Report on CLEF-2001 Experiments

Jacques Savoy

Institut interfacultaire d'informatique, Université de Neuchâtel, Switzerland
Jacques.Savoy@unine.ch
Web site: www.unine.ch/info/

Abstract. For our first participation in CLEF retrieval tasks, our first 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 them. Our second aim was to suggest a combined approach that might be implemented in order to facilitate effective access to multilingual collections.

1. Monolingual indexing and search

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 by adding suffixes to the end of a root, etc.). Any adaptation of indexing or search 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 for use with languages other than English. Section 1.3 depicts the Okapi probabilistic model together with the description of the runs submitted by us in the 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* and *Frankfurter Rundschau* (German) and *EFE* (Spanish) and various news items edited by 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), and this number is little bit higher for the English collection (167.33) and clearly higher for the German corpora (509.131).

From the original documents and during the indexing process, we retained only 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 with the French collections and one with the Italian corpora within which we retained the following tags; for the French collections: <DE>, <KW>, <TB>, <CHA1>, <SUBJECTS>, <NAMES>, <NOM1>, <NOTE>, <GENRE>, <PEOPLE>, <SU11>, <SU21>, <GO11>, <GO12>, <GO13>, <GO14>, <GO24>, <TI01>, <TI02>, <TI03>, <TI04>, <TI05>, <TI06>, <TI07>, <TI08>, <TI09>, <ORT1>, <SOT1>, <SYE1> and <SYF1>. In the Italian corpora, and for one experiment, we used the following tags: <DE>, <KW>, <TB>, <ARGUMENTS>, <NAMES>, <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 …".

To evaluate our approaches, we used the SMART system as a test bed for implementing the OKAPI probabilistic model [Robertson 2000]. This year our experiments were conducted on an Intel Pentium III/600 (memory: 1 GB, swap: 2 GB, disk: 6 x 35 GB).

	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	140.476	129.908	509.131	120.245
standard error	126.315	118.605	97.602	431.527	60.148
median	138	102	92	396	107
maximum	1,812	1,723	1,394	8,136	682
minimum	2	3	1	1	5
max df	69,082	42,983	48,805	129,562	215,151
number of indexing terms / document					
mean	273.846	208.709	173.477	703.068	183.658
standard error	246.878	178.907	130.746	712.416	87.873
median	212	152	125	516	163
maximum	6,087	3,946	3,775	17,213	1,073
minimum	2	8	2	1	13
number of queries	47	48	47	49	49
no rel. for queries	#q:54 #q:57 #q:60	#q:64, #q:87	#q:43 #q:52 #q:64	#q:44	#q:61
number rel. items	856	1,193	1,246	2,238	2,694
mean rel. / request	18.21	24.85	26.51	42.04	54.97
standard error	22.56	24.57	24.37	47.77	63.68
median	10	17	18	27	26
maximum	107 (#q:50)	90 (#q:60)	95 (#q:50)	212 (#q:42)	261 (#q:42)
minimum	1 (#q:59)	1 (#q:43)	2 (#q:44)	1 (#q:64)	1 (#q:64)

Table 1: Test collection statistics

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 and French languages [Fox 1990], [Savoy 1999]. For the three other languages, we established a general stopword list by following the guidelines described in [Fox 1990]. Firstly, we sorted all word forms appearing in our corpora according to their frequency of occurrence and we 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 the term "política" (ranking 131) from the Spanish corpora were 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", "to be" in Spanish). The presence of homographs represents another debatable issue, and to some extent, we had to make arbitrary decisions concerning 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), and 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, an indexing procedure tries to conflate word variants into the same stem or root using a stemming algorithm. 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 [Sproat 1992]. As a first approach, we intended 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 [1968], which is based on a list of over 260 suffixes, while that of Porter [1980] looks for about 60 suffixes.

A "quick and dirty" stemming procedure has already been developed for the French language [Savoy 1999]. Based on the same concept, we have 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 Italian, the main inflectional rule 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, the suggested algorithms do not account for person and tense variations used by verbs or other derivational constructions.

Finally, the morphology of most European languages manifests other aspects that are not taken into account by our approach, with compound word constructions being just one example (e.g., handgun, worldwide). In German compound words are widely used and this causes more difficulties than does English. For example, a life insurance company employee would be "Lebensversicherungsgesellschaftsangeteller" (Leben + S + versicherung + S + gesellschaft +S + angeteller for life + insurance + company + employee). Also the morphological marker («S») is not always present (e.g., "Bankangetellenlohn" built as Bank + angetellen + lohn (salary)). Finally, diacritic characters are usually not present in an English collection (with some exceptions, such as "à la carte" or "résumé"); such characters are replaced by their corresponding non-accentuate letter.

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 [McNamee 2000], [Mayfield 2001]. This value of 5 was chosen for two reasons; it returns a better performance on CLEF-2000 corpora [Savoy 2001a], and, on the other hand, it is closed to the mean word length of our German corpora (mean word length: 5.87; standard error: 3.7).

1.3. Indexing and searching strategy

For the CLEF-2001 experiments, we conducted different experiments using the OKAPI probabilistic model [Robertson 2000] in which the weight w_{ij} assigned to a given term t_j in a document D_i was computed according to the following formula:

$$w_{ij} = \frac{(k_1 + 1) \cdot tf_{ij}}{K + tf_{ij}} \qquad \text{with } K = k_1 \cdot \left[(1 - b) + b \cdot \frac{l_i}{avdl}\right]$$

where tf_{ij} indicates the within-document term frequency, and b, k_1 are constants (fixed at b = 0.75 and $k_1 = 1.2$). K represents the ratio between the length of D_i measured by l_i (sum of tf_{ij}) and the collection mean denoted by advl (fixed at 900).

To index a keyword contained in a request Q, the following formula was used:

$$w_{qj} = tf_{qj} \cdot \ln[(n - df_j) / df_j]$$

where tf_{qj} indicates the search term frequency, df_j the collection-wide term frequency, n the number of documents in the collection.

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 [Buckley 1996] with $\alpha = 0.75$, $\beta = 0.75$ where the system was allowed to add to the original query generally 10 search keywords, extracted from the 5-best ranked documents.

In the monolingual track, we submitted six runs along with their corresponding descriptions as listed in Table 2. Four of them were fully automatic using the request's Title and Descriptive logical sections while the last two used more logical sections from the documents 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.

As a retrieval effectiveness indicator, we adopted the non-interpolated average precision (computed on the basis of 1,000 retrieved items per request by the TREC-EVAL program) allow for both precision and recall using a single number. These values (unofficial) are depicted in the last column of Table 2

Run name	Language	Query	Form	Query expansion	average precision
UniNEmofr	French	T-D	automatic	10 terms from 5 best docs	(50.00)
UniNEmoit	Italian	T-D	automatic	10 terms from 5 best docs	(48.65)
UniNEmoge	German	T-D	automatic	30 terms from 5 best docs	(42.32)
UniNEmoes	Spanish	T-D	automatic	10 terms from 5 best docs	(58.00)
UniNEmofrM	French	T-D-N	manual	no expansion	(51.84)
UniNEmoitM	Italian	T-D-N	manual	10 terms from 5 best docs	(54.18)

Table 2: Monolingual run descriptions

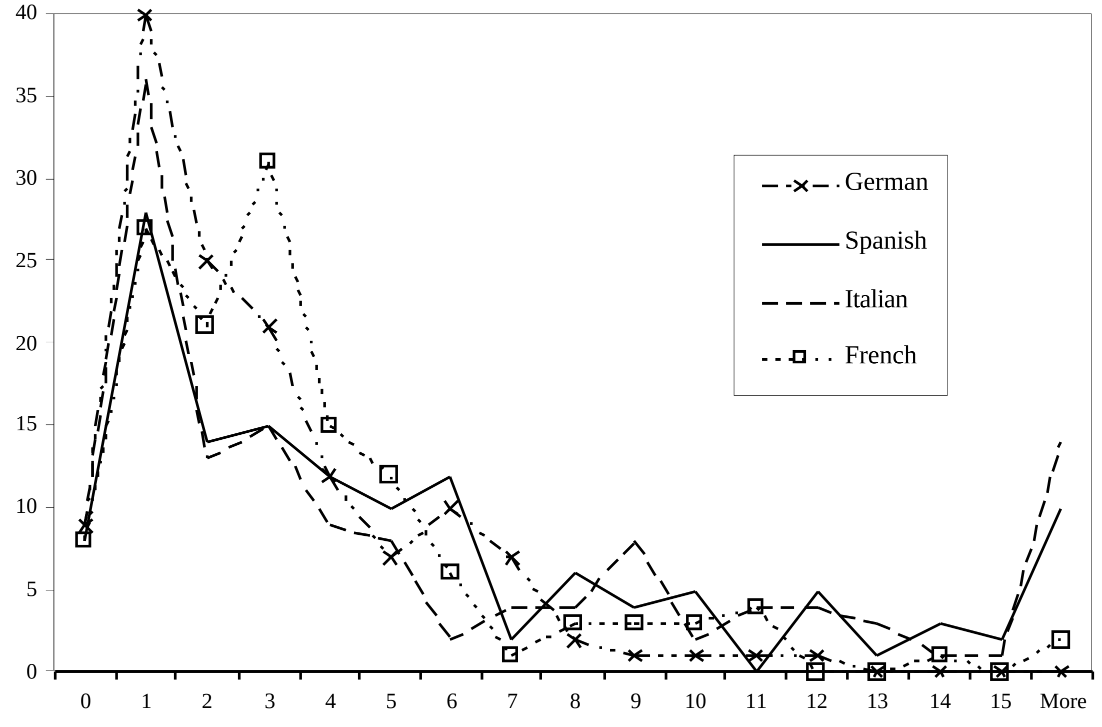
2. Multilingual information retrieval

In order to overcome language barriers [Oard 1996], [Grefenstette 1998], we based our approach on free and readily available translation resources that automatically provide translations to queries submitted 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 our merging strategy and a description of our runs submitted in the multilingual track.

2.1. Query automatic translation

In order to develop a fully automatically approach, we chose to translate the requests using the SYSTRAN® system [Gachot 1998] (available for free at http://www.systran.com) and to translate query terms word-by-word using the BABYLON bilingual dictionary (available at http://www.babylon.com) [Hull 1996]. In the latter case, the bilingual dictionary may suggest not only one, but several terms for the translation of each word. In our experiments, we decide to pick the first translation available (under the heading "babylon1") or the first two terms (indicated under the label "babylon2").

Figure 1: Distribution of the number of translation alternatives



In order to obtain a quantitative picture of a term's ambiguity, we analyze the number of translation alternatives generated by BABYLON's bilingual dictionaries. For this study, we do not take into account for determinants (e.g., "the"), conjunctions and prepositions (e.g., "and", "in", "of") or words appearing in our English stopword list (e.g., "new", "use"), 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.

From the data depicted in Table 3, we can see that the mean number of translations provided by BABYLON dictionaries varies according to language, 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 values of these distributions is rather small, varying from 2 for German to 4 for Spanish. Thus when considering 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 the Spanish language. Figure 1 shows more clearly how the number of translation alternatives is relatively concentrated around one.

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

	Number of translation alternatives			
Query (Title only)	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

Table 3: Number of translations given by the Babylon system for the English keywords appearing in the Title section of our queries

2.2. Examples of failures

Thus, in order to obtain a preliminary picture of the relative merit of each query translation-based strategy, we analyzed some queries by comparing the translations produced by our two machine-based tools with the request formulation written by an human being (examples are given in Table 4). 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 analyzed as the chemical symbol of illinium (chemical element #61 "found" by two at the University of Illinois in 1926; however this discovery was not confirmed and the chemical element #61 was finally found in 1947 and was named promethium). Moreover, the proper name "Sung" was analyzed 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 given in Italian and Spanish (where a more appropriate translation might be "mezzi finali" in Italian or "semifinales" in Spanish).

In query #48 "Peace-keeping forces in Bosnia" or in the request #57 "Tainted-blood trial", our automatic system was unable to decipher compound word constructions using the "-" symbol and 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 stage represents only the first step in proposing cross-language information retrieval systems. We also need to investigate situations where users write a request 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 these corpora and obtaining five results lists, we needed to merge them in order to provide users with a single list of retrieved articles.

Recent works have suggested various solutions to merge separate results list obtained from separate collections or distributed information services. As a first 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. Based solely on the rank of the retrieved records, we can interleave the results in a round-robin fashion. According to previous studies [Voorhees 1995], [Callan 1995], the retrieval effectiveness of such interleaving scheme is around 40% below that achieved from a single retrieval scheme working with a single huge collection that represents the entire set of documents. However, this decrease may diminish (around -20%) when using other collections [Savoy 2001b].

<num> C070 (both 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 BYBYLON> mort de Kim Il chanter

<IT-title manually translated> Morte di Kim Il Sung
<IT-title SYSTRAN> Morte di Kim Il cantata
<IT-title BYBYLON> morte di Kim ilinio cantare

<DE-title manually translated> Tod von Kim Il Sung
<GE-title SYSTRAN> Tod von Kim Il gesungen
<GE-title BYBYLON> Tod von Kim Ilinium singen

<ES-title manually translated> Muerte de Kim Il Sung
<ES-title SYSTRAN> Muerte de Kim Il cantada
<ES-title BYBYLON> muerte de Kim ilinio cantar

<num> 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 BYBYLON> Russe intervention dans Chechnya

<num> C054 (both 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 BYBYLON> final quatre résultat

<IT-title manually translated> Risultati della "Final Four"
<IT-title SYSTRAN> Risultati Di Finale Quattro
<IT-title BYBYLON> ultimo quattro risultato

<DE-title manually translated> Ergebnisse im Halbfinale
<GE-title SYSTRAN> Resultate Der Endrunde Vier
<GE-title BYBYLON> abschliessend Vier Ergebnis

<ES-title manually translated> Resultados de la Final Four
<ES-title SYSTRAN> Resultados Del Final Cuatro
<ES-title BYBYLON> final cuatro resultado

Table 4: Examples of unsucessful query translations

To take account of 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 the similarity values are therefore directly comparable [Kwok 1995], [Moffat 1995]. 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 [1994], 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 the document score 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.* [2000] suggest normalizing the document score rsv_j according to the following formula:

$$\text{rsv}'_j = \left(\text{rsv}_j - \text{rsv}_{\min}\right) / \left(\text{rsv}_{\max} - \text{rsv}_{\min}\right)$$

in which rsv_j is the original retrieval status value (or document score), and $\text{rsv}_{\max}$ and $\text{rsv}_{\min}$ are the maximum and minimum document score values that a collection could achieve for the current request. In this study, the $\text{rsv}_{\max}$ is given by the document score achieved by the first retrieved item and the retrieval status value obtained by the 1000th retrieved record gives the value of $\text{rsv}_{\min}$.

This merging strategy was used for our four runs that formed a part of the multilingual track. As a baseline for comparison, we used the manually translated requests in the "UniNEmum" and "UniNEmuLm" runs. In order to retrieve more relevant items from the various corpora, the "UniNEmuL" and "UniNEmuLm" runs were based on long request (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.

Run name	English	French	Italian	German	Spanish
UniNEmum expand	original 5 docs | 10 terms	original 5 docs | 10 terms	original 5 docs | 10 terms	original 5 docs | 30 terms	original 5 docs | 10 terms
UniNEmu expand	original 5 docs | 10 terms	systran+bybylon1 10 docs | 15 terms	systran+babylon2 5 docs | 50 terms	systran+babylon2 10 docs | 40 terms	systran+babylon2 10 docs | 15 terms
UniNEmuLm expand	original 5 docs | 10 terms	original no	original 10 docs | 15 terms	original 10 docs | 100 terms	original 5 docs | 10 terms
UniNEmuL expand	original 5 docs | 10 terms	systran+bybylon1 10 docs | 10 terms	systran+babylon2 5 docs | 50 terms	systran+bybylon1 10 docs | 30 terms	systran+bybylon1 10 docs | 15 terms

Table 5: Descriptions of our multilingual runs

As indicated in Table 5, 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 (unofficial) achieved by these runs are depicted in Table 6.

Run name	average precision	% change	Prec@5	Prec@10	Prec@20
UniNEmum	40.21	-	65.60	61.20	59.30
UniNEmu	33.28	-17.23%	60.40	59.80	55.10
UniNEmuLm	41.77	-	70.80	66.60	60.10
UniNEmuL	36.85	-11.78%	69.20	63.00	58.60

Table 6: Average precision (unofficial) of our multilingual runs

Conclusion

In this our first participation in CLEF retrieval tasks, we are suggesting a general stopword list for the Italian, German and Spanish languages. Based on our experiments with the French language [Savoy 1999], we would suggest simple and efficient stemming procedures for these three languages. Although we are convinced that these stopword lists and stemming procedures are not perfect, based on the relevance assessments of the CLEF-2001 corpora we should be able to improve upon these two retrieval tools.

For the German language and its high frequency of compound word constructions, it could still be worthwhile to find out whether n-gram indexing approaches might produce higher levels of retrieval performance relative to an enhanced word segmentation heuristic, without requiring a German dictionary.

Moreover, we could consider additional sources of evidence when translating a request (e.g., based on the EuroWordNet [Vossen 1998]) or logical approaches that would appropriately weight translation alternatives. Finally, when searching in multiple collections containing documents written in various languages, it might be worthwhile to look into better results merging strategies 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 by the SNSF (Swiss National Science Foundation) under grant 21-58 813.99.

Appendix 1. Queries

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

References

[Buckley 1996] Buckley, C., Singhal, A., Mitra, M. & Salton, G. (1996). New retrieval approaches using SMART. In Proceedings of TREC'4, (pp. 25-48). Gaithersburg: NIST Publication #500-236.

[Callan 1995] Callan, J. P., Lu, Z. & Croft, W. B. (1995). Searching distributed collections with inference networks. In Proceedings of the 18th International Conference of the ACM-SIGIR'95 (pp. 21-28). New York: The ACM Press.

[Dumais 1994] Dumais, S. T. (1994). Latent semantic indexing (LSI) and TREC-2. In Proceedings of TREC'2, (pp. 105-115). Gaithersburg: NIST Publication #500-215.

[Fox 1990] Fox C. (1990). A stop list for general text. *ACM-SIGIR Forum*, 24, 19-35.

[Gachot 1998] Gachot, D. A., Lange, E. & Yang, J. (1998). The SYSTRAN NLP browser: An application of machine translation technology. In Grefenstette G. (Ed.), Cross-language information retrieval, (pp. 105-118). Boston: Kluwer.

[Grefenstette 1998] Grefenstette, G. (Ed.) (1998). *Cross-language information retrieval*. Amsterdam: Kluwer.

[Hull 1996] Hull, D. & Grefenstette, G. (1996). Querying across languages: A dictionary-based approach to multilingual information retrieval. In Proceedings of the 19th International Conference of the ACM-SIGIR'96, (pp. 49-57). New York: The ACM Press.

[Kwok 1995] Kwok, K. L., Grunfeld L. & Lewis, D. D. (1995). TREC-3 ad-hoc, routing retrieval and thresholding experiments using PIRCS. In Proceedings of TREC'3, (pp. 247-255). Gaithersburg: NIST Publication #500-225.

[Lovins 1968] Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11(1), 22-31.

[Mayfield 2001] Mayfield, J., McNamee, P. & Piatko, J. (2001). The JHU/APL HAIRCUT system at Trec-8. In Proceedings TREC-8, (pp. 445-452). Gaithersburg: NIST Publication #500-246.

[McNamee 2000] McNamee, P. & Mayfield, J. (2000). A language-independent approach to European text retrieval. In Proceedings CLEF-2000, http://www.iei.pi.cnr.it/DELOS/CLEF/apl.doc.

[Moffat 1995] Moffat, A. & Zobel, J. (1995). Information retrieval systems for large document collections. In Proceedings of TREC'3, (pp. 85-93). Gaithersburg,: NIST Publication #500-225.

[Oard 1996] Oard, D. & Dorr, B. J. (1996). A survey of multilingual text retrieval. Institute for advanced computer studies and computer science department, University of Maryland, http://www.clis.umd.edu/dlrg/filter/papers/mlir.ps.

[Porter 1980] Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14, 130-137.

[Powell 2000] Powell, A. L., French, J. C., Callan, J., Connell, M. & Viles, C. L. (2000). The impact of database selection on distributed searching. In Proceedings of the 23rd International Conference of the ACM-SIGIR'2000, (pp. 232-239). New York: The ACM Press.

[Robertson 2000] Robertson, S. E., Walker, S. & Beaulieu, M. (2000). Experimentation as a way of life: Okapi at TREC. *Information Processing & Management*, 36(1), 95-108.

[Savoy 1999] Savoy, J. (1999). A stemming procedure and stopword list for general French corpora. *Journal of the American Society for Information Science*, 50(10), 944-952.

[Savoy 2001a] Savoy, J. (2001). Bilingual information retrieval: CLEF-2000 experiments. In Proceedings ECSQARU-2001 Workshop. Toulouse, France: to appear.

[Savoy 2001b] Savoy, J. & Rasolofo, Y. (2001). Report on the TREC-9 experiment: Link-based retrieval and distributed collections. In Proceedings TREC-9. Gaithersburg, MD: to appear.

[Sproat 1992] Sproat, R. (1992). *Morphology and computation.* Cambridge: The MIT Press.

[Voorhees 1995] Voorhees, E. M., Gupta, N. K. & Johnson-Laird, B. (1995). The collection fusion problem. In Proceedings of TREC'3, (pp. 95-104). Gaithersburg: NIST Publication #500-225.

[Vossen 1998] Vossen, P. (1998). *EuroWordNet: A multilingual database with lexical semantic networks*. Dordrecht: Kluwer.