In*: *Cross-language information retrieval and evaluation. Cross-language evaluation forum workshop, CLEF 2000, Lisbon, Portugal, September 22-23, 2000, Revised Papers*. pp. 211-225. Heidelberg: Springer, (Lecture Notes in Computer Science, Vol. 2069)
- Järvelin, K. and Kekäläinen, J. (2000) "IR evaluation methods for retrieving highly relevant documents". *Proceedings of the 23th Annual International ACM SIGIR on Research and Development in Information Retrieval, Athens, July 24-28, 2000*, pp. 41-48. New York, NY: ACM Press.
- Karlsson, F. (1983) *Suomen kielen äänne-ja muotorakenne*. [Phonological and morphological structures in Finnish]. Porvoo - Hki - Juva: WSOY. [In Finnish]
- Kosinov, S. (2001) Evaluation of n-grams conflation approach in text-based information retrieval. Paper delivered at *Infotech Oulu, International Workshop on Information Retrieval*, Oulu, Finland, 19.-21.9. 2001.
- O'Rourke, A.J., Robertson, A.M. and Willett, P. (1997) Word variant identification in old French. *Information Research*, **2**(4).
- Oard, D. and Diekema, A. (1998) "Cross-Language Information Retrieval". *Annual Review of Information Science and Technology (ARIST)*, **33**, 223-256.
- Pfeifer, U., Poersch, T. and Fuhr, N. (1996) "Retrieval effectiveness of proper name search methods". *Information Processing & Management*, **32**(6), 667-679.
- Pirkola, A. (1998) "The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval". *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, Aug. 24-28, 1998*, pp. 55-63. New York, NY: ACM Press.
- Pirkola, A. (2001) "Morphological typology of languages for IR". *Journal of Documentation*, **57** (3), 330-348.
- Pirkola, A. and Järvelin, K. (2001) "Employing the resolution power of search keys". *Journal of the American Society for Information Science and Technology*, **52**(7), 575 -583.
- Pirkola, A., Hedlund, T., Keskustalo, H., and Järvelin, K. (2001) "Dictionary-based cross-language information retrieval: problems, methods, and research findings". *Information Retrieval*, **4**(3/4), 209-230.
- Robertson, A.M. and Willett, P. (1998) "Applications of n-grams in textual information systems". *Journal of Documentation*, **54**(1), 48-69.
- Salton, G. (1989) *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Reading, MA: Addison-Wesley.
- Sormunen, E. (2000) "A novel method for the evaluation of Boolean query effectiveness across a wide operational range". *Proceedings of the 23 rd Annual ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, July 24-28, 2000*, pp. 25-32. New York, NY: ACM Press,
- Ullmann, J.R. (1977) "A binary n-gram technique for automatic correction of substitution, deletion, insertion, and reversal errors in words". *Computer journal*, **20**(2), 141-147.

Xu, J., and Croft W.B. (1998) "Corpus-based stemming using cooccurrence of word variants". *ACM Transactions on Information Systems*, **16**(1), 61-81.

- Zobel, J. and Dart, P. (1995) "Finding approximate matches in large lexicons". *Software - practice and experience*, **25**(3), 331-345.

How to cite this paper

Pirkola, A, Keskustalo, Heikki, Leppänen, Erkka, Känsälä, Antti-Pekka and Järvelin, Kalervo (2002) "Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants." *Information Research*, **7**(2) [Available at http://InformationR.net/ir/7-2/paper126.html]

Contents **Home**
