Report on CLEF-2001 Experiments: Effective Combined Query-Translation Approach*

Jacques Savoy

Institut interfacultaire d'informatique, Université de Neuchâtel,
Pierre-à-Mazel 7, 2000 Neuchâtel, Switzerland
Jacques.Savoy@unine.ch http://www.unine.ch/info/clef/

Abstract. In our first participation in CLEF retrieval tasks, the primary objective was to define a general stopword list for various European languages (namely, French, Italian, German and Spanish) and also to suggest simple and efficient stemming procedures for these languages. Our second aim was to suggest a combined approach that could facilitate effective access to multilingual collections.

1 Monolingual Indexing and Searching

Most European languages (including French, Italian, Spanish, German) share many of the same characteristics as does the language of Shakespeare (e.g., word boundaries marked in a conventional manner, variant word forms generated generally by adding suffixes to the end of roots, etc.). Any adaptation of indexing or searching strategies thus means the elaboration of general stopword lists and fast stemming procedures. Stopword lists contain non-significant words that are removed from a document or a request before the indexing process is begun. Stemming procedures try to remove inflectional and derivational suffixes in order to conflate word variants into the same stem or root.

This first chapter will deal with these issues and is organized as follows: Section 1.1 contains an overview of our five test collections while Section 1.2 describes our general approach to building stopword lists and stemmers to be used with languages other than English. Section 1.3 depicts the Okapi probabilistic model together with various vector-space models and also evaluates them using the five test collections written in five different languages (monolingual track).

1.1 Overview of the Test Collections

The corpora used in our experiments included newspapers such as the *Los Angeles Times*, *Le Monde* (French), *La Stampa* (Italian), *Der Spiegel*, *Frankfurter Rundschau* (German) together with various articles edited by news agencies such

* To appear in Peters, C. (ed.): "Cross-Language Information Retrieval and Evaluation." Lecture Notes in Computer Science. Springer-Verlag, Berlin Heidelberg New York (2002)"

as *EFE* (Spanish) and the Swiss news agency (available in French, German and Italian but without parallel translation). As shown in Table 1, these corpora are of various sizes, with the English, German and Spanish collections being twice the volume of the French and Italian sources. On the other hand, the mean number of distinct indexing terms per document is relatively similar across the corpora (around 130), while this number is a little bit higher for the English collection (167.33). From those original documents, during the indexing process we only retained the following logical sections in our automatic runs: <TITLE>, <HEADLINE>, <TEXT>, <LEAD>, <LEAD1>, <TX>, <LD>, <TI>, and <ST>, On the other hand, we conducted two experiments (indicated as manual runs), one on the French collection and another on the Italian corpora within which we retained the following tags: for the French collections: <DE>, <KW>, <TB>, <SUBJECTS>, <CHA1>, <NAMES>, <NOM1>, <NOTE>, <GENRE>, <ORT1>, <SU11>, <SU21>, <GO11>, <GO12>, <GO13>, <GO14>, <GO24>, <TI01>, <TI02>, <TI03>, <TI04>, <TI05>, <TI06>, <TI07>, <TI08>, <PEOPLE>, <TI09>, <SOT1>, <SYE1>, and <SYF1>; while for the Italian corpora, and for one experiment, we used the following tags: <DE>, <KW>, <TB>, <NAMES>, <ARGUMENTS>, <LOCATIONS>, <TABLE>, <PEOPLE>, <ORGANISATIONS>, and <NOTE>.

From topic descriptions, we automatically removed certain phrases such as "Relevant document report ...", "Find documents that give ...", "Trouver des documents qui parlent ...", "Sono valide le discussioni e le decisioni ...", "Relevante Dokumente berichten ..." or "Los documentos relevantes proporcionan información ...".

In order to evaluate our approaches, we used the SMART system as a test bed for implementing the Okapi probabilistic model [1] and other vector-space strategies. This year our experiments were conducted on an Intel Pentium III/600 (memory: 1 GB, swap: 2 GB, disk: 6 x 35 GB).

1.2 Stopword Lists and Stemming Procedures

In order to define general stopword lists, we knew that such lists were already available for the English [2] and French languages [3]. For the three other languages, we established general stopword lists by following the guidelines described in [2]. Firstly, we sorted all word forms appearing in our corpora according to frequency of occurrence and extracted the 200 most frequently occurring words. Secondly, we inspected this list to remove all numbers (e.g., "1994", "1"), plus all nouns and adjectives more or less directly related to the main subjects of the underlying collections. For example, the German word "Prozent" (ranking 69), the Italian noun "Italia" (ranking 87) or from the Spanish corpora the term "política" (ranking 131) was removed from the final list. From our point of view, such words can be useful as indexing terms in other circumstances. Thirdly, we included some non-information-bearing words, even if they did not appear in the first 200 most frequent words. For example, we added various personal or possessive pronouns (such as "meine", "my" in German), prepositions ("nello", "in the" in Italian), conjunctions ("où", "where" in French) or verbs ("estar",

Table 1. Test collection statistics

	English	French	Italian	German	Spanish
Size (in MB)	425 MB	243 MB	278 MB	527 MB	509 MB
# of documents	113,005	87,191	108,578	225,371	215,738
Number of distinct indexing terms / document					
Mean	167.33	140.48	129.91	129.26	120.25
Standard deviation	126.3	118.6	97.6	119.8	60.1
Median	138	102	92	96	107
Maximum	1,812	1,723	1,394	2,593	682
Minimum	2	3	1	1	5
Number of queries	47	49	47	49	49
Number of rel. items	856	1,212	1,246	2,130	2,694
Mean rel. items / request	18.21	24.73	26.51	43.47	54.97
Standard error	3.29	3.47	3.56	6.97	9.09
Median	10	17	18	27	26
Maximum	107	90	95	212	261
Minimum	1	1	2	1	1
With 5- rel. docs	18	10	9	4	4
With 10- rel. docs	28	15	16	13	10

"to be" in Spanish). Another debatable issue was the presence of homographs and to some extent, we had to make arbitrary decisions relative to their inclusion in stopword lists. For example, the French word "son" can be translated as "sound" or "his".

The resulting stopword lists thus contained a large number of pronouns, articles, prepositions and conjunctions. As in various English stopword lists, there were also some verbal forms ("sein", "to be" in German; "essere", "to be" in Italian; "sono", "I am" in Italian). In our experiments we used the stoplist provided by the SMART system (571 English words) along with our 217 French words, 431 Italian words, 294 German words and 272 Spanish terms (these stopword lists are available at `http://www.unine.ch/info/clef/`).

After removing high frequency words, as an indexing procedure we used a stemming algorithm that tries to conflate word variants into the same stem or root. In developing this procedure for the French, Italian, German and Spanish languages, it is important to remember that these languages have more complex morphologies than does the English language [4]. As a first stage we decided to remove only inflectional suffixes such that singular and plural word forms or feminine and masculine forms conflate to the same root. More sophisticated schemes have already been proposed for the removal of derivational suffixes (e.g., "-ize", "-ably", "-ship" in the English language), such as the stemmer developed by Lovins [5], based on a list of over 260 suffixes, while that of Porter [6] looks for about 60 suffixes. For the Spanish language for example, Figuerola [7] described two different ones and their experiments showed that removing only inflectional

suffixes (88 different inflectional suffixes were defined) seemed to provide better retrieval levels, compared with removing both inflectional and derivational suffixes (this extended stemmer included 230 suffixes).

A "quick and efficient" stemming procedure had already been developed for the French language [3]. Based on this same concept, we implemented a stemming algorithm for the Italian, Spanish and German languages (the C code for these stemmers can be found at `http://www.unine.ch/info/clef/`). In our approach, we only tried to remove inflectional suffixes attached to nouns or adjectives. In this context, the main inflectional rule in Italian is to modify the final character (e.g., "-o", "-a" or "-e") into another (e.g., "-i", "-e"). As a second rule, Italian morphology may also alter the final two letters (e.g., "-io" in "-o", "-co" in "-chi", "-ga" in "-ghe"). In Spanish, the main inflectional rule is to add one or two characters to denote the plural form of nouns or adjectives (e.g., "-s", "-es" like in "amigo" and "amigos" (friend) or "rey" and "reyes" (king)) or to modify the final character (e.g., "-z" in "-ces" in "voz" and "voces" (voice)). In German, a few rules may be applied to obtain the plural form of words (e.g., "Sängerin" into "Sängerinnen" (singer), "Boot" into "Boote" (boat), "Gott" into "Götter" (god)). However, our suggested algorithms cannot handle person and tense variations found in verbs or other derivational constructions.

Finally, most European languages contain other morphological characteristics that our approach does not consider, with just one example being compound word constructions (e.g., handgun, worldwide). In German, compound words are widely used and hence causes many more difficulties than in English. For example, a life insurance company employee would be "Lebensversicherungsgesellschaftsangestellter" (Leben + s + versicherung + s + gesellschaft + s + angestellter for life + insurance + company + employee). Also morphological markers ("s") are not always present (e.g., "Bankangestelltenlohn" built as Bank + angestellten + lohn (salary)). According to Monz & de Rijke [8] or Chen [9], including both compounds and their composite parts (only noun-noun decompositions in [8]) in queries and documents can provide better performance. However, according to Molina-Salgado [10], decomposition of German words causes the average precision to be reduced.

Finally, diacritic characters are usually not present in English collections (with some exceptions, such as "à la carte" or "résumé"); and these characters are replaced by their corresponding non-accentuate letter in the Italian, German and Spanish language.

Given that French, Italian and Spanish morphology is comparable to that of English, we decided to index French, Italian and Spanish documents based on word stems. For the German language and its more complex compounding morphology, we decided to use a 5-gram approach [11], [12]. However, and contrary to [11], our generation of 5-grams indexing terms does not span word boundaries. This value of 5 was chosen for two reasons; it results in better performance when using the CLEF-2000 corpora [13], and it is also close to the mean word length of our German corpora (mean word length: 5.87; standard error: 3.7). Using this indexing scheme, the compound "das Hausdach" (the roof of the house) will

generate the following indexing terms: "das", "hausd", "ausda", "usdac" and "sdach".

1.3 Indexing and Searching Strategy

In order to obtain a broader view of the relative merit of various retrieval models [14], we first adopted a binary indexing scheme where each document (or request) is represented by a set of keywords without any weights. To measure the similarity between documents and requests we counted the number of common terms, computed from the inner product (retrieval model denoted "doc=bnn, query=bnn" or "bnn-bnn"). Binary logical restrictions are however often too limiting for document and query indexing. In order to weight the presence of each indexing term in a document surrogate (or in a query), we might take the term occurrence frequency into account, thus providing better term distinction and increasing indexing flexibility (retrieval model notation: "doc=nnn, query=nnn" or "nnn-nnn").

Those terms that do occur very frequently in the collection are not however believed to be very helpful in discriminating between relevant and non-relevant items. Thus we might count their frequency in the collection, or more precisely the inverse document frequency (denoted by idf), resulting in larger weights for sparse words and smaller weights for more frequent ones. Moreover, a cosine normalization could prove beneficial and each indexing weight might vary within the range of 0 to 1 (retrieval model notation: "ntc-ntc", Table 2 depicts the exact weighting formulation).

Table 2. Weighting schemes

bnn	$w_{ij} = 1$	npn	$w_{ij} = tf_{ij} \cdot ln\left[\frac{n - df_j}{df_j}\right]$
nnn	$w_{ij} = tf_{ij}$		
ntc	$w_{ij} = \frac{tf_{ij} \cdot idf_j}{\sqrt{\sum_{k=1}^{t}(tf_{ik} \cdot idf_k)^2}}$	atn	$w_{ij} = idf_j \cdot \left[\frac{0.5 + 0.5 \cdot tf_{ij}}{max\ tf_{i.}}\right]$
Okapi	$w_{ij} = \frac{(k_1+1) \cdot tf_{ij}}{K + tf_{ij}}$ with $K = k_1 \cdot \left[(1-b) + b \cdot \frac{l_i}{advl}\right]$		
dtu	$w_{ij} = \frac{(ln(ln(tf_{ij})+1)+1) \cdot idf_j}{(1-slope) \cdot pivot + (slope \cdot nt_i)}$		
dtc	$w_{ij} = \frac{(ln(ln(tf_{ij})+1)+1) \cdot idf_j}{\sqrt{\sum_{k=1}^{t}[(ln(ln(tf_{ik})+1)+1) \cdot idf_k]^2}}$		
Lnu	$w_{ij} = \frac{\frac{ln(tf_{ij})+1}{pivot + 1}}{(1-slope) \cdot pivot + (slope \cdot nt_i)}$		

Other variants could also be created, especially in situations when the occurrence of a given term in a document is a rare event. Thus, it may be a good

practice to give more importance to the first occurrence of this word as compared to any successive, repeated occurrences. Therefore, the tf component may be computed as 0.5 + 0.5 [tf / max tf in a document] (retrieval model denoted "doc=atn").

Finally we should assume that a term's presence in a shorter document provides stronger evidence than in a longer document. To account for this, we integrated document length within the weighting formula, leading to more complex IR models; those denoted for example by "doc=Lnu" [15] and "doc=dtu" [16]. Finally for CLEF-2001, we also conducted various experiments using the Okapi probabilistic model [1]. In our experiments, the constants b, k_1, advl, pivot and slope shown in Table 2 are fixed at b = 0.75, k_1 = 1.2, advl = 900, pivot = 125, and slope = 0.1. To measure the length of document i, we used the notation l_i corresponding to the sum of tf_{ij}.

Table 3. Average precision of various indexing and searching strategies based on monolingual requests and documents

	Average precision				
Title only	English 47 queries	French 49 queries	Italian 47 queries	German 49 queries	Spanish 49 queries
Okapi–npn	**48.50**	**43.79**	**39.60**	**32.62**	**48.87**
Lnu–ltc	44.36	40.35	38.18	29.23	45.13
atn–ntc	44.30	40.99	36.81	31.56	45.38
dtu–dtc	46.47	41.88	39.00	31.22	45.68
ntc–ntc	23.65	28.76	26.32	24.31	32.90
bnn–bnn	22.98	21.90	23.31	20.95	25.48
nnn nnn	13.33	16.00	19.04	11.62	21.71
Title-Desc					
Okapi–npn	**54.17**	**49.88**	**45.88**	**39.51**	**54.71**
Lnu–ltc	51.05	47.43	43.60	36.71	51.37
atn–ntc	51.09	47.97	41.62	37.54	51.31
dtu–dtc	53.26	48.97	43.49	36.72	50.59
ntc–ntc	31.25	32.21	30.01	30.08	36.83
bnn–bnn	25.51	17.91	25.66	18.79	28.68
nnn–nnn	12.06	14.13	20.78	9.83	24.74
Title-Desc-Narr					
Okapi–npn	**58.13**	**51.16**	**48.92**	**42.71**	**55.85**
Lnu–ltc	57.35	50.43	47.21	40.07	52.28
atn–ntc	54.52	50.78	45.21	39.26	54.82
dtu dtc	54.49	51.49	46.47	36.79	51.90
ntc–ntc	36.13	36.69	32.74	31.10	40.17
bnn–bnn	20.36	11.71	19.90	5.75	21.86
nnn–nnn	13.28	16.58	22.52	5.38	25.10

To evaluate the retrieval performance of these various IR models, we adopted

non-interpolated average precision (computed on the basis of 1,000 retrieved items per request by the TREC-EVAL program), thus allowing a single number to represent both precision and recall. Our evaluation results in Table 3 show that the Okapi probabilistic model provides the best performance when considering five different languages and three different query formulations. In the second position, we cannot see any clear distinction between three vector-space models, namely "doc=Lnu, query=ltc", "doc=atn, query=ntc" or "doc=dtu, query=dtc". For example for the French corpus, the second best approach is always "doc=dtu, query=dtc". For the Spanish collection however, the IR model "doc=dtu, query=dtc" reveals the second best performance when using a query based on only the Title section, the "doc=Lnu, query=ltc" when using a query based on Title and Descriptive logical sections and "doc=atn, query=ntc" when using the longest query formulation. Finally, the traditional tf-idf weighting scheme ("ntc-ntc") does not provide very satisfactory results, and the simple term-frequency weighting scheme ("nnn-nnn") or the simple coordinate match ("bnn-bnn") results in poor retrieval performance.

Table 4. Average precision using blind query expansion

	Average precision				
Title-Desc	English 47 queries	French 49 queries	Italian 47 queries	German 49 queries	Spanish 49 queries
Okapi–npn	54.17	49.88	45.88	39.51	54.71
10 terms / 5 docs	**54.81**	**50.21**	48.65	41.36	**58.00**
15 terms / 5 docs	52.81	49.91	48.85	41.87	57.85
20 terms / 5 docs	52.18	48.70	48.79	**42.29**	57.59
10 terms / 10 docs	51.91	50.00	48.54	40.99	57.17
15 terms / 10 docs	51.39	49.86	48.86	41.42	57.41
20 terms / 10 docs	50.27	49.28	**49.25**	41.81	57.24

It has been observed that pseudo-relevance feedback (blind expansion) seems to be a useful technique for enhancing retrieval effectiveness. In this study, we adopted Rocchio's approach [15] with $\alpha = 0.75$, $\beta = 0.75$ where the system was allowed to add to the original query m terms extracted from the n best ranked documents. To evaluate this proposition, we used the Okapi probabilistic model and enlarged the query by 10 to 20 terms, provided by the best 5 or 10 articles retrieved. The results depicted in Table 4 indicate that the optimal parameter setting seems to be collection-dependant, with a slight preference for extracting 10 terms from the best 5 ranked documents. Moreover, performance improvement also seems to be collection-dependant (or language-dependant), with an increase of only of 1.18% for the English corpus (average precision increased from 54.17 to 54.81) while for the Spanish language, enhancement is around 6% (average precision increased from 54.71 to 58.00).

In the monolingual track, we submitted six runs along with their corresponding descriptions, as listed in Table 5. Four of them were fully automatic using the request's Title and Descriptive logical sections, while the last two used more of the document's logical sections and were based on the request's Title, Descriptive and Narrative sections. These last two runs were labeled "manual" because we used logical sections containing manually assigned index terms. For all runs, we did not use any manual interventions during the indexing and retrieval procedures.

Table 5. Official monolingual run descriptions

Run name	Language	Query	Form	Query expansion	Average pr.
UniNEmofr	French	T-D	automatic	10 terms / 5 docs	50.21
UniNEmoit	Italian	T-D	automatic	10 terms / 5 docs	48.65
UniNEmoge	German	T-D	automatic	30 terms / 5 docs	43.09
UniNEmoes	Spanish	T-D	automatic	10 terms / 5 docs	58.00
UniNEmofrM	French	T-D-N	manual	no expansion	51.88
UniNEmoitM	Italian	T-D-N	manual	10 terms / 5 docs	54.18

2 Multilingual Information Retrieval

In order to overcome language barriers [17], [18], [19], we based our approach on free and readily available translation resources that automatically provide translations of queries in the desired target language. More precisely, the original queries were written in English and we did not use any parallel or aligned corpora to derive statistically or semantically related words in the target language. The first section of this chapter describes our combined strategy for cross-lingual retrieval while Section 2.2 provides some examples of translation errors. Finally, Section 2.3 presents different merging strategies along with their evaluations (multilingual track).

2.1 Query Translation

In order to develop a fully automatic approach, we chose to translate requests using the SYSTRAN system [20] (available at `http://babel.altavista.com`) and to translate query terms word-by-word using the BABYLON bilingual dictionary (available at `http://www.babylon.com`). The bilingual dictionary is able to provide not only one but several options for the translation of each word [21]. In our experiments, we decide to pick the first translation available (listed under "BABYLON 1") or the first two terms (listed under "BABYLON 2").

In order to obtain a quantitative picture of term ambiguity, we analyzed the number of translation alternatives generated by BABYLON's bilingual dictionaries. This study did not take determinants into account (e.g., "the"), conjunctions and prepositions (e.g., "and", "in", "of") or words appearing in our English stopword list (e.g., "new", "use"), and terms generally having a larger number of translations. Based on the Title section of the English requests, we found 137 search keywords to be translated.

The data in Table 6 shows how the mean number of translations provided by BABYLON dictionaries can vary depending on language, for example from 2.94 for German to 5.64 for Spanish. We found the maximum number of translation alternatives for the word "fall" in French and German (the word "fall" can be viewed as a noun or a verb), for the term "court" in Italian and for the word "attacks" in Spanish. The median value of their distributions is rather small, varying from 2 for German to 4 for Spanish. Thus for the first two translation alternatives, we covered around 54% of the keywords to be translated in German, 40.9% in French, 42.3% in Italian and 36.5% for Spanish.

Table 6. Number of translations provided by the BABYLON system for English keywords appearing in the Title section of our queries

	Number of translation alternatives			
	French	Italian	German	Spanish
Mean number of translations	3.63	5.48	2.94	5.64
Standard deviation	3.15	5.48	2.41	5.69
Median	3	3	2	4
Maximum	17	19	12	24
with word	"fall"	"court"	"fall"	"attacks"
No translation	8	9	9	8
Only one alternative	27	36	40	28
Two alternatives	21	13	25	14
Three alternatives	31	15	21	15

In order to improve search performance, we tried combining the SYSTRAN system's machine translation with a bilingual dictionary. In this case, we translated a query using the SYSTRAN system and for each English search term we added the first or the first two translated words obtained from a bilingual dictionary look-up.

Table 7 provides an overview of the relative performance of our three automatic query translation approaches, depicting average precision achieved by manually translated queries (labeled "monolingual") in Column 2. Column 3 lists the retrieval performance achieved by the machine translation SYSTRAN system and Column 4 the mean precision obtained using only the first translation candidate provided by BABYLON's bilingual dictionary. Column 5 accounts for the first two translations alternatives provided by the bilingual dictionary,

Table 7. Average precision using different query translation strategies (Title-Desc)

	Average precision				
French	monolingual	SYSTRAN	BABYLON 1	BABYLON 2	combined
Okapi–npn	49.88	44.79	35.06	31.07	**48.62**
Lnu–ltc	47.43	42.74	34.81	34.32	**45.83**
atn–ntc	47.97	41.52	29.50	27.17	**45.65**
dtu–dtc	48.97	43.01	29.47	28.67	**46.39**
ntc–ntc	32.21	27.92	24.49	24.10	**31.04**
Mean difference		-11.81%	-31.72%	-35.07%	-3.93%
Italian					
Okapi–npn	45.88	33.10	31.41	27.11	**37.00**
Lnu–ltc	43.60	31.94	33.28	28.95	**38.27**
atn–ntc	41.62	28.44	28.65	25.26	**33.29**
dtu–dtc	43.49	29.93	33.58	29.90	**37.37**
ntc–ntc	30.01	23.26	24.11	22.98	**27.64**
Mean difference		-27.99%	-25.76%	-33.70%	-14.71%
German					
Okapi–npn	39.51	29.64	27.74	27.86	**35.06**
Lnu–ltc	36.71	25.80	25.61	28.49	**32.75**
atn–ntc	37.54	25.96	25.39	23.53	**31.47**
dtu–dtc	36.72	27.24	25.72	27.12	**31.53**
ntc–ntc	30.08	23.46	19.93	20.07	**27.65**
Mean difference		-26.67%	-31.22%	-29.72%	-12.09%
Spanish					
Okapi–npn	54.71	41.56	35.94	32.59	**45.77**
Lnu–ltc	51.37	39.92	36.51	34.13	**43.49**
atn–ntc	51.31	37.25	35.65	30.49	**43.39**
dtu–dtc	50.59	38.03	36.86	31.98	**44.09**
ntc–ntc	36.83	25.99	24.90	24.54	**29.81**
Mean difference		-25.60%	-30.66%	-36.94%	-15.80%

and finally Column 6 shows our combined approach, where a query is translated automatically by the machine translation system and the first translation candidate for each search keyword is added to the translated request.

For each language, Table 7 lists the mean difference between manually translated queries and our various automatic translation strategies. These values indicate that the manual approach always performs better than the four automatic schemes, while the machine translation approach provides better retrieval performance when compared to the bilingual dictionary. For this latter approach, choosing only the first translation candidate seems to provide better results than choosing the first two. As shown in the last column, the retrieval effectiveness of our combined translation strategy usually provides the best automatic performance. However, the average difference between the manual translation approach and our combined scheme is usually around 14%, except for the French collec-

tion, where the difference is only 3.43%. Moreover, for the French corpus and the Okapi model, the average precision for our combined solution is 48.62, only -2.5% below the retrieval performance of manually translated queries (average precision of 49.88).

2.2 Examples of Translation Failures

In order to obtain a preliminary picture of the difficulties underlying our automatic translation approaches, we analyzed some queries by comparing the translations produced by our two machine-based tools with those written by a humans being (see Table 8 for examples). As a first example, the title of Query #70 is "Death of Kim Il Sung" (in which the number "II" is written as the letter "i" followed by the letter "l"). This couple of letters "IL" is interpreted as the chemical symbol for illinium (chemical element #61 "found" by two University of Illinois researchers in 1926; a discovery not confirmed until the chemical element #61 was finally found in 1947, and named promethium). Moreover, the proper name "Sung" was interpreted as the past participle of the verb "to sing".

As another example, we analyzed Query #54 "Final four results" translated as "demi-finales" in French or "Halbfinale" in German. This request resulted in the incorrect identification of a multi-word concept (namely "final four") both by our two automatic translation tools and by the manual translation provided in Italian and Spanish (where a more appropriate translation might be "semifinali" in Italian or "semifinales" in Spanish).

In Query #48 "Peace-keeping forces in Bosnia" or in Query #57 "Tainted-blood trial", our automatic system was unable to decipher compound word constructions using the "-" symbol and thus failed to translate the term "peace-keeping" or "tainted-blood".

In Query #74 "Inauguration of Channel Tunnel", the term "Channel Tunnel" was translated into French as "Eurotunnel". In the Spanish news test there were various translations for this proper name, including "Eurotúnel" (which appears in the manually translated request), as well as the term "Eurotunel" or "Eurotunnel".

2.3 Merging Strategies

Using our combined approach to automatically translate a query, we were able to search a document collection for a request written in English. However, this represents only the first stage our proposed cross-language information retrieval systems. We also needed to investigate situations where users write requests in English in order to retrieve pertinent documents in English, French, Italian, German and Spanish. To deal with this multi-language barrier, we divided our document sources according to language and thus formed five different collections. After searching in each corpora and the five result lists, they had to be merged so that users would be provided with a single list of retrieved articles.

Table 8. Examples of unsucessful query translations

C070 (query translations failed in French, Italian, German and Spanish)
<EN-TITLE> Death of Kim Il Sung
<FR-TITLE manually translated> Mort de Kim Il Sung <FR-TITLE SYSTRAN> La mort de Kim Il chantée <FR-TITLE BABYLON> mort de Kim Il chanter
<IT-TITLE manually translated> Morte di Kim Il Sung <IT-TITLE SYSTRAN> Morte di Kim Il cantata <IT-TITLE BABYLON> morte di Kim ilinio cantare
<GE-TITLE manually translated> Tod von Kim Il Sung <GE-TITLE SYSTRAN> Tod von Kim Il gesungen <GE-TITLE BABYLON> Tod von Kim Ilinium singen
<SP-TITLE manually translated> Muerte de Kim Il Sung <SP-TITLE SYSTRAN> Muerte de Kim Il cantada <SP-TITLE BABYLON> muerte de Kim ilinio cantar
C047 (both query translations failed in French)
<EN-TITLE> Russian Intervention in Chechnya
<FR-TITLE manually translated> L'intervention russe en Tchéchénie <FR-TITLE SYSTRAN> Interposition russe dans Chechnya <FR-TITLE BABYLON> Russe intervention dans Chechnya
C054 (query translations failed in French, Italian, German and Spanish)
<EN-TITLE> Final Four Results
<FR-TITLE manually translated> Résultats des demi-finales <FR-TITLE SYSTRAN> Résultats De la Finale Quatre <FR-TITLE BABYLON> final quatre résultat
<IT-TITLE manually translated> Risultati della "Final Four" <IT-TITLE SYSTRAN> Risultati Di Finale Quattro <IT-TITLE BABYLON> ultimo quattro risultato
<GE-TITLE manually translated> Ergebnisse im Halbfinale <GE-TITLE SYSTRAN> Resultate Der Endrunde Vier <GE-TITLE BABYLON> abschliessend Vier Ergebnis
<SP-TITLE manually translated> Resultados de la Final Four <SP-TITLE SYSTRAN> Resultados Del Final Cuatro <SP-TITLE BABYLON> final cuatro resultado

Recent works suggested various solutions to merge separate results list obtained from separate collections or distributed information services. As a preliminary approach, we will assume that each collection contains approximately the same number of pertinent items and that the distribution of the relevant documents is similar across the result lists. We could interleave the results in a round-robin fashion, based solely on the rank of the retrieved records. According to previous studies [22], [23], the retrieval effectiveness of such interleaving

schemes is around 40% below that of single retrieval schemes working with a single huge collection representing the entire set of documents. However, this decrease was found to diminish (around -20%) when using other collections [24].

Table 9. Average precision using different merging strategies, based on manually translated queries (top half) or automatically translated queries (bottom half)

	Average precision (% change)			
Title-Desc Original	round-robin baseline	raw-score	CORI	normalized score
Okapi–npn	34.23	15.87 (-53.6%)	13.00 (-62.0%)	**38.02** (+11.1%)
Lnu–ltc	32.09	31.41 (-2.1%)	21.23 (-33.8%)	**34.36** (+7.1%)
atn–ntc	31.31	23.03 (-26.4%)	17.15 (-45.2%)	**33.81** (+8.0%)
dtu–dtc	31.80	32.72 (+2.9%)	23.77 (-25.3%)	**34.60** (+8.8%)
ntc–ntc	20.97	17.30 (-17.5%)	15.37 (-26.7%)	**22.77** (+8.6%)
Mean difference		-19.36%	-38.61%	+8.70%
Translated queries				
Okapi–npn	29.59	13.08 (-55.8%)	11.19 (-62.2%)	**31.27** (+5.7%)
Lnu–ltc	28.84	25.41 (-11.9%)	17.30 (-40.0%)	**29.80** (+3.3%)
atn–ntc	27.32	17.56 (-35.7%)	13.49 (-50.6%)	**28.78** (+5.3%)
dtu–dtc	28.25	26.59 (-5.9%)	18.58 (-34.2%)	**30.21** (+6.9%)
ntc–ntc	19.16	13.14 (-31.4%)	11.60 (-39.5%)	**20.23** (+5.6%)
Mean difference		-28.14%	-45.30%	+5.37%

To account for the document score computed for each retrieved item (or the similarity value between the retrieved record and the request denoted score rsv_j), we might formulate the hypothesis that each collection is searched by the same or a very similar search engine and that similarity values are therefore directly comparable [25], [26]. Such a strategy, called raw-score merging, produces a final list sorted by the document score computed by each collection. However, as demonstrated by Dumais [27], collection-dependent statistics in document or query weights may vary widely among collections, and therefore this phenomenon may invalidate the raw-score merging hypothesis.

To account for this fact, we might normalize document scores within each collection by dividing them by the maximum score (e.i. the document score of the retrieved record in the first position). As a variant of this normalized score merging scheme, Powell et al. [28] suggest normalizing the document score rsv_j according to the following formula:

$rsv'_j = (rsv_j - rsv_{min}) / (rsv_{max} - rsv_{min})$

in which rsv_j is the original retrieval status value (or document score), and rsv_{max} and rsv_{min} are the maximum and minimum document score values that a collection could achieve for the current request. In this study, rsv_{max} is provided by the document score obtained by the first retrieved item and the retrieval

status value obtained by the 1000th retrieved record becomes the value of rsv_{min}.

Finally, we might use the CORI approach [23] within which each collection is viewed as a single gigantic document. In a first step, this system computes a collection score for each corpus in a manner similar to that used by an IR system to define a document score, according to a given request. In a second step, instead of using document scores directly as in the raw-score merging strategy, each document score is multiplied by the corresponding collection score and the system uses the value of this product as a key to sort the merged lists.

Table 9 provides an overview of retrieval performances for these various merging strategies by depicting average precision for the round-robin, raw-score and normalized score merging strategies, together with the performance achieved by the CORI approach. From studying this table, it seems that the best merging approach is the normalized score merging strategy. However, we must recall that in our experiments we used whole words when indexing English, French, Italian and Spanish collections and 5-grams when indexing German documents. Document scores are not really comparable across collections, thus penalizing both the raw-score merging and CORI approaches.

We used the normalized score merging strategy for our three official runs of the multilingual track, using the manually translated requests in the "UniNEmum" and "UniNEmuLm" runs as a baseline for comparison. In order to retrieve more relevant items from the various corpora, the "UniNEmuL" and "UniNEmuLm" runs were based on long requests (using the Title, Descriptive and Narrative sections) while the "UniNEmu" and "UniNEmum" runs were based on queries built with the Title and Descriptive logical sections.

Table 10. Descriptions of our official multilingual runs

Run name	English	French	Italian	German	Spanish
UniNEmum	original	original	original	original	original
expand	5 doc/10 ter	5 doc/10 ter	5 doc/10 ter	5 doc/30 ter	5 doc/10 ter
UniNEmu	original	syst+baby1	syst+baby2	syst+baby2	syst+baby2
expand	5 doc/10 ter	10 doc/15 ter	5 doc/50 ter	10 doc/40 ter	10 doc/15 ter
UniNEmuLm	original	original	original	original	original
expand	5 doc/10 ter	no	10 doc/15 ter	10 doc/100 ter	5 doc/10 ter
UniNEmuL	original	syst+baby1	syst+baby2	syst+baby1	syst+baby1
expand	5 doc/10 ter	10 doc/10 ter	5 doc/50 ter	10 doc/30 ter	10 doc/15 ter

As indicated in Table 10, our automatic "UniNEmu" and "UniNEmuL" runs used both the query translation furnished by the SYSTRAN system and one or two translation alternatives given by the BABYLON bilingual dictionary. The average precision achieved by these runs is depicted in Table 11.

Table 11. Average precision of our official multilingual runs

Run name	average prec.	% change	prec@5	prec@10	prec@20
UniNEmum	40.50	-	66.00	61.60	59.70
UniNEmu	33.73	-16.72%	61.20	60.40	55.60
UniNEmuLm	42.11	-	71.20	67.00	60.50
UniNEmuL	37.32	-11.37%	70.00	63.40	59.40

3 Conclusion

As our first participation in the CLEF retrieval tasks, we would suggest a general stopword list for the Italian, German and Spanish languages. Based on our experiments with the French language [3], we suggest a simple and efficient stemming procedure be used for these three languages. In this case and after comparing our approach with those used by others, removing inflectional suffixes attached only to nouns or adjectives seems to be worthwhile.

For the German language and its high frequency of compound word constructions, it might still be worthwhile to determine whether n-gram indexing approaches might produce higher levels of retrieval performance relative to an enhanced word segmentation heuristic, where a German dictionary is not required.

Moreover, we might also consider additional evidence sources when translating a request (e.g., based on statistical translation models [29] or on the EuroWordNet [30]) or logical approaches that could appropriately weight translation alternatives. Finally, when searching in multiple collections containing documents written in various languages, it might be worthwhile to look into those merging strategies that provide better results or include intelligent selection procedures in order to avoid searching in a collection or in a language that does not contain any relevant documents.

Acknowledgments. The author would like to thank C. Buckley from SabIR for giving us the opportunity to use the SMART system, without which this study could not have been conducted. This research was supported in part by the SNF (grant 21-58 813.99).

References

1. Robertson, S.E., Walker, S., Beaulieu, M.: Experimentation as a Way of Life: Okapi at TREC. Information Processing & Management **36** (2000) 95–108
2. Fox, C.: A Stop List for General Text. ACM-SIGIR Forum **24** (1999) 19–35
3. Savoy, J.: A Stemming Procedure and Stopword List for General French Corpora. Journal of the American Society for Information Science **50** (1999) 944–952
4. Sproat, R.: Morphology and Computation. The MIT Press, Cambridge (1988)

5. Lovins, J.B.: Development of a Stemming Algorithm. Mechanical Translation and Computational Linguistics **11** (1968) 22–31
6. Porter, M.F.: An Algorithm for Suffix Stripping. Program **14** (1980) 130–137
7. Figuerola, C.G., Gómez, R., Zazo Rodríguez, A.F.: Stemming in Spanish: A First Approach to its Impact on Information Retrieval. In *this volume*
8. Monz, C., de Rijke, M.: The University of Amsterdam at CLEF 2001. In *this volume*
9. Chen, A.: Multilingual Information Retrieval using English and Chinese Queries. In *this volume*
10. Molina-Salgado, H., Moulinier, I., Knutson, M., Lund, E., Sekhon, K.: Thomson Legal and Regulatory at CLEF 2001: Monolingual and Bilingual Experiments. In *this volume*
11. McNamee, P., Mayfield, J.: A Language-Independent Approach to European Text Retrieval. In: Peters, C. (ed.): Cross-Language Information Retrieval and Evaluation. Lecture Notes in Computer Science, Vol. 2069. Springer-Verlag, Berlin Heidelberg New York (2001) 131–139
12. McNamee, P., Mayfield, J.: JHU/APL Experiments at CLEF: Translation Resources and Score Normalization. In *this volume*
13. Savoy, J.: Bilingual Information Retrieval: CLEF-2000 Experiments. In Proceedings ECSQARU-2001 Workshop. IRIT, Toulouse (2001) 53–63
14. Salton, G.: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, Reading (1989)
15. Buckley, C., Singhal, A., Mitra, M., Salton, G.: New Retrieval Approaches Using SMART. In Proceedings TREC-4. NIST, Gaithersburg (1996) 25–48
16. Singhal, A., Choi, J., Hindle, D., Lewis, D.D., Pereira, F.: AT&T at TREC-7. In Proceedings TREC-7. NIST, Gaithersburg (1999) 239–251
17. Oard, D., Dorr, B.J.: A Survey of Multilingual Text Retrieval. Institute for Advanced Computer Studies and Computer Science Department, University of Maryland (1996), `http://www.clis.umd.edu/dlrg/filter/papers/mlir.ps`
18. Grefenstette, G. (ed.): Cross-Language Information Retrieval. Kluwer, Amsterdam (1998)
19. Peters, C. (ed.): Cross-Language Information Retrieval and Evaluation. Lecture Notes in Computer Science, Vol. 2069. Springer-Verlag, Berlin Heidelberg New York (2001)
20. Gachot, D.A., Lange, E., Yang, J.: The SYSTRAN NLP Browser: An Application of Machine Translation Technology. In: Grefenstette, G. (ed.): Cross-Language Information Retrieval. Kluwer, Boston (1998) 105–118.
21. Hull, D., Grefenstette, G.: Querying Across Languages. In Proceedings of the ACM-SIGIR'1996. The ACM Press, New York (1996) 49–57
22. Voorhees, E.M., Gupta, N.K., Johnson-Laird, B.: The Collection Fusion Problem. In Proceedings of TREC-3. NIST, Gaithersburg (1995) 95–104
23. Callan, J.P., Lu, Z., Croft, W.B.: Searching Distributed Collections with Inference Networks. In Proceedings of the ACM-SIGIR'1995. The ACM Press, New York (1995) 21–28
24. Savoy, J., Rasolofo, Y.: Report on the TREC-9 Experiment: Link-Based Retrieval and Distributed Collections. In Proceedings TREC-9. NIST, Gaithersburg (2001)
25. Kwok, K.L., Grunfeld L., Lewis, D.D.: TREC-3 Ad-hoc, Routing Retrieval and Thresholding Experiments Using PIRCS. In Proceedings of TREC-3. NIST, Gaithersburg (1995) 247–255
26. Moffat, A., Zobel, J.: Information Retrieval Systems for Large Document Collections. In Proceedings of TREC-3. Gaithersburg, NIST (1995) 85–93

27. Dumais, S.T.: Latent Semantic Indexing (LSI) and TREC-2. In Proceedings of TREC-2. NIST, Gaithersburg (1994) 105–115
28. Powell, A.L., French, J.C., Callan, J., Connell, M., Viles, C.L.: The Impact of Database Selection on Distributed Searching. In Proceedings of ACM-SIGIR'2000. The ACM Press, New York (2000) 232-239
29. Nie, J.Y., Simard, M.: Using Statistical Translation Models for Bilingual IR. In *this volume*
30. Vossen, P.: EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer, Dordrecht (1998)

This article was processed using the LaTeX macro package with LLNCS style

Table 12. Title of the queries of the CLEF-2001 test collection

C041 <EN-TITLE> Pesticides in Baby Food
C042 <EN-TITLE> U.N./US Invasion of Haiti
C043 <EN-TITLE> El Niño and the Weather
C044 <EN-TITLE> Indurain Wins Tour
C045 <EN-TITLE> Israel/Jordan Peace Treaty
C046 <EN-TITLE> Embargo on Iraq
C047 <EN-TITLE> Russian Intervention in Chechnya
C048 <EN-TITLE> Peace-Keeping Forces in Bosnia
C049 <EN-TITLE> Fall in Japanese Car Exports
C050 <EN-TITLE> Revolt in Chiapas
C051 <EN-TITLE> World Soccer Championship
C052 <EN-TITLE> Chinese Currency Devaluation
C053 <EN-TITLE> Genes and Diseases
C054 <EN-TITLE> Final Four Results
C055 <EN-TITLE> Swiss Initiative for the Alps
C056 <EN-TITLE> European Campaigns against Racism
C057 <EN-TITLE> Tainted-Blood Trial
C058 <EN-TITLE> Euthanasia
C059 <EN-TITLE> Computer Viruses
C060 <EN-TITLE> Corruption in French Politics
C061 <EN-TITLE> Siberian Oil Catastrophe
C062 <EN-TITLE> Northern Japan Earthquake
C063 <EN-TITLE> Whale Reserve
C064 <EN-TITLE> Computer Mouse RSI
C065 <EN-TITLE> Treasure Hunting
C066 <EN-TITLE> Russian Withdrawal from Latvia
C067 <EN-TITLE> Ship Collisions
C068 <EN-TITLE> Attacks on European Synagogues
C069 <EN-TITLE> Cloning and Ethics
C070 <EN-TITLE> Death of Kim Il Sung
C071 <EN-TITLE> Vegetables, Fruit and Cancer
C072 <EN-TITLE> G7 Summit in Naples
C073 <EN-TITLE> Norwegian Referendum on EU
C074 <EN-TITLE> Inauguration of Channel Tunnel
C075 <EN-TITLE> Euskirchen Court Massacre
C076 <EN-TITLE> Solar Energy
C077 <EN-TITLE> Teenage Suicides
C078 <EN-TITLE> Venice Film Festival
C079 <EN-TITLE> Ulysses Space Probe
C080 <EN-TITLE> Hunger Strikes
C081 <EN-TITLE> French Airbus Hijacking
C082 <EN-TITLE> IRA Attacks in Airports
C083 <EN-TITLE> Auction of Lennon Memorabilia
C084 <EN-TITLE> Shark Attacks
C085 <EN-TITLE> Turquoise Program in Rwanda
C086 <EN-TITLE> Renewable Power
C087 <EN-TITLE> Inflation and Brazilian Elections
C088 <EN-TITLE> Mad Cow in Europe
C089 <EN-TITLE> Schneider Bankruptcy
C090 <EN-TITLE> Vegetable Exporters