Authors:	Keskustalo Heikki, Pirkola Ari, Visala Kari, Leppänen Erkka, Järvelin Kalervo
Name of article:	Non-adjacent Digrams Improve Matching of Cross-Lingual Spelling Variants
Name of work:	String Processing and Information Retrieval : 10th International Symposium, SPIRE 2003, Manaus, Brazil
Editors of work:	Nascimento Mario A, De Moura Edleno S, Oliveira Arlindo L
Year of publication:	2003
ISBN:	3-540-20177-7
Publisher:	Springer
Pages:	252-265
Series name and number:	Lecture Notes in Computer Science 2857
ISSN:	0302-9743
Discipline:	Natural sciences / Computer and information sciences
Language:	en
School/Other Unit:	School of Information Sciences

URL: http://www.springerlink.com/content/v7x75bdup9ta6n3g/fulltext.pdf
URN: http://urn.fi/urn:nbn:uta-3-871
DOI: http://dx.doi.org/10.1007/978-3-540-39984-1_19

Additional infromation:
The original publication is available at www.springerlink.com.

In: Nascimento, M.A., de Moura, E.S., Oliveira, A.L, (Eds.). Proceedings of the 10th International Symposium, SPIRE 2003 Manaus, Brazil, October 2003. Berlin: Springer, Lecture Notes in Computer Science 2857, pp. 252 - 265.

Non-adjacent Digrams Improve Matching of Cross-Lingual Spelling Variants

Heikki Keskustalo, Ari Pirkola, Kari Visala, Erkka Leppänen, Kalervo Järvelin

Department of Information Studies, University of Tampere, Finland

Abstract. Untranslatable query keys pose a problem in dictionary-based cross-language information retrieval (CLIR). One solution consists of using approximate string matching methods for finding the spelling variants of the source key among the target database index. In such a setting, it is important to select a matching method suited especially for CLIR. This paper focuses on comparing the effectiveness of several matching methods in a cross-lingual setting. Search words from five domains were expressed in six languages (French, Spanish, Italian, German, Swedish, and Finnish). The target data consisted of the index of an English full-text database. In this setting, we first established the best method among six baseline matching methods for each language pair. Secondly, we tested novel matching methods based on binary digrams formed of both adjacent and non-adjacent characters of words. The latter methods consistently outperformed all baseline methods.

1 Introduction

In dictionary-based cross-language information retrieval (CLIR) a source query is typically translated word-by-word into the target language by using machine-readable dictionaries. However, due to the terminology missing from the dictionaries, untranslatable keys often appear in the queries thus posing a source for translation errors [3]. A trivial solution for handling the untranslatable keys is to use them as such in the target query. This solution succeeds sometimes, e.g., in case of some acronyms and proper names, while failing in many cases. A more advanced solution is to use approximate string matching to find the most similar word or words for the source keys from the target index, which can be placed into the target query [6].

The goal in approximate string matching is to rank or identify similar strings with respect to the given key. What is meant by similarity depends on the characteristics of the particular application. For example, human keyboard operators introduce reversal, insertion, deletion and substitution errors, while optical character recognition machines typically introduce substitution and reject errors [10]. Thus, it makes sense to consider different string pairs being similar in different usage contexts. Specifically, in case of untranslatable words in CLIR, the goal is to identify cross-lingual spelling variants. For example, by using the Spanish key *escleroterapia* we may wish to find its English variant *sclerotherapy* from the database. It is not clear whether the similarity measures developed for other

In: Nascimento, M.A., de Moura, E.S., Oliveira, A.L, (Eds.). Proceedings of the 10th International Symposium, SPIRE 2003. Manaus, Brazil, October 2003. Berlin: Springer, Lecture Notes in Computer Science 2857, pp. 252 - 265.

aims than CLIR are optimal for finding spelling variants. The similarity measure should take into account the special characteristics of the cross-lingual spelling variant strings.

Previous research has shown many successful applications of approximate string matching in information retrieval, see, e.g., [9] for a review of the usage of n-grams in textual information systems. Approximate matching improved proper name searching as compared to identical matching, and digrams performed best among the tested single methods among the top results in [5]. In [1] a trigram similarity measure was used for successfully identifying dictionary spellings of misspelled word forms. The study by [2] describes a multilingual retrieval system based on a vector space model in which the documents were represented by using 5-grams or 6-grams. In [11] several similarity methods for phonetic matching were tested and the best method was found to be a variant of edit distance utilizing letter groupings. Recently, [8] proposed a novel method utilizing automatically derived character transformation rules together with conventional n-grams for improving cross-lingual spelling variant matching. Also, combining evidence resulting from distinct matching methods seems to further improve the matching results [5] [11].

In this paper, we will utilize a cross-lingual research setting containing test words from five domains. Each word is expressed in seven languages (six source languages, and English as the target language). By using the words in the source languages as search keys we will compare the effectiveness of several approximate string matching methods. The main research question is to measure the effectiveness of several novel matching methods. These methods utilize non-adjacent binary digrams and we compare their effectiveness to the baseline results. The baseline matching methods include conventional n-grams of several lengths, longest common subsequence, edit distance, and exact match. The rest of the paper is organized as follows. Section 2 introduces the methodology, Section 3 presents the findings, and Section 4 contains the discussion and conclusions.

2 Preliminaries

2.1 Skip-grams

The concept of *skip-grams* (binary digrams formed from non-adjacent letters) as a solution specifically for cross-lingual spelling variation problems was introduced in [7]. This paper contributes to the issue by testing the effectiveness of several novel skip-gram types and reporting the effectiveness of several baselines, using six source languages with respect to one target language, and by using query keys from several domains. Next, we will present a notation generalized from [7], defining how the skip-gram similarity between two strings is computed.

Let the *gram class (GC)*, expressed by a set of non-negative integers, indicate the number of skipped characters when digrams are formed from the string $S = s_1 s_2 s_3 ... s_n$. In other words, the gram class defines how one digram set *(DS)* is formed from the string S. For example, if $GC = \{0,1\}$ then for string $S = s_1 s_2 s_3 s_4$

In: Nascimento, M.A., de Moura, E.S., Oliveira, A.L, (Eds.). Proceedings of the 10th International Symposium, SPIRE 2003. Manaus, Brazil, October 2003. Berlin: Springer, Lecture Notes in Computer Science 2857, pp. 252 - 265.

we form the DS by skipping both zero and one characters in S when the digrams are formed, thus $DS_{\{0,1\}}(S) = \{s_1s_2, s_1s_3, s_2s_3, s_2s_4, s_3s_4\}$. We call the largest value in GC the spanning length, e.g., for GC=$\{0,1\}$ the spanning length is one. Let the *character combination index (CCI)* be a set of gram classes enumerating all the digram sets to be produced from S. For example, if $CCI = \{\{0\}, \{1, 2\}\}$ then for the string $S = s_1s_2s_3s_4$ we form two digram sets, namely $DS_{\{0\}}(S)$ $= \{s_1s_2, s_2s_3, s_3s_4\}$ (by zero skipping) and $DS_{\{1,2\}}(S) = \{s_1s_3, s_1s_4, s_2s_4\}$ (by skipping both one and two characters). Finally, the similarity measure *(SIM)* is defined between two strings S and T with respect to the given CCI in the following way:

$$SIM_{CCI}(S,T) = \frac{\sum_{i \in CCI} |DS_i(S) \cap DS_i(T)|}{\sum_{i \in CCI} |DS_i(S) \cup DS_i(T)|} \; . \tag{1}$$

For example, if S = *abcd,* T = *apcd,* and CCI = {{*0*},{*1,2*}}, we apply the set operations pairwise to digram sets $DS_{\{0\}}(S)$ = {ab, bc, cd} and $DS_{\{0\}}(T)$ = {ap, pc, cd}, and then to $DS_{\{1,2\}}(S)$ = {ac, ad, bd} and $DS_{\{1,2\}}(T)$ = {ac, ad, pd}, thus $SIM_{\{\{0\},\{1,2\}\}}(abcd, apcd)$ = (1+2)/(5+4) $\approx$ 0.33. The basis for the formula above is the similarity measure for two sets given in [5]. Other similarity measures could also be used analogously for a pair of sets, e.g., Dice or Overlap coefficients [4].

2.2 Cross-Lingual Spelling Variation

Cross-lingual spelling variation refers to word variation where a language pair shares words written differently but having the same origin, for example, technical terms derived from Latin or Greek, or proper names. At the string level, this variation often involves single character insertions, deletions and substitutions, or combinations of them [7]. For instance, transforming an Italian variant *ematome* into the English variant *hematoma* involves a single character insertion (*h*) and substitution (*e*⇒*a*), while transforming the corresponding Finnish variant *hematooma* involves a single character deletion (*o*). On the other hand, transforming Swedish variant *heksaklorid* into English *hexachloride* involves combinations of deletion and substitution (*ks*⇒*x*), and substitution and insertion (*k*⇒*ch*), and a single insertion (*e*). Also more complex combinations of operations occur, like between Italian and English term variants *ginecofobia* and *gynephobia*.

In [7] the effectiveness of two combinations of skip-gram classes (CCI = {{0,1}} and CCI = {{0},{1,2}}) was tested in a cross-lingual setting using English, Swedish and German search keys and Finnish as the target language. In most cases the skip-grams outperformed conventional digrams. However, the performance levels of several gram class combinations were not tested. Therefore, we will next hypothesize some novel CCI values for the experimental testing by considering the properties of cross-lingual spelling variation presented above, and by associating these properties to the skip-gram classes.

In: Nascimento, M.A., de Moura, E.S., Oliveira, A.L, (Eds.). Proceedings of the 10th International Symposium, SPIRE 2003. Manaus, Brazil, October 2003. Berlin: Springer, Lecture Notes in Computer Science 2857, pp. 252 - 265.

2.3 Gram Classes and Spelling Variation

Cross-lingual spelling variation typically involves single character insertions, deletions and substitutions, or their two-character combinations. Therefore, in this research we restrict our attention to gram classes having spanning length two or less. The gram classes can be interpreted in the following way considering the kind of evidence they carry forward from their host string. Gram class {0} is a special case of skip-grams expressing conventional digrams formed from adjacent letters of the host string.

Gram class {1} allows one substitution, for example, substrings *gin* and *gyn* share class {1} digram *gn* although they do not share any common digrams or trigrams. This gram class is possibly meaningful from the CLIR point of view, as single character substitutions occur frequently between cross-lingual spelling variants. The gram class {0,1} allows one insertion between adjacent letters or a deletion of one letter separating two characters. For example, substrings *ic* and *isc* share class {0,1} digram *ic*.

The class {1,2} allows one insertion between letters separated by one character, or a deletion of one of the two characters separating two characters. For example, substrings *eksa* and *exa* share class {1,2} digram *ea*. Thus classes {0}, {1}, {0,1}, and {1,2} can be considered as having potential importance in CLIR. For research economical reasons, we left outside of testing some gram classes that we concluded to be less meaningful from CLIR point of view. These include the gram class {2} allowing substitution of exactly two characters, class {0,2} allowing substitution, insertion or deletion of exactly two characters, and class {0,1,2} allowing several types of combinations of substitutions, insertions and deletions.

Also negative effects may be introduced by the utilization of the novel gram classes. Hence we proceed next on running tests in order to evaluate their effectiveness in practice.

2.4 Test Data

The test data consists of three parts: the search keys, the target words, and the set of correct answers (relevance judgments).

Altogether more than 1600 search keys were used in the experiment. The search key lists were formed as follows. The first word set of 217 English words was selected from the database index and translated into the six search languages intellectually by one of the researchers. Several translation resources were used for performing this task. Thus 217 word tuples in seven languages were formed. These words were scientific terms, mostly medical or biological, called *bio* terms in the tables, or geographical place names (*geo*). As an example, the tuple *(hybridoma, hybridooma, hybridom, hybridzelle, hybridome, ibridoma, hibridoma)* contains word variants ordered by languages English, Finnish, Swedish, German, French, Italian, Spanish. We used a 26 letter alphabet augmented by letters å, ä, ö, and ü. All the translations of the first word set were checked by native speakers or advanced students majoring in each particular language. Very few corrections took place. Also, every third tuple from this set was selected as the

In: Nascimento, M.A., de Moura, E.S., Oliveira, A.L, (Eds.). Proceedings of the 10th International Symposium, SPIRE 2003. Manaus, Brazil, October 2003. Berlin: Springer, Lecture Notes in Computer Science 2857, pp. 252 - 265.

training data to be used exclusively in the analyses performed prior to final test runs. Therefore, 72 training word tuples and 145 final test word tuples were obtained from the first word set. The second search word set was gathered by first collecting 126 supplementary English words. These were from the domains of economics (abbreviated as *econ* in the tables), technology (*tech*), and miscellaneous (*misc*) containing common foreign words. The words were translated into the six search languages by one of the researchers. Thus, altogether 271 final test word tuples were formed, each containing the English word variant and its corresponding search key variants in six languages.

The target words consisted of a list containing all words of an English full-text database index (Los Angeles Times used in CLEF 2000 experiments) [11]. It contains around 189,000 unique word forms. They are either in basic forms as recognized by the morphological analyser ENGTWOL used in indexing, or in case of unrecognised word forms, the original words as such. All words are written in monocase.

The set of relevance judgments consisted of the English word variants in the tuples. For each search key there was precisely one correct English counterpart, but in some cases there were more than one search key variants with respect to one English word. All English variants of the first search word set occurred in the original Los Angeles Times index. For the second search word set, three English keys were not found from the index list originally and they were added to the list prior to final runs.

2.5 Matching Methods

The effectiveness of each matching method was measured by calculating the average precision at 100 % recall point. Sometimes several target words gained the same similarity value with respect to the key. Therefore, we evaluated the precision by using two methods. In the worst case method we assumed the correct word to be the last word among the group of words (cohort) having the same SIM value. In the average case method we assumed the correct word to be in the middle of the cohort. In practise, these two methods gave almost the same values. This is because typically the cohorts were small. Therefore, in Tables 1-13 we report only average case results.

The following *baseline methods* were tested:

- Exact match
- Edit distance
- Longest common subsequence
- Digrams (conventional digrams, i.e., skip-grams with CCI ={{0}})
- Trigrams
- Tetragrams

The exact match and edit distance were used as similarity measures as such. The longest common subsequence (*lcs*) as such would have favoured long index words, therefore, we substracted the value of lcs from the mean length of the two

In: Nascimento, M.A., de Moura, E.S., Oliveira, A.L, (Eds.). Proceedings of the 10th International Symposium, SPIRE 2003. Manaus, Brazil, October 2003. Berlin: Springer, Lecture Notes in Computer Science 2857, pp. 252 - 265.

words compared to make the similarity measure more meaningful. The digrams, trigrams and tetragrams were utilized by applying the formula (1) to the sets of n-grams derived from the strings.

Secondly, the following *skip-gram methods* were tested:

- CCI = {{0},{1}}
- CCI = {{0},{0,1}}
- CCI = {{0},{1,2}}
- CCI = {{0},{1},{0,1}}
- CCI = {{0},{1},{1,2}}
- CCI = {{0},{0,1},{1,2}}
- CCI = {{0},{1},{0,1},{1,2}}

Formula (1) was applied for computing the skip-gram similarity. We compare the skip-gram results with several baseline results for each language pair.

Word beginnings and endings deserve special attention in approximate string matching [4]. Therefore, we used training data with conventional digrams, trigrams, and skip-grams with CCI = {{0},{1,2}} (as found to be successful in CLIR in [7]) and tested three matching variations: (i) word starts are padded with an appropriate number of special characters, (ii) both word starts and endings are padded, and (iii) no padding is performed: only the characters of the string itself are considered. The choice (ii) turned out to give the best results in most cases, although in some cases the choice (i) gave slightly better results. The choice (iii) gave consistently the worst results. On the basis of these results we decided to utilize both word start and end padding in all of the final runs.

3 Findings

Next, we will discuss the results of each language pair individually. In Sections 3.1 - 3.6 we first establish the best one of the six baseline methods for each language pair. Secondly, we present the results of the three best skip-gram methods and compare them to the best baseline method based on the average precision over all domains. This is a very conservative approach. Therefore, we also compare the effectiveness of the best skip-gram method with respect to *trigrams* in Sections 3.1 - 3.6, as a well-known method. The effectiveness of the methods is discussed also at the individual terminological domains.

3.1 Finnish-English

The best baseline method in Finnish-to-English matching was edit distance with average precision 45.9 % (Table 1). Edit distance was the best method also in each single domain, except in case of of *miscellanous words*, where the digrams slightly outperformed it. Digram based matching was in the second place, followed by our variation of the longest common subsequence. Trigrams and tetragrams performed poorly. Variation between the domain results is considerable (32.5 % to 67.7 % for edit distance). Finnish and English word variants were

In: Nascimento, M.A., de Moura, E.S., Oliveira, A.L, (Eds.). Proceedings of the 10th International Symposium, SPIRE 2003. Manaus, Brazil, October 2003. Berlin: Springer, Lecture Notes in Computer Science 2857, pp. 252 - 265.

Table 1. Finnish-English baseline results (Precision %).

Domain	Edit distance	Digrams	LCS	Trigrams	Tetragrams	Exact match
Bio (N=92)	67.7	61.4	54.0	49.2	45.6	0.0
Geo (N=55)	35.4	30.0	27.3	29.3	29.6	9.1
Econ (N=31)	36.5	32.2	27.5	30.7	24.8	0.0
Tech (N=36)	36.2	31.6	30.2	21.2	16.7	0.0
Misc (N=59)	32.5	33.8	31.5	28.9	26.3	0.0
Avg. (N=273)	45.9	41.9	37.6	35.0	32.0	1.8

rarely identical, except some *place names*, as reflected by the low precision figure (1.8 %) for exact matching. Edit distance is selected as the comparison basis for Table 2 as the best baseline method. As we can see in Table 2, the skip-

Table 2. Finnish-English, the best baseline results (Best BL) and the results of the three best skip-gram methods (Precision %). Improvements are marked with respect to the best baseline, and with respect to the trigrams (in parantheses).

Domain	Best BL	{{0},{0,1},{1,2}}	{{0},{1,2}}	{{0},{1},{0,1},{1,2}}
Bio (N=92)	67.7	69.0 +1.9% (+40.2%)	68.6	67.3
Geo (N=55)	35.4	36.1 +2.0% (+23.2%)	36.7	36.1
Econ (N=31)	36.5	43.5 +19.2% (+41.7%)	40.0	43.7
Tech (N=36)	36.2	49.6 +37.0% (+134.0%)	49.7	48.5
Misc (N=59)	32.5	36.5 +12.3% (+26.3%)	37.5	36.5
Avg. (N=273)	45.9	49.9 +8.7% (+42.6%)	49.7+8.2%	49.2 +7.2%

gram methods outperformed the baseline methods. The best skip-gram method, using CCI = {{0},{0,1},{1,2}}, outperformed edit distance on the average by +8.7 %. In the domain of *technology* the improvement was most notable, +37.0 %, in *economics* +19.2 %, and with *miscellaneous words* +12.3 %. In case of *biological* and *geographical* terms, the improvement was inconsequential. The improvements gained by the best skip-gram method are even higher if compared to digrams (average improvement of +19.1 %), trigrams (+42.6 %) or tetragrams (+55.9 %). Thus, conventional n-grams, especially with large values of n, perform poorly with Finnish as the source language, although n-grams work well elsewhere (see French-English results). The skip-grams generally improve spelling variant matching in case of Finnish-to-English matching. The formula (1) itself allows more finegrained similarity values than edit distance or longest common subsequence. One should notice that the skip-gram methods outperformed edit distance although the conventional digrams did not. Because of this, we conclude that the evidence that the skip-grams carry forward from the host string is better suited for matching spelling variants than the evidence gained by ordinary digrams, trigrams or tetragrams (see Sections 2.2 and 2.3).

In: Nascimento, M.A., de Moura, E.S., Oliveira, A.L, (Eds.). Proceedings of the 10th International Symposium, SPIRE 2003. Manaus, Brazil, October 2003. Berlin: Springer, Lecture Notes in Computer Science 2857, pp. 252 - 265.

3.2 French-English

Both the general level of performance and the order of the best baseline methods are different for the French results (Table 3) as compared to the Finnish results (Table 1). The level of the average precision is much higher (73.4 %)

Table 3. French-English baseline results (Precision %).

Domain	Digrams	Trigrams	Tetragrams	Edit distance	LCS	Exact match
Bio (N=92)	88.3	87.7	87.2	89.2	87.4	41.3
Geo (N=59)	52.5	53.3	52.8	52.4	49.3	27.1
Econ (N=31)	80.1	78.1	76.3	72.3	69.7	41.9
Tech (N=36)	78.4	78.8	78.7	76.4	73.9	66.7
Misc (N=59)	64.6	64.7	64.6	62.8	60.7	37.3
Avg. (N=277)	73.4	73.2	72.7	72.2	69.9	40.8

for French key matching than in case of Finnish keys (45.9 %). Also, in each domain, the results of the different methods are close to each other, except for the exact matching. The single best baseline method for French was conventional digram matching method (average precision 73.4 %), closely followed by trigrams, tetragrams, edit distance and longest common subsequence. On the basis of these results, conventional n-grams are rather well suited for French-to-English spelling variant matching. The English variant in many cases had a rather long common start with the French variant (e.g. *catalytic/catalytique; glycogen/glycogene*), whilst in case of the Finnish variant, character substitutions and insertions were typical also in the beginning and in the middle of the word (*katalyyttinen; glykogeeni*). One can notice from the exact match results that the proportion of identical spelling variants between French and English is relatively high. As the best baseline, digrams are selected as the comparison basis for the skip-gram runs below (Table 4). Also in French-to-English

Table 4. French-English, the best baseline results (Best BL) and the results of the three best skip-gram methods (Precision %). Improvements are marked with respect to the best baseline, and with respect to the trigrams (in parantheses).

Domain	Best BL	{{0},{0,1},{1,2}}	{{0},{1},{0,1},{1,2}}	{{0},{1},{1,2}}
Bio (N=92)	88.3	90.0 +1.9% (+2.6%)	90.0	90.0
Geo (N=59)	52.5	54.5 +3.8% (+2.3%)	54.7	55.0
Econ (N=31)	80.1	83.5 +4.2% (+6.9%)	81.9	81.6
Tech (N=36)	78.4	77.0 -1.8% (-2.3%)	77.0	78.3
Misc (N=59)	64.6	68.9 +6.7% (+6.5%)	69.2	67.3
Avg. (N=277)	73.4	75.5 +2.9% (+3.1%)	75.5 +2.9%	75.2 +2.5%

In: Nascimento, M.A., de Moura, E.S., Oliveira, A.L, (Eds.). Proceedings of the 10th International Symposium, SPIRE 2003. Manaus, Brazil, October 2003. Berlin: Springer, Lecture Notes in Computer Science 2857, pp. 252 - 265.

matching the skip-gram methods outperformed the best baseline method (Table 4), but the improvement was inconsequential (+2.9 % in the best case). The best results were again attained by using CCI={{0},{0,1},{1,2}}. Although the improvements generally were small, in the domain of miscellaneous words the improvement was notable (+6.7 %).

3.3 German-English

The results for the German-English baseline runs are given in Table 5. The best

Table 5. German-English baseline results (Precision %).

Domain	Edit distance	Digrams	Trigrams	Tetragrams	LCS	Exact match
Bio (N=97)	76.6	77.8	75.5	71.1	73.8	15.5
Geo (N=62)	45.5	41.4	42.7	42.1	36.0	14.5
Econ (N=31)	51.1	52.6	52.3	50.6	44.9	9.7
Tech (N=36)	65.6	60.5	58.3	55.3	60.3	27.8
Misc (N=59)	53.3	54.0	52.3	51.4	50.6	22.0
Avg. (N=285)	60.8	60.0	58.9	56.5	55.9	17.6

baseline method was again edit distance (average precision 60.8 %), closely followed by digrams (60.0 %). Also for each single domain the best results were gained by either edit distance or digrams. Compared to Finnish-English runs, trigrams and tetragrams performed rather well. The skip-gram methods out-

Table 6. German-English, the best baseline results (Best BL) and the results of the three best skip-gram methods (Precision %). Improvements are marked with respect to the best baseline, and with respect to the trigrams (in parantheses).

Domain	Best BL	{{0},{1,2}}	{{0},{1},{1,2}}	{{0},{0,1},{1,2}}
Bio (N=97)	76.6	83.6 +9.1% (+10.7%)	83.6	83.9
Geo (N=62)	45.5	46.4 +2.0% (+9.1%)	46.5	47.1
Econ (N=31)	51.1	59.0 +15.5% (+12.8%)	58.9	58.9
Tech (N=36)	65.6	69.0 +5.2% (+18.4%)	69.1	67.2
Misc (N=59)	53.3	57.9 +8.6% (+10.7%)	57.1	57.1
Avg. (N=285)	60.8	65.7 +8.1% (+11.5%)	65.5 +7.7%	65.5 +7.7%

performed the baselines (Table 6). In the best case (CCI = {{0},{1,2}}) the improvement was +8.1 %. The largest domain improvements took place with *economics* and *biology* (+15.5 % and +9.1 %, respectively). The best skip-gram method for Finnish and French (CCI={{0},{0,1},{1,2}}) performed well also with the German keys (average improvement +7.7 %).

In: Nascimento, M.A., de Moura, E.S., Oliveira, A.L, (Eds.). Proceedings of the 10th International Symposium, SPIRE 2003. Manaus, Brazil, October 2003. Berlin: Springer, Lecture Notes in Computer Science 2857, pp. 252 - 265.

3.4 Italian-English

The results for the Italian-English baseline runs are given in Table 7. Edit dis-

Table 7. Italian-English baseline results (Precision %).

Domain	Edit distance	Digrams	LCS	Trigrams	Tetragrams	Exact match
Bio (N=98)	67.2	62.1	57.7	56.0	50.9	6.1
Geo (N=65)	53.5	53.7	49.3	54.9	55.0	27.7
Econ (N=31)	39.4	41.3	35.4	32.7	26.7	0.0
Tech (N=36)	50.5	39.8	46.1	39.5	36.5	19.4
Misc (N=59)	38.3	35.9	39.5	35.8	33.3	10.2
Avg. (N=289)	53.2	49.9	48.3	47.1	43.8	12.8

tance was the best baseline method for Italian keys (average precision 53.2 %) followed by digrams (49.9 %). The results indicate that nature of the domain terminology may have a strong impact on the selection of the appropriate matching method. For example, with *geographical terms* all matching methods perform quite alike, even the tetragrams, while on the other hand tetragrams perform very poorly, e.g., with the domain of *economics*. Edit distance is competitive in each domain. All skip-gram methods again outperformed all baseline methods

Table 8. Italian-English, the best baseline results (Best BL) and the results of the three best skip-gram methods (Precision %). Improvements are marked with respect to the best baseline, and with respect to the trigrams (in parantheses).

Domain	Best BL	{{0},{1,2}}	{{0},{1},{1,2}}	{{0},{1},{0,1},{1,2}}
Bio (N=98)	67.2	70.2 +4.5% (+25.4%)	69.8	69.3
Geo (N=65)	53.5	60.6 +13.3% (+10.4%)	60.2	59.1
Econ (N=31)	39.4	45.1 +14.5% (+37.9%)	45.0	47.6
Tech (N=36)	50.5	48.5 -4.0% (+22.8%)	49.1	49.1
Misc (N=59)	38.3	43.5 +13.6% (+21.5%)	44.1	43.1
Avg. (N=289)	53.2	57.2 +7.5% (+21.4%)	57.2 +7.5%	56.8 +6.8%

(Table 8). The largest improvement (+7.5 %) with respect to the baseline results was gained by using CCI = {{0},{1,2}} or CCI = {{0},{1}, {1,2}}. This improvement figure for the Italian keys is of the same magnitude as the corresponding figure for Finnish (+8.7 %) and German (+8.1 %). Looking at the individual domains, for *technological* terms the performance dropped (-4.0 %) from the best baseline by using skip-grams, but it improved with all other domains, especially with the words in *economics* (+14.5 %), *miscellaneous* words (+13.6 %) and *geographical* words (+13.3 %).

In: Nascimento, M.A., de Moura, E.S., Oliveira, A.L, (Eds.). Proceedings of the 10th International Symposium, SPIRE 2003. Manaus, Brazil, October 2003. Berlin: Springer, Lecture Notes in Computer Science 2857, pp. 252 - 265.

3.5 Spanish-English

The results for the Spanish-English baseline runs are given in Table 9. With

Table 9. Spanish-English baseline results (Precision %).

Domain	Edit distance	Digrams	Trigrams	Tetragrams	LCS	Exact match
Bio (N=94)	72.8	67.6	63.2	57.7	63.1	6.4
Geo (N=57)	54.4	55.5	54.9	55.0	51.7	31.6
Econ (N=31)	44.5	45.5	45.9	45.6	38.1	6.5
Tech (N=36)	59.5	57.5	57.9	55.9	51.3	22.2
Misc (N=59)	39.6	41.1	41.9	40.8	40.6	6.8
Avg. (N=277)	57.0	55.7	54.3	52.0	51.6	13.7

Spanish keys, edit distance was again the best baseline method (average precision 57.0 %), followed by digrams (55.7 %) and trigrams (54.3 %). Effectiveness of the matching methods is again sensitive to the terminological domain. For example, with *geographical* and *miscellaneous* terms, and with terms in *economics*, n-grams worked at least as well as edit distance, but for bio terms the performance level goes down rapidly as n increases. The best skip-gram method outperformed the

Table 10. Spanish-English, the best baseline results (Best BL) and the results of the three best skip-gram methods (Precision %). Improvements are marked with respect to the best baseline, and with respect to the trigrams (in parantheses).

Domain	Best BL	{{0},{1,2}}	{{0},{1},{1,2}}	{{0},{0,1},{1,2}}
Bio (N=94)	72.8	72.7 -0.1% (+15.0%)	72.6	72.5
Geo (N=57)	54.4	61.2 +12.5% (+11.5%)	61.8	59.4
Econ (N=31)	44.5	48.4 +8.8% (+5.4%)	48.1	48.3
Tech (N=36)	59.5	63.1 +6.1% (+9.0%)	62.1	62.0
Misc (N=59)	39.6	42.6 +7.6% (+1.7%)	42.6	42.6
Avg. (N=277)	57.0	60.0 +5.3% (+10.5%)	59.9 +5.1%	59.4 +4.2%

best baseline by +5.3 % (Table 10). The highest average precision was gained by the method with CCI = {{0},{1,2}}. At individual domains, the performance improved especially in case of *geographical* words (+12.5 %). Other domains with notable improvement by using the Spanish keys were *economics* (+8.8 %) and *miscellaneous* words (+7.6 %). The performance improvements were remarkable also in case of Italian keys with these three domains. Moreover, the Spanish result improved also in case of *technological* words (+6.1 %), but not in the domain of *biology*.

In: Nascimento, M.A., de Moura, E.S., Oliveira, A.L, (Eds.). Proceedings of the 10th International Symposium, SPIRE 2003. Manaus, Brazil, October 2003. Berlin: Springer, Lecture Notes in Computer Science 2857, pp. 252 - 265.

3.6 Swedish-English

The results for the Swedish-English baseline runs are presented in Table 11. With

Table 11. Swedish-English baseline results (Precision %).

Domain	Digrams	Edit distance	Trigrams	Tetragrams	LCS	Exact match
Bio (N=92)	73.8	68.7	69.9	66.8	62.9	10.9
Geo (N=56)	48.4	47.0	48.2	47.4	43.6	26.8
Econ (N=31)	41.7	40.2	37.5	36.9	31.4	12.9
Tech (N=36)	58.4	59.1	55.9	53.0	53.5	25.0
Misc (N=59)	48.1	51.1	48.3	47.9	44.8	27.1
Avg. (N=274)	57.5	56.0	55.3	53.6	50.3	19.7

Swedish as the source language, conventional digrams were the single best baseline method (average precision 57.5 %), closely followed by edit distance (56.0 %) and trigrams (55.3 %). Based on average precision, the skip-gram methods

Table 12. Swedish-English, the best baseline results (Best BL) and the results of the three best skip-gram methods (Precision %). Improvements are marked with respect to the best baseline, and with respect to the trigrams (in parantheses).

Domain	Best BL	{{0},{0,1},{1,2}}	{{0},{1},{1,2}}	{{0},{1,2}}
Bio (N=92)	73.8	79.9 +8.3% (+14.3%)	79.6	79.3
Geo (N=56)	48.4	49.2 +1.7% (+2.1%)	49.1	49.6
Econ (N=31)	41.7	46.6 +11.8% (+24.3%)	46.6	47.3
Tech (N=36)	58.4	63.6 +8.9% (+13.8%)	63.6	63.6
Misc (N=59)	48.1	53.9 +12.1% (+11.6%)	54.4	53.4
Avg. (N=274)	57.5	62.1 +8.0% (+12.3%)	62.1 +8.0%	62.0 +7.8%

again outperformed the best baseline method (Table 12). The greatest improvement (+8.0 %) was due to skip-grams with CCI = {{0}, {0,1},{1,2}}, but the four best skip-gram methods outperformed the best baseline by at least +7.5 %. Considering the individual domains, the biggest improvement took place with *miscellaneous* words (+12.1%) as well as with the words of *economics* (+11.8 %), *technology* (+8.9 %) and *biology* (+8.3 %).

4 Discussion and Conclusions

In this paper, we have explored cross-language spelling variant matching. We studied first the effectiveness of six baseline methods and then of seven skip-gram methods. The research setting contained more than 1600 source keys partitioned into six languages and five term domains, and about 189,000 target

In: Nascimento, M.A., de Moura, E.S., Oliveira, A.L, (Eds.). Proceedings of the 10th International Symposium, SPIRE 2003. Manaus, Brazil, October 2003. Berlin: Springer, Lecture Notes in Computer Science 2857, pp. 252 - 265.

words in English. We found that among all tested matching methods, the skip-gram techniques were the most effective one for finding cross-lingual spelling variants in languages based on Latin alphabet.

Table 13 presents a summary of the results. Here we compare the results of skip-grams to conventional digrams. In our study, conventional digrams were always the best or the second best baseline, and they always outperformed trigrams. Even in absolute terms, the improvements gained by skip-grams were

Table 13. Average performance of the best skip-gram matching methods, as compared to conventional digrams (Precision %).

Source language	Digrams	{{0},{0,1},{1,2}}	{{0},{1,2}}	{{0},{1},{1,2}}
Finnish (N=273)	41.9	49.9 +19.1%	49.7 +18.6%	49.1 +17.2%
French (N=277)	73.4	75.5 +2.9%	74.7 +1.8%	75.2 +2.5%
German (N=285)	60.0	65.5 +9.2%	65.7 +9.5%	65.5 +9.2%
Italian (N=289)	49.9	56.5 +13.2%	57.2 +14.6%	57.2 +14.6%
Spanish (N=277)	55.7	59.4 +6.6%	60.0 +7.7%	59.9 +7.5%
Swedish (N=274)	57.5	62.1 +8.0%	62.0 +7.8%	62.1 +8.0%

notable as compared to the digrams. As we can see in Table 13, in case of Finnish as the source language, the average precision improved from 41.9 % to 49.9 % by using CCI={{0},{0,1},{1,2}}. Other large improvements took place by using CCI={{0},{1,2}} for German (60.0 % to 65.7 %) and Italian (49.9 % to 57.2 %).

Skip-grams seem to be well suited especially for some individual domains. For example, in *economics*, compared to the best baseline, an improvement of +19.2 % (Table 2), +15.5 % (Table 6), +14.5 % (Table 8), +8.8 % (Table 10), and +11.8 % (Table 12) was achieved with Finnish, German, Italian, Spanish, and Swedish, respectively, as the source languages.

Although a study on time and space aspects is beyond the scope of the present study, we mention that our recent implementation in C for the skip-grams (tuned presently for formula (1) and for CCI = {{0,1}}) has an average response time of 0.08 seconds for finding the best match from among 192,000 words (CLEF 2000 *LA Times* collection; 1000 key word sample, average key length of 8.9 characters; Sun Ultra-10 workstation, 333 MHz, 512 MB RAM).

Skip-grams may be of interest to areas where conventional n-grams are used today, including, e.g., monolingual and cross-lingual spelling variant matching and music retrieval based on pitch sequencies.

We propose the following research agenda for applying skip-grams in novel areas. Following the inference process presented in Section 2.3 for the cross-lingual spelling variants, one infers the CCI values which could be useful in the domain of interest. In other words, instead of using, e.g., simply CCI = {{0}} (conventional digrams) one makes a *CCI hypothesis*, and runs the empirical tests using both conventional digrams and the novel CCI value. Because skip-grams

In: Nascimento, M.A., de Moura, E.S., Oliveira, A.L, (Eds.). Proceedings of the 10th International Symposium, SPIRE 2003. Manaus, Brazil, October 2003. Berlin: Springer, Lecture Notes in Computer Science 2857, pp. 252 - 265.

are a simple technique, replacing conventional digrams with them is attractive in case they give better results. On the basis of this study, skip-grams do give better results in CLIR than the well-known conventional methods tested.

Our future plans include implementing the skip-gram matching as part of a real-time CLIR system and developing more advanced matching methods by utilizing, e.g., character transformation rule information [8].

Acknowledgements

This work was partly funded by *Clarity* - Proposal/Contract no.: IST-2000-25310. The target index was processed by using ENGTWOL (Morphological Transducer Lexicon Description of English): Copyright ©1989-1992 Atro Voutilainen and Juha Heikkilä. TWOL-R (Run-Time Two-Level Program): Copyright ©1983-1992 Kimmo Koskenniemi and Lingsoft Oy. We wish to thank the anonymous referees and the members of the FIRE research group for useful suggestions.

References

1. Angell, R. C., Freund, G. E., Willett, P. (1983) Automatic Spelling Correction Using a Trigram Similarity Measure, *Information Processing & Managament*, 4, 1983, 255 - 261.
2. Damashek, M.(1995) Gauging Similarity with n-Grams: Language-Independent Sorting, Categorization, and Retrieval of Text, *Science*, Vol. 267, February 1995, 843 - 848.
3. Hull, D., Grefenstette, G. (1996) Querying Across Languages: A Dictionary-Based Approach to Multilingual Information Retrieval. *Proc. ACM SIGIR*, Zürich, Switzerland, 1996, 49 - 57.
4. Peters, C. (2002) Cross Language Evaluation Forum. [http://clef.iei.pi.cnr.it]
5. Pfeifer, U., Poersch, T., Fuhr, N. (1995) Searching Proper Names in Databases. *HIM*, 1995, 259 - 275.
6. Pirkola, A., Hedlund, T., Keskustalo, H., Järvelin, K. (2001) Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings. *Information Retrieval*, 4 (3/4), 2001, 209 - 230.
7. Pirkola, A., Keskustalo, H., Leppänen, E., Känsälä, A-P., Järvelin, K. (2002) Targeted s-Gram Matching: a Novel n-Gram Matching Technique for Cross- and Monolingual Word Form Variants. *Information Research*, 7 (2) 2002. [Available at http://InformationR.net/ir/7-2/paper126.html]
8. Pirkola, A., Toivonen, J., Keskustalo, H., Visala, K., Järvelin, K. (2003) Fuzzy Translation of Cross-Lingual Spelling Variants. Accepted for *ACM SIGIR* 2003.
9. Robertson, A.M., Willet, P. (1998) Applications of N-Grams in Textual Information Systems. *Journal of Documentation*, 1, 1998, 48 - 69.
10. Ullman, J.R. (1977) A Binary n-Gram Technique for Automatic Correction of Substitution, Deletion, Insertion and Reversal Errors in Words. *Computer Journal*, 2, 1977, 141 - 147.
11. Zobel, J., Dart, P. (1996) Phonetic String Matching: Lessons from Information Retrieval. *Proc. ACM SIGIR*, Zürich, Switzerland, August 1996, 166 - 173.