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

The RATF formula (Kwok's formula): exploiting average term frequency in cross-language retrieval

Ari Pirkola, Erkka Leppänen and Kalervo Järvelin
Department of Information Studies
University of Tampere
Finland

Abstract

In an earlier study, we presented a query key goodness scheme, which can be used to separate between good and bad query keys. The scheme is based on the relative average term frequency (RATF) values of query keys. In the present paper, we tested the effectiveness of the scheme in Finnish to English cross-language retrieval in several experiments. Query keys were weighted and queries were reduced based on the RATF values of keys. The tests were carried out in TREC and CLEF document collections using the InQuery retrieval system. The TREC tests indicated that the best RATF-based queries delivered substantial and statistically significant performance improvements, and performed as well as syn-structured queries shown to be effective in many CLIR studies. The CLEF tests indicated the limitations of the use of RATF in CLIR. However, the best RATF-based queries performed better than baseline queries also in the CLEF collection.

Introduction

Standard best-match retrieval systems are based on the tf.idf weighting scheme (e.g., Robertson, *et al.*, 1995; Salton, 1989; Singhal, *et al.*, 1996; Turtle, 1990). The basic idea of tf.idf weighting is that the more occurrences a query key has in a document and the fewer there are documents where the key occurs, the more likely the document is to be relevant with respect to the query. However, tf.idf does not necessarily indicate the goodness of keys, e.g., in some relevant documents an important key may occur just once, though it generally has high tf in documents.

In an earlier study we demonstrated that the use of *average term frequencies* of keys as key weights contributes to better retrieval performance in monolingual IR (Pirkola and Järvelin, 2001b). We developed *a query key goodness scheme* that calculates goodness weights for keys on the basis of cf/df (average term frequency) and df statistics of a document collection. Cf stands for *collection frequency* and refers to the number of occurrences of a key in a collection, and df stands for *document frequency* and refers to the number of documents in which the key occurs. The key goodness scheme (called RATF formula, where RATF refers to *relative average term frequency*) was tested through several experiments in monolingual retrieval in a TREC collection. The highest ranked keys by the RATF scheme were weighted higher than other keys structurally and using the RATF values as query key and subquery weights. The tests indicated that RATF-based key weighting as well as RATF-based query structuring delivered substantial and statistically significant performance improvements.

In the title of this paper the term *Kwok's formula* is used as a synonym to the term *RATF formula.* This is because the basic idea of the RATF formula is the same as that behind the query key weighting formula presented by Kwok in 1996: important keys often have high average term frequency and low document frequency - therefore it is often useful to weight high such keys in queries. We developed the RATF formula independently based on our findings in Pirkola and Järvelin (2001a), which showed that typically 1-2 query keys have far higher cf/df and far lower df than the other keys of a query. Kwok (1996) used his scheme for query key weighting in monolingual retrieval and was able to show clear improvements in retrieval performance owing to the use of his formula. In this study we will test

whether the RATF formula is useful in *cross-language information retrieval*.

Cross-language information retrieval (CLIR) refers to an information retrieval task where the language of queries is other than that of the retrieved documents. Different approaches to cross-language retrieval are discussed in Oard and Diekema (1998). A standard method in dictionary-based CLIR is to replace each source language key by all of its target language equivalents included in a translation dictionary (Pirkola, 1998; Pirkola, *et al*., 2001). Dictionaries typically give several translations for one source language word, and the number of mistranslated keys, i.e., the keys that have wrong meanings in the context of the topic, in a CLIR query (the final translated query) is usually high. We argued that the RATF formula could be exploited in the same way in cross-language retrieval as in monolingual retrieval to improve query performance. We thus assumed that by applying RATF values as key and subquery weights CLIR queries will perform well despite the abundance of mistranslated and other bad keys. This is because many of the bad keys are general words whose RATFs are low, and when RATF values are used as key weights bad keys are downweighted with respect to the more specific important keys which typically have high RATFs.

This approach as a starting point, we will examine in this paper the utilization of the RATF formula in CLIR. We will test the effectiveness of the formula by using the same RATF weighting method as in Pirkola and Järvelin (2001b), where RATF values of keys as such were used key weights. We will also develop and test new RATF-based key weighting methods that particularly are suited for CLIR. We will also explore whether the removal of keys with low RATFs will improve retrieval performance. The RATF-based CLIR queries are compared with *syn-queries*. In many studies the syn-queries of the *Inquery retrieval system* have been demonstrated to perform very well in CLIR (Ballesteros and Croft, 1998; Gollins, 2000; Hedlund, *et al*., 2001a; Meng, *et al*., 2000; Pirkola, 1998; Pirkola, *et al*., 2000; Oard and Wang, 2001; Sperer and Oard, 2000).

The tests presented in this paper were performed using TREC and CLEF (Peters, 2000) document collections as test collections. The TREC collection contained 515,000 documents and the CLEF collection some 112,000 documents. As test requests we used 50 TREC and 50 CLEF topics. The title and description fields of the topics were translated by a professional translator into Finnish. The Finnish words were translated back to English by means of an automatic translation system. The automatically translated words were used as query keys in the CLIR queries. Several RATF-based weighting methods were tested. As a test system we used the Inquery retrieval system. We will demonstrate that query key weighting based on the RATF formula will yield substantial performance improvements in cross-language retrieval with respect to queries where no disambiguation method is applied. However, we will also show that there are restrictions in applying RATF in CLIR.

The remainder of this paper is structured as follows. Section 2 presents the RATF formula. Section 3 describes the methodology, and Section 4 the findings. Sections 5 and 6 contain the discussion and conclusions.

The RATF formula

The RATF scheme computes *relative average term frequency* values for keys. The scheme is defined as the function RATF as follows:

Let k denote some key of a collection and cf_k its collection frequency, and df_k its document frequency. Let SP be a collection dependent scaling parameter, and p the power parameter. The function RATF(k) gives the relative average term frequency of the key i.

$$RATF(k) = (cf_k / df_k) * 10^3 / ln(df_k + SP)^p$$

The RATF formula gives high values for the keys whose atf (i.e., cf/df) is high and df low. The scaling parameter SP is used to downweight rare words. In our training experiments, SP = 3000 and p =3, gave the best query performance. SP = 3000 was used in the TREC tests of this study. In the CLEF tests, the SP value of 800 was used based on the relative collection sizes of the TREC and CLEF collections. RATF values were assigned automatically to the words using a program produced for this study.

Table 1 presents an example of a RATF value set, showing RATF values for the words of the description field of the TREC topic 51 (the stop-words *to, or, a, between, and, over, the, of* as well as duplicates were first removed), as well as document frequencies and average term frequencies of the words. The description of the topic 51 is as follows:

- *Document will discuss government assistance to Airbus Industrie, or mention a trade dispute between Airbus and a United States aircraft producer over the issue of subsidies.*

Table 2 presents the RATF values for the retranslated words of the Topic 51. The words above (excluding stop-words) were first manually translated into Finnish and then translated back to English using an automatic translation system and a program that automatically assigns RATF values to words. (The dictionary gave 60 translations; in Table 2 only the words with the highest and lowest RATFs are presented).

In Tables 1 and 2, words are sorted by their RATF values. In both cases the word *Airbus* is ranked high. Note that its atf is high and df low (Table 1). If key weighting is solely based on the tf.idf weighting scheme, the word *mention* might have a strong influence on search results because of its low df, though it apparently is non-topical. The RATF formula, however, ranks it low because of its low atf.

The figures in Table 2 suggest that the RATF formula is effective in CLIR. The formula gives high values for the important topic words (*Airbus*, *subsidy*), and low values for mistranslated and other bad words. However, automatic translation provides some pecularities. The form *industrielle* is a Finnish inflectional form of the word *Industrie*. The morphological analyzer did not recognize the form, and it occurs in its Finnish form in the list of Table 2, as well as in the CLIR query. The word *interstice* is an example of a mistranslated word whose RATF is high.

Word	RATF	DF	ATF
airbus	3,74	663	2,07
subsidies	2,86	3063	1,89
industrie	2,39	262	1,27
trade	2,27	38039	2,73
aircraft	2,09	9719	1,76
document	1,62	17306	1,58
government	1,59	90080	2,39
assistance	1,53	13940	1,41
dispute	1,48	11359	1,30
states	1,39	139126	2,32
united	1,38	63172	1,89
mention	1,34	8731	1,11
producer	1,17	45584	1,47
issue	1,16	63640	1,58
discuss	0,91	48523	1,16

Table 1: Df, atf, and RATF values for the words of the TREC topic 51

Word	RATF	Word	RATF
industrielle	5,29	bargain	1,91
airbus	3,74	aid	1,87
engine	2,93	shop	1,86

subsidy	2,86	gap	1,85
interstice	2,36	merit	1,75
flight	2,32	originate	1,72
trade	2,27	machine	1,68
transaction	2,09	controversy	1,67
motor	2,05	distance	1,64
contrivance	2,04	...	-
talent	1,97	existence	1,11
apparatus	1,96	help	1,06
collaborate	1,95	come	1,02
virtue	1,92	develop	1,01
space	1,91	time	0,99
interval	1,91	relation	0,98

Table 2: RATF values for the re-translated words of the TREC topic 51

Methods and data

Test collections, requests, and query formulation

In this study we used the same *training request set* as in the monolingual retrieval tests in Pirkola and Järvelin (2001b). The training request set consisted of the TREC topics 51-75 and 101-125. In this study the training queries were used in the development of different RATF-based query types, and in determining the threshold RATFs.

The TREC test request set consisted of the TREC Topics 76-100 and 126-150. The test queries used in the tests of this paper were formed from the test requests. The TREC topics 101-150 are narrower in scope and have fewer relevant documents than the topics 51-100 (Harman, 1993). The average performance level of the queries 51-100 is higher than that of the queries 101-150. To get a representative sample of different types of topics (narrow and broad), the test request set (as well as the training request set) was formed from two different TREC topic sets with topics of different characteristics. The words of the (1) title and (2) title and description fields of the TREC request sets were used as query keys. The former queries are called *short queries* and the latter ones *long queries*.

The CLEF test request set contained CLEF 2001 topics (50 topics). The words of the title and description fields of the topics were used as query keys. The TREC and CLEF queries differed from each other, first, in that unlike TREC queries duplet keys were retained in the CLEF queries. Second, in the CLEF tests proper names and other words not contained in the translation dictionary were translated by an n-gram matching method. The method is described in detail in the paper by Pirkola and colleagues in this issue of *Information Research*.

A professional translator translated the TREC and CLEF topics (title and description fields) into Finnish according to the guidelines provided by CLEF (Peters, 2000). The Finnish words were retranslated to English using an automatic query translation and construction system developed in the Information Retrieval Laboratory at the University of Tampere (UTA). The different query versions tested in this study were formulated manually, however. The language and query processing components of the automatic query translation and construction system are described in detail in Hedlund, *et al*. (2001a) and Pirkola, *et al*. (2001).

Retrieval system and query operators

As a test system, the study used the *Inquery retrieval system* (Allan, *et al.*, 2000; Broglio, *et al.*, 1994). English keys and the words of documents were normalized using the morphological analyzer *Kstem*, which is part of Inquery. Inquery is based on Bayesian inference networks. All keys are attached with a *belief value*, which is calculated by the tf.idf modification of the system (Allan, *et al.*, 2000):

$$0.4 + 0.6 \times \left[\frac{tf_{ij}}{tf_{ij} + 0.5 + 1.5 \times \left[\frac{dl_j}{adl} \right]} \right] \times \left[\frac{\log \left[\frac{N + 0.5}{df_i} \right]}{\log(N + 1.0)} \right]$$

tf_{ij} = the frequency of the query key i in the document j
df_i = the number of documents which contain the query key i
dl_j = the length of the document j (in words)
adl = average document length in the collection (in words)
N = the number of documents in the collection

The value 0.4 is a default value given to a key not occurring in a document.

The Inquery query language provides a set of operators to specify relations between quey keys.

For the *sum*-operator, the system computes the average of key (or subquery) weights. The keys contained in the *weighted sum* (*wsum*) operator are weighted according to the weight value associated with each key (or subexpression). The final belief score, a weighted average, is scaled by the weight value associated with the *wsum* itself.

The *syn*-operator treats its operand search keys as instances of the same search key. For the keys linked by the syn-operator, an aggregate document frequency is computed instead of individual document frequencies for every key (Sperer and Oard, 2000).

The *uwn*-operator (unordered window) is a proximity operator. It only retrieves the documents, which contain the arguments of the operator within the defined window.

In queries, the operators are marked by the hash sign "#", e.g., #sum and #wsum. Commas are not needed between query keys or their weights. Therefore, #sum(south africa sanctions) and #wsum(1 2 south 2 africa 1 sanctions) are well-formed expressions in the InQuery query language.

Baseline and test queries

In this section, we will describe the baseline and test queries (CLIR queries) investigated in the study. The following notations will be used:

The translation equivalents (included in a dictionary) of source language keys A, B, C ... are the sets $\{a_1, ..., a_n\}$, $\{b_{1, ..., bm}\}$, $\{c_{1, ..., ck}\}$, ..., respectively. Similarly, the translation equivalents of the components of a source language compound *AB* are the sets $\{a_1, ..., a_n\}$ and $\{b_{1, ..., bm}\}$. In query language expressions below the translation equivalents are just enumerated, without intervening commas and braces, following the InQuery query language syntax.

The RATF values of the words (translation equivalents) $\{a_1, ..., a_n\}$, $\{b_1, ..., b_m\}$,..., are denoted by RATF(a_1), ...,

RATF(a_n), RATF(b_1), ..., RATF(b_m), ...

The average of the RATF values of the translation equivalents in an equivalent set, e.g., $S_A = \{a_1, ..., a_n\}$ is denoted by *avgRATF*(S_A). Its value is given by:

$$| S_A |^{-1} * S_{aeS_A} RATF(a)$$

AekvRATF($\{a_1, ..., a_n\}$) denotes key weighting where the average RATF of the words $\{a_1, ..., a_n\}$is reduced on the basis of the number of translation equivalents in an equivalent set (Section 3.3.2.). The rationale behind aekvRATF(S_A) is that that those keys (concepts) that have many translations in another language probably are harmful or less important than the keys that only have a few translations. Its value is given by:

aekvRATF(S_A) = *avgRATF*(S_A), if $|S_A| < c$

aekvRATF(S_A) = *avgRATF*(S_A) - 0.2*($|S_A|$-*c*), if $|S_A| > c$

Here *c* is a constant parameter setting the limit for excessive number of translations. It is collection dependent and derived experimentally. In our experiments we employed *c* = 3.

For example, the following 8 words appear in the same equivalent set (their RATF values are in the parentheses): *interstice* (2,36), *interval* (1,91), *space* (1,91), *gap* (1,85), *distance* (1,64), *stretch* (1,59), *time* (0,99), *relations* (0,98). For each word (and a syn-set containing these words), the avgRATF is 1,65. For each word (syn-set), aekvRATF is 1,65 - (5*0.20) = 0.65.

In the case of equivalent sets containing 1-3 equivalents, *aekvRATF* (S_A) is the same as *avgRATF*(S_A).

Baseline queries

We compared the CLIR queries that are presented in Section 3.3.2 with the following three types of baseline queries. All the baseline and test query types were run in the TREC collection. Undisambiguated (unstructured, unweighted) and syn-queries (types 2 and 3 below) and the best test query types based on the results of the TREC tests were run in the CLEF collection. Duplicates were not removed from the CLEF queries. Therefore the CLEF queries are marked by (QKF) in Table 10, where QKF refers to *query key frequency*. In fact, in the case of CLEF queries 'undisambiguated' queries were disambiguated, in part, through QKF.

(1) The original English queries contained as query keys the title words (short queries) and the title and description words (long queries) of the TREC Topics (76-100 and 126-150). The English queries were run in the study to show the performance level of the test queries. The original English queries were flat sum-queries:

#sum(A B ...)

(2) Undisambiguated CLIR queries were flat sum-queries:

#sum(a_1... a_n b_1... b_m ...)

The undisambiguated CLIR queries included the same query keys as the test queries, but no disambiguation method was applied. However, as mentioned above duplicate keys were retained in the CLEF queries. The use of query key frequency is a kind of disambiguation technique.

(3) Syn-queries

In *syn-based structuring* the translation equivalents of each source language key are grouped together by the *syn*-operator of the Inquery retrieval system. Query structuring using the syn-operator has been shown to be an effective disambiguation method in many CLIR studies (Ballesteros and Croft, 1998; Gollins, 2000; Hedlund, *et al.*, 2001a; Meng, *et al.*, 2000; Oard and Wang, 2001; Pirkola, 1998; Pirkola, *et al.*, 2000; Sperer and Oard, 2000). In many studies in which the effects of the syn-operator have been tested, the proximity operator (uwn) has been used

in syn-queries to handle the compound words; the translation equivalents that correspond to the first part of a source language compound are combined to the equivalents that correspond to the second part of the compound using the uwn-operator.

In earlier studies the contribution of the uwn-operator on the effectiveness of syn-queries have not been tested. We, however, tested the effects of the uwn-operator. The results are presented in a later paper. The findings showed that syn-queries performed better than any of the combined (i.e., syn + uwn) query types tested. Therefore in this study we will use the best syn-query type (i.e., syn-queries without uwn) as baseline for RATF-based queries.

The results of the syn/uwn tests are published later but shortly discussed here. An important point is that one has to make difference between the disambiguation effect of a proximity operation and phrase-based searching. The disambiguation effect of the uwn-operator refers to the fact that normally a combination of two mistranslated keys does not make any sense. Therefore, the proximity combination method applied for the translation equivalents probably has a clear disambiguation effect. On the other hand, many studies on monolingual retrieval which have used sophisticated linguistic analysis or statistical methods have shown that phrase-based searching does not improve retrieval performance, or that improvements are just small (Buckley, *et al.*, 1995; Mitra, *et al.*, 1997; Smeaton, 1998). It should be noted that this monolingual component is involved in CLIR. One should also note that there are different types compounds, which probably should be handled in different ways for improved CLIR performance. This issue is investigated by Hedlund and colleagues at UTA.

For a source language query containing the keys A, B, and C, the syn-query was formed as follows:

#sum(#syn(a_1... a_n) #syn(b_1... b_m) #syn(c_1... c_k))

For a source language query including a compound AB and a single word C, the syn-query is as follows:

#sum(#syn(a_1...a_n b_1...b_m) #syn(c_1... c_k))

Test queries (CLIR queries)

Reduced Queries

In the first test we examined whether the RATF scheme is helpful in recognizing *bad query keys*, i.e., keys which tend to lower query performance. Bad keys involve, particularly, mistranslated keys and marginal topical and performative words used in natural language queries (as well as in Description fields of TREC Topics). Based on the results in the training data, the keys of RATF < 1.4 (the first experiment) and aekvRATF < 0.8 (the second experiment) were removed from the undisambiguated CLIR queries (baseline 2).

Single Key Weighting

In the second test we investigated whether the RATF scheme applies for the weighting of single keys. The following four weighting methods were tested. (RATF values were multiplied by 100.)

RATF. Query keys were weighted by their RATF values.

- #wsum(100 100*RATF(a_1) a_1 ...100*RATF(a_n) a_n 100*RATF(b_1) b_1 ...
- 100*RATF(b_m) b_m ...)

A sample RATF-weighted query is demonstrated below. The query is formed on the basis of the title of the TREC Topic 52 (*South African Sanctions*). Table 3 shows the RATF values for these keys.

#wsum(100 382 africa 249 sanction 177 south)

Key	RATF
africa	3,82

sanction	2,49
south	1,77

Table 3: RATF values for the words of the title of the Topic 52

RATF/nil-parameter. In this experiment, the values of SP = 0 and p = 1 were used in the RATF formula. In other words, the formula was reduced to RATF(k) = (cf_k / df_k)/ln(df_k). Query keys were weighted by their RATF/nil-parameter values.

AvgRATF. Keys were weighted by their avgRATF values.

- #wsum(100 100*avgRATF(S_A) a_1... 100* avgRATF(S_A) a_n 100*avgRATF(S_B) b_1...100* avgRATF(S_B) b_m ...)

AekvRATF. Keys were weighted by their aekvRATF values.

- #wsum(100 100*aekvRATF(S_A) a_1... 100* aekvRATF(S_A) a_n 100*aekvRATF(S_B) b_1... 100* aekvRATF(S_B) b_m ...)

AekvRATF was calculated by reducing avgRATF values of keys by the RATF unit of 0.20 per equivalent after three equivalents (for equivalents contained in equivalent sets of more than three equivalents).

Syn-Set Weighting

In this test, syn-sets of syn-queries (baseline 3) were weighted using avgRATF and aekvRATF.

AvgRATF. Syn-sets were weighted by their avgRATF values.

- #wsum(100 100*avgRATF(S_A) #syn(a_1... a_n) 100*avgRATF(S_B) #syn(b_1... b_m) ...)

AekvRATF. Syn-sets were weighted by their aekvRATF values.

- #wsum(100 100*aekvRATF(S_A) #syn(a_1... a_n) 100*aekvRATF(S_B) #syn(b_1... b_m) ...)

Findings

The effectiveness of the test queries was evaluated as precision at 10% recall (Pr. at 10% R) and average precision over 10%-100% recall levels (avg. precision). The former is a user-oriented measure for high-precision queries while the latter is a system-oriented average performance measure.Statistical significance of the difference between the performance of the test queries and that of unstructured queries was tested using *Wilcoxon signed ranks test.* The Wilcoxon 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.

The TREC tests

Table 4 shows the results of the baseline TREC runs. Tables 5-9 show the results of the test TREC runs. In Tables 5-9 the performance of test queries is compared with that of undisambiguated queries.

The findings of query key removal are shown in Table 5. As can be seen, long RATF-reduced queries perform slightly better than unreduced (undisambiguated) queries, but in the case of short queries the removal of low-RATF keys actually results in performance drop. An obvious reason for this is that, though most keys with low RATFs apparently are harmful, some keys with low RATFs are important. AekvRATF seems to attack this problem, since for both query types the queries where the keys of aekvRATF < 0.8 were removed perform clearly better than the

baseline queries.

Tables 6 (long queries) and 7 (short queries) present the results of single key weighting. As shown in the tables, all the weighting methods give substantial performance improvements. For *long queries*, the relative improvement figures (avg. precision) due to single key weighting are 94.1% (RATF), 98.0% (avgRATF), and 123.5% (aekvRATF). In the case of precision at 10% recall, improvements are somewhat smaller (58.8% - 85.3%), but still substantial. RATF/non-parameter queries perform worse than the queries where key weighting is based on the basic RATF formula. The improvement potentials for *short RATF-based* queries are limited, since short undisambiguated queries perform very well. Nevertheless, for short queries, the improvements in average precision are clear, i.e., 7.0% - 15.0% (avg. precision) and 11.4% - 24.3% (Pr. at 10% R). In all cases, the most effective RATF-based weighting method is aekvRATF. Performance differences between aekvRATF-queries and syn-queries are small (Tables 6-7).

The results for the question whether the RATF-weighting of syn-sets of syn-queries will improve performance are shown in Tables 8 (long queries) and 9 (short queries). As can be seen, in the case of long queries the RATF-based syn-queries perform slightly better than plain syn-queries. However, short avgRATF/syn-queries and short aekvRATF/syn-queries perform slightly worse than short plain syn-queries.

The CLEF tests

The CLEF results are presented in Table 10. As shown, single key weighting using RATF results in the decrease in retrieval performance. Probably, this is largely due to n-gram matching. N-gram matching typically gives one correct correspondent whose RATF is high and five false correspondents whose RATFs also are high (see the Discussion section). The correct correspondent of a source language proper name often occurred in the set of 1-10 best matching words in the ranked word list of n-gram matching. Therefore in our CLEF 2001 tests we selected the six best correspondents for the final query (Hedlund, *et al*., 2001b). The correspondents were combined by the syn-operator. This method was useful improving query performance, but in RATF-weighting seems to be harmful.

Single key weighting based on aekvRATF gives better performance than baseline. In the case of average precision performance improvements are clear (12.2%) but statistically insignificant. AekvRATF reduces the weights of false correspondents of n-gram matching. This is because the six best corespondents of n-gram matching were grouped together, and aekvRATF was computed for this key group of the six keys.

Syn-queries perform markedly better than the baseline queries, with the improvement percentages being 12,6% and 18,4%. Plain syn-queries perform better than syn/avgRATF and syn/aekvRATF queries. Thus, using avgRATF and aekvRATF as syn-set weights does not improve performance. The factors that affect the performance of syn-queries are discussed in Pirkola, *et al*. (2001). In addition, the performace of syn-queries depends on query length, as shown in this study and in Sperer and Oard (2000).

Discussion

One of the main problems associated with dictionary-based CLIR is *translation ambiguity*, which refers to the abundance of mistranslated keys in CLIR queries. The techniques to handle translation ambiguity involve corpus-based query expansion to reduce the effects of mistranslated and other bad keys (Ballesteros and Croft, 1997; Chen, *et al*., 1999), the use of word co-occcurrence statistics for selecting the best or correct translations (Chen, *et al*., 1999), the selection of translation equivalents on the basis of aligned sentences (Davis, 1996; Davis and Dunning, 1995), the selection of translation equivalents on the basis of word frequencies in the target-corpus (Kwok, 2000) and query structuring using Inquery's syn-operator (Pirkola, 1998). Syn-based structuring alters (relatively) tf.idf weights of keys in a query:

- In unstructured queries mistranslated keys with low document frequency may ruin query performance. In structured queries in syn-sets they are downweighted because of the aggregate document frequency (Sperer and Oard 2000)
- Important keys often have 1-2 translations only, and have relatively more weight in structured than in unstructured CLIR queries.

Kwok (2000) used a similar kind of method in which the keys in a translation equivalent set were considered as synonyms. For the equivalent set an aggregate *collection term frequency* was computed. The researcher

demonstrated that the method is useful in English to Chinese CLIR.

In this study we tested the use of the RATF formula in CLIR. The formula calculates goodness values for query keys on the basis of document frequency and collection frequency statistics of words in a document collection. The rationale behind using the RATF formula in CLIR is that many of the bad translation equivalents are common words whose RATFs are low, whereas the actual topic words often are more specific words with their RATFs being higher.

The tests in the TREC collection showed that there are many effective RATF-based weighting methods. The best one was aekvRATF, which takes into account both the number of translation equivalents of a source language key and the RATF values of the equivalents. AekvRATF-queries performed as well as, or even somewhat better than syn-queries which have been reported to perform well for different language pairs. The fact that RATF-based queries, in particular aekvRATF-queries are effective in CLIR is significant in that document and collection frequencies often are standard records in retrieval system packages. This allows an easy integration of a RATF-type key goodness evaluation method into cross-language retrieval systems. The syn-operator (or that kind of operator) is not a standard operator. Therefore, syn-queries can be run only in some systems.

Nevertheless, the experiments in the CLEF collection showed that the utilization of RATF in CLIR has limitations. The CLEF experiments differed from the TREC experiments in three main points. The first one was a collection size. The size of the CLEF collection was around one fourth of that of the TREC collection (Section 3.1).

Second, duplet query keys were not removed from the CLEF queries. In other words, in the CLEF experiments the effectiveness of RATF-based queries was tested against that of queries in which query key frequencies were applied. In CLEF topics important topic words often have 1-3 occurrences (in the title and description fields). The results suggest that RATF-based weighting is not useful in queries where important keys are weighted through query key frequencies. RATF-based weighting and query key frequency weighting seem to be competitive methods.

Third, in the CLEF tests source query keys not found in the dictionary were translated by an n-gram matching technique. The six best matching keys were used in the final CLIR queries. Selecting several best matching words was necessary, because the correct key is often found in the word set of 1-10 best words in the ranked word list of n-gram matching. However, from the RATF weighting perspective the use of several best matching keys is harmful, since it disturbs query balance. For instance, in the query 43 the RATF formula gave the value of 5,64 for the important key *nino* (referring to *El Nino*) and correctly ranked it higher than the less important keys of the query 43, such as *effect* (1,77), *impression* (2,55), and *influence* (2,12). However, n-gram translation also gave false correspondents whose RATFs were high, e.g., *nio* (11,45) and *annino*(5,72).

In aekvRATF-queries, however, the effects of such kinds of keys are depressed. It should be noted that aekvRATF-queries performed better than baseline queries both in the TREC and CLEF tests.

The next step in the development of our automatic CLIR system at UTA is to develop a more effective n-gram translation technique. We have developed a collocation identification method (called RKA and presented in Pirkola and Järvelin, 2001b) that may be useful in separating the correct correspondents of proper names and other untranslatable words from false correspondents. The preliminary tests have been encouraging. It is possible that n-gram matching together with RKA will effectively recognize the correct correspondents. This in turn may allow an effective use of RATF in CLIR also when n-gram based translation is applied. In particular, the effectiveness of aekvRATF can be expected to improve.

Conclusions

In Pirkola and Järvelin (2001b) we proposed a query key goodness scheme, which can be used to identify the best keys among the words of a natural language request. The scheme is based on the relative average term frequency (RATF) values of query keys. It gives high weights to words whose average term frequency is high and document frequency low (but not very low). The parameters for RATF calculation were learned through extensive tests using a training set of 50 requests and several parameter value combinations.

In this study the RATF formula was tested in cross-language retrieval. We conclude that *RATF as such* is useful in CLIR queries formed from such source language queries in which each key has one occurrence (such queries typically used, for example, in the Web). RATF as such is not useful in queries in which important keys are

weighted high using query key frequencies.

Neither is RATF useful in CLIR queries in which proper names are translated through n-gram matching. N-gram translation gives many bad words whose RATFs are high. However, it is possible to improve the effectiveness of n-gram translation. This in turn may allow an effective use of RATF also in the case of n-gram translation.

An important result is that *aekvRATF*, which takes into account the RATF values of keys and the number of translation equivalents of source keys was effective in all tests of this study. The importance of this finding is in that aekvRATF can be computed easily. It thus seems to apply for many types of cross-language retrieval systems.

Query type	Pr. at 10% R	Avg. Pr.
Longqueries		
OriginalEnglish	27,0	10,4
Undisambiguated (unweighted)	13,6	5,1
Syn	26,0	10,6
Short queries		
Original English	29,0	12,0
Undisambiguated (unweighted)	21,0	10,0
Syn	25,8	11,8

Table 4: The performance of the baseline TREC queries

Query type	Pr. at 10% R	Avg. Pr.
Long queries		
Undisambiguated - baseline	13,6	5,1
Removal of keys, threshold RATF 1,4	16,5	6,8
Change %	+21,3	+33,3
Statistical sign. level	0,01	0,001
Removalof keys, threshold aekvRATF 0,8	19,2	7,8
Change %	+41,2	+52,9
Statistical sign. level	0,001	0,001
Short queries		
Undisambiguated - baseline	21,0	10,0
Removal of keys, threshold RATF 1,4	19,9	9,5
Change %	-5,2	-5,0
Statistical sign. level	-	-
Removal of keys, threshold aekvRATF 0,8	23,3	10,8
Change %	+11,0	+8,0
Statistical sign. level	-	-

Table 5: The performance of reduced queries (TREC)

Query type	Pr. at 10% R	Avg. Pr.
Undisambiguated - baseline	13,6	5,1
Original English	27,0	10,4
Syn	26,0	10,6
Single key weighting, RATF	22,1	9,9
Change %	+62,5	+94,1
Statistical sign. level	0,01	0,001
Single key weighting, RATF/nil-parameter	20,1	9,5
Change %	+47,8	+86,3
Statistical sign. level	0,01	0,001
Single key weighting, avgRATF	21,6	10,1
Change %	+58,8	+98,0
Statistical sign. level	0,001	0,001
Single key weighting, aekvRATF	25,2	11,4
Change %	+85,3	+123,5
Statistical sign. level	0,001	0,001

Table 6: The performance of single key weighted queries. Long TREC queries

Query type	Pr. at 10% R	Avg. Pr.
Undisambiguated-baseline	21,0	10,0
Original English	29,0	12,0
Syn	25,8	11,8
Single key weighting, RATF	23,4	10,7
Change %	+11,4	+7,0
Statistical sign. level	-	-
Single key weighting, RATF/nil-parameter	22,7	10,5
Change %	+8,1	+5,0
Statistical sign. level	-	-
Single key weighting, avgRATF	24,0	11,1
Change %	+14,3	+11,0
Statistical sign. level	-	-

Single key weighting, aekvRATF	26,1	11,5
Change %	+24,3	+15,0
Statistical sign. level	0,01	0,01

Table 7: The performance of single key weighted queries. Short TREC queries

Query type	Pr. at 10% R	Avg. Pr.
Undisambiguated-baseline	13,6	5,1
Original English	27,0	10,4
Syn	26,0	10,6
Syn-set weighting, avgRATF	26,1	11,8
Change %	+91,9	+131,4
Statistical sign. level	0,001	0,001
Syn-set weighting, aekvRATF	26,3	11,8
Change %	+93,4	+131,4
Statistical sign. level	0,001	0,001

Table 8: The performance of RATF-weighted syn-queries. Long TREC queries

Query type	Pr. at 10% R	Avg. Pr.
Undisambiguated-baseline	21,0	10,0
Original English	29,0	12,0
Syn	25,8	11,8
Syn-set weighting, avgRATF	25,5	11,6
Change %	+21,4	+16,0
Statistical sign. level	0,01	0,01
Syn-set weighting, aekvRATF	24,7	11,2
Change %	+17,6	+12,0
Statistical sign. level	-	0,01

Table 9: The performance of RATF-weighted syn-queries. Short TREC queries

Query type	Pr. at 10% R	Avg. Pr.
Query key frequency (QKF) - baseline	48,6	29,4
Syn (QKF)	54,7	34,8
Change %	+12,6	+18,4

Statistical sign. level	-	-
Single key weighting, RATF (QKF)	41,9	26,7
Change %	-13,8	-9,2
Statistical sign. level	-	-
Single key weighting, aekvRATF (QKF)	49,2	33,0
Change %	+1,2	+12,2
Statistical sign. level	-	-
Syn-set weighting, avgRATF (QKF)	49,8	32,1
Change %	+2,5	+9,2
Statistical sign. level	-	-
Syn-set weighting, aekvRATF (QKF)	49,6	31,8
Change %	+2,1	+8,2
Statistical sign. level	-	-

Table 10: The performance of CLEF queries

Acknowledgements

The *Inquery* search engine was provided by the Center for Intelligent Information Retrieval at the University of Massachusetts. 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

- Allan, J., Connell, M.E., Croft, W.B., Feng, F.-F, Fisher, D., and Li, X. (2000) "Inquery and TREC-9." *The Ninth Text REtrieval Conference (TREC-9)*, Gaithesburg, MD. Available at: http://trec.nist.gov/pubs/trec5/papers/umass-trec96.ps.gz [Accessed 21 January 2002]
- Ballesteros, L. and Croft, W.B. (1997) "Phrasal translation and query expansion techniques for cross-language information retrieval". *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Philadelphia, PA, pp. 84-91. New York, NY: Association for Computing Machinery.
- Ballesteros, L. and Croft, W.B. (1998). "Resolving ambiguity for cross-language retrieval". *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, pp. 64-71. New York, NY: Association for Computing Machinery.
- Broglio, J., Callan, J. and Croft, W.B. (1994). "Inquery system overview". *Proceedings of the TIPSTER Text Program (Phase I)*, pp. 47-67. San Francisco, CA: Morgan Kaufman Publishers Inc.
- Buckley, C., Singhal, A., Mitra, M. and Salton, G. (1995) "New retrieval approaches using SMART: TREC-4. *The Fourth Text REtrieval Conference (TREC-4)"*, Gaithesburg, MD. Available at: http://trec.nist.gov/pubs/trec4/papers/Cornell_trec4.ps.gz [Accessed 21 January 2002]
- Conover, W.J. (1980) *Practical non-parametric statistics*. New York: John Wiley & Sons.
- Davis, M. and Dunning, T. (1995). "A TREC evaluation of query translation methods for multi-lingual text retrieval. *TheFourth Text REtrieval Conference (TREC-4)"*, Gaithesburg, MD. Available from:

http://trec.nist.gov/pubs/trec4/papers/nmsu.ps.gz [Accessed 20 January 2002]
- Davis, M. (1996). "New experiments in cross-language text retrieval at NMSU's Computing Research Lab." *The Fifth Text REtrieval Conference (TREC-5)*, Gaithesburg, MD. Available from: http://trec.nist.gov/pubs/trec5/papers/nmsu.davis.paper.ps.gz [Accessed 21 January 2002]
- Gollins, T.J. (2000). *Dictionary based transitive cross-language information retrieval using lexical triangulation*. Sheffield: University of Sheffield. (Master of Science Thesis).
- Harman, D. (1993). "Overview of the Second Text REtrieval Conference (TREC-2)." *The Second Text REtrieval Conference (TREC-2)*, Gaithesburg, MD. Available at: http://trec.nist.gov/pubs/trec2/papers/txt/01.txt [Accessed 21 January 2002]
- Hedlund T, Keskustalo H, Pirkola A, Sepponen M & Järvelin K, (2001a). "Bilingual tests with Swedish, Finnish and German queries: dealing with morphology, compound words and query structure". In: Carol Peters, ed. *Cross-Language Information Retrieval and Evaluation: Proceedings of the CLEF 2000 Workshop, Lecture Notes in Computer Science*, 2069, pp. 211-225. Heidelberg: Springer.
- Hedlund, T., Keskustalo, H., Pirkola, A., Airio, E., and Järvelin, K. (2001b). "UTACLIR @ CLEF 2001." *Working Notes for CLEF 2001 Workshop*. Available at: http://www.ercim.org/publication/ws-proceedings/CLEF2/hedlund.pdf
- Kwok, K.L. (1996). "A new method of weighting query terms for ad-hoc retrieval." *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Switzerland, pp. 187-195. New York, NY: Association for Computing Machinery.
- Kwok, K.L. (2000). Exploiting a Chinese-English bilingual wordlist for English-Chinese cross language information retrieval. *Proceedings of the 5th International Workshop on Information Retrieval with Asian languages, IRAL2000*, pp. 173-179.
- Meng, H., Chen, B., Grams, E., Khudanpur, S., Lo, W-K., Levow, G-A, Oard, D., Schone, B., Tang, K., Wang, H-M., and Wang, J.Q. (2000). "Mandarin-English Information (MEI): Investigating Translingual Speech Retrieval". *HLT 2001, Human Language Technology Confererence*, March 18-21, 2001, San Diego, California. Available at http://hlt2001.org/papers/hlt2001-50.pdf [Accessed 21 January 2002]
- Mitra, M., Buckley, C., Singhal, A. and Cardie, C. (1997). "An analysis of statistical and syntactic phrases". *Proceedings of RIAO'97, Computer Assisted Information Searching on the Internet*, Montreal, Canada, pp., 200-214.
- Oard, D. and Wang, J. (2001). "NTCIR-2 experiments at Maryland: Comparing structured queries and balanced translation". *The Second NTCIR Workshop*, March 7-9, Tokyo, Japan. Available at http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings2/jianqiang.pdf [Accessed 21 January 2002]
- Oard, D. and Diekema, A. (1998). "Cross-Language Information Retrieval". *Annual Review of Information Science and Technology (ARIST)*, **33**, 223-256.
- Peters, C. (2000). *CLEF - Cross-Language Evaluation Forum*. Available at http://galileo.iei.pi.cnr.it/DELOS/CLEF/clef.html [Accessed 21 January 2002]
- 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, pp. 55-63. New York, NY: Association for Computing Machinery.
- Pirkola, A., Hedlund, T., Keskustalo, H., Järvelin, K. (2000). "Cross-lingual Information Retrieval Problems: Methods and findings for three language pairs". In: Irene Wormell, ed. *ProLISSa Progress in Library and Information Science in Southern Africa. First biannual DISSAnet Conference*. Pretoria, 26-27 October 2000. Pretoria : Centre for Information Development, University of Pretoria
- Pirkola, A. and Järvelin, K. 2001a. "Employing the resolution power of search keys". *Journal of the American Society for Information Science and Technology*, **52**(7), 575 -583.
- Pirkola, A. and Järvelin, K. (2001b). Exploiting average term frequency and word distribution statistics in text retrieval. Submitted to *ACM Transactions of Information Systems*.
- 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, S.E., Walker, S., Beaulieu, M.M., Gatford, M., and Payne, A. (1995). "Okapi at TREC-4<" *The Fourth Text REtrieval Conference (TREC-4)*, Gaithesburg, MD. Available at: http://trec.nist.gov/pubs/trec4/papers/city.ps.gz [Accessed 20 January 2002]
- Salton, G. (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Reading, MA: Addison-Wesley.
- Singhal, A., Buckley, C. and Mitra, M. (1996). Pivoted document length normalization. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*,

Zurich, Switzerland, pp. 21-29. New York, NY: Association for Computing Machinery
- Smeaton, A.F. (1998). "User-chosen phrases in interactive query formulation for information retrieval". *Proceedings of the 20th BCS-IRSG Colloquium on IR Research*, Grenoble, France. Available at: ftp://ftp.compapp.dcu.ie/pub/w-papers/1998/CA0898.ps.Z [Accessed 20 January 2002]
- Sperer, R. and Oard, D. (2000) Structured translation for cross-language IR. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Athens, Greece, pp. 120-127. New York, NY: Association for Computing Machinery.
- Turtle, H.R. 1990. *Inference networks for document retrieval*. Amherst, MA: University of Massachusetts, Computer and Information Science Department. PhD Dissertation. (COINS Technical Report 90-92)

How to cite this paper:

Pirkola, A, Leppänen, E & Järvelin, K (2002) "The RATF formula (Kwok's formula): exploiting average term frequency in cross-language retrieval" *Information Research*, **7**(2) [Available at http://InformationR.net/ir/7-2/paper127]

Last updated: 18th January, 2001

Check for citations, using Google Scholar

Contents **Home**