ITC-irst at CLEF 2003: Monolingual, Bilingual, and Multilingual Information Retrieval

Nicola Bertoldi and Marcello Federico

ITC-irst - Centro per la Ricerca Scientifica e Tecnologica
I-38050 Povo, Trento, Italy.

Abstract

This paper reports on the participation of ITC-irst in the Cross Language Evaluation Forum 2003; in particular, in the monolingual, bilingual, small multilingual, and spoken document retrieval tracks. Considered languages were English, French, German, Italian, and Spanish. With respect to our CLEF 2002 system, the statistical models for bilingual document retrieval have been improved, more languages have been considered, and a novel multilingual information retrieval system has been developed, which combines several bilingual retrieval models into a statistical framework. As in the last CLEF, bilingual models integrate retrieval and translation scores over the set of N-best translations of the source query.

1. Introduction

This paper reports on the participation of ITC-irst in the Cross Language Evaluation Forum (CLEF) 2003. Several tracks were faced: monolingual document retrieval in Italian, French, German and Spanish; bilingual document retrieval from German to Italian and from Italian to Spanish; small multilingual document retrieval from English to English, German, French, and Spanish; and, finally, cross-language spoken document retrieval from French, German, Italian, Spanish to English.

The statistical cross-language information retrieval (CLIR) model presented in the 2002 CLEF evaluation (Federico and Bertoldi, 2002) was extended in order to cope with a multilingual target collection. Moreover, better query-translation probabilities were obtained by exploiting bilingual dictionaries and statistics from monolingual corpora. Basically, the ITC-irst system presented in the 2002 CLEF evaluation was expanded with a module for merging document rankings of different document collections generated by different bilingual systems.

Each bilingual system features a statistical model, which generates a list of the N-best query translations, and a basic IR engine, which integrates scores, computed by a standard Okapi model and a statistical language model, over multiple translations. Remarkably, training of the system's parameters just requires a bilingual dictionary, the target document collection, and a document collection in the source language.

This paper is organized as follows. Section 2 introduces the statistical approach to multilingual IR. Sections 3 briefly summarizes main features of our system, and describes the retrieval procedure. Section 4 and 5 present experimental results for each tracks we participated in. Section 6 closes the paper.

2. Multilingual Information Retrieval: Statistical approach

Multilingual information retrieval can be defined as the task of finding and ranking documents, which are relevant for a given topic, within a collection of texts in several language. As we know the language of each document, we may view the multilingual target collection as the union of distinct monolingual collections.

2.1. Multilingual retrieval model

Let a multilingual collection $\mathcal{D}$ contain documents in $\mathcal{L}$ different languages, where $\mathcal{D}$ results from the union of $\mathcal{L}$ monolingual sub-collections $\mathcal{D}_1, \ldots, \mathcal{D}_{\mathcal{L}}$. Let $\mathbf{f}$ be a query in a given source language, eventually different from any of the $\mathcal{L}$ languages. One would like to rank documents d within the multilingual collection $\mathcal{D}$, according to the posterior probability:

$$\Pr(d \mid \mathbf{f}) \propto \Pr(\mathbf{f}, d) \tag{1}$$

where the right term of formula (1) follows from the constancy of $\Pr(\mathbf{f})$, with respect to the ranking of documents.
A hidden variable l is introduced, which represents the language of either a sub-collection or a document.

$$\begin{aligned}\Pr(\mathbf{f}, d) &= \sum_l \Pr(l, \mathbf{f}, d) \\ &= \sum_l \Pr(l) \Pr(\mathbf{f}, d \mid l) \end{aligned} \tag{2}$$

where $\Pr(l)$ is an a-priori distribution over languages, which can be estimated from the multilingual collection or taken uniform. Formula (2) shows a weighted mixture of bilingual IR models depending on the sub-collection. However, given that we know the language each document is written in, we can assume that the probability $\Pr(\mathbf{f}, d \mid l)$ is larger than zero only if d belongs to the sub-collection $\mathcal{D}_l$.

Next, a hidden variable $\mathbf{e}$ is introduced, which represents a (term-by-term) translation of $\mathbf{f}$ into one of the $\mathcal{L}$ languages. Hence, we derive the following decomposition:

$$\begin{aligned}\Pr(\mathbf{f}, d \mid l) &= \sum_{\mathbf{e}} \Pr(\mathbf{f}, \mathbf{e}, d \mid l) \\ &\approx \sum_{\mathbf{e}} \Pr(\mathbf{f}, \mathbf{e} \mid l) \Pr(d \mid \mathbf{e}, l)\end{aligned} \tag{3}$$

In deriving formula (3), we make the assumption (or approximation) that the probability of document d given query $\mathbf{f}$, translation $\mathbf{e}$ and language l, does not depend on $\mathbf{f}$. Formula (3) puts in evidence a language-dependent query-translation model, $\Pr(\mathbf{f}, \mathbf{e} \mid l)$, and a collection-dependent query-document model, $\Pr(d \mid \mathbf{e}, l)$.

The language-dependent query-translation model is defined as follows:

$$\begin{aligned}\Pr(\mathbf{f}, \mathbf{e} \mid l) &= \Pr(\mathbf{f} \mid l) \Pr_l(\mathbf{e} \mid \mathbf{f}) \\ &\propto \begin{cases} \dfrac{\Pr_l(\mathbf{f}, \mathbf{e})}{\displaystyle\sum_{\mathbf{e}' \in \mathcal{T}_l(\mathbf{f})} \Pr_l(\mathbf{f}, \mathbf{e}')} & \text{if } \mathbf{e} \in \mathcal{T}_l(\mathbf{f}) \\ 0 & \text{otherwise} \end{cases}\end{aligned}$$

where $\mathcal{T}_l(\mathbf{f})$ is the set of all translations of $\mathbf{f}$ into language l. For practical reasons, this set is approximated with the set of the N most probable translations computed by the basic query-translation model $\Pr_l(\mathbf{f}, \mathbf{e})$. The term $\Pr(\mathbf{f} \mid l)$ can be considered independent from l and hence be discarded. The normalization introduced in formula (4) is needed in order to obtain ranking scores, which are comparable among different languages.

The collection-dependent query-document model is derived from a basic query-document model $\Pr_l(d \mid \mathbf{e})$ as follows:

$$\Pr(d \mid \mathbf{e}, l) = \begin{cases} \dfrac{\Pr_l(d, \mathbf{e})}{\displaystyle\sum_{d' \in \mathcal{I}(\mathbf{e}, l)} \Pr_l(d', \mathbf{e})} & \text{if } \mathbf{d} \in \mathcal{I}(\mathbf{e}, l) \\ 0 & \text{otherwise} \end{cases}$$

where $\mathcal{I}(\mathbf{e}, l)$ is the set of documents in $\mathcal{D}_l$ containing at least a word of $\mathbf{e}$.

The basic query document and query translation models are now briefly described; more details can be found in (Bertoldi and Federico, 2002). The subscript l, which refers to the specific language or collection the models are estimated on, will be omitted without loss of generality.

2.2. Basic Query-Document Model

The query-document model computes the joint probability of a query $\mathbf{e}$ and a document d, written in the same language. The query-document model considered in the experiments results from the combination of two different models: a language model and an Okapi based scoring function.

Language Model The joint probability can be factored out as follows:

$$\Pr(\mathbf{e}, d) = \Pr(\mathbf{e} \mid d) \Pr(d) \tag{4}$$

where the a-priori probability of d, $\Pr(d)$, is assumed to be uniform, and the probability of $\mathbf{e}$ given d to be an order-free multinomial (bag-of-word) model:

$$\Pr(\mathbf{e} = e_1, \ldots, e_n \mid d) = \prod_{k=1}^{n} p(e_k \mid d) \tag{5}$$

Okapi The joint probability can be obtained through the normalization over queries and documents of a generic scoring function $s(\mathbf{e}, d)$:

$$\Pr(\mathbf{e}, d) = \frac{s(\mathbf{e}, d)}{\sum_{\mathbf{e}', d'} s(\mathbf{e}', d')} \tag{6}$$

The denominator is considered only for the sake of normalization, but can be disregarded in the computation of equation (3).

A scoring function derived from the standard Okapi formula, is used

$$s(\mathbf{e} = e_1, \ldots, e_n, d) = \prod_{k=1}^{n} idf(e_k)^{W_d(e_k)} \tag{7}$$

Combination Previous work (Bertoldi and Federico, 2001) showed that the two models rank documents almost independently. Hence, information about the relevant documents can be gained by integrating the scores of both methods. Combination of the two models is implemented by just taking the sum of scores, after a suitable normalization.

2.3. Basic Query-Translation Model

The query-translation model computes the probability of any query-translation pair. This probability is modeled by an HMM (Rabiner, 1990) in which the observable variable is the query $\mathbf{f}$ in the source language, and the hidden variable is its translation $\mathbf{e}$ in the target language. According to the HMM, the joint probability of a pair $(\mathbf{f}, \mathbf{e})$ is decomposed as follows:

$$\begin{aligned}&Pr(\mathbf{f} = f_1, \ldots, f_n, \mathbf{e} = e_1, \ldots, e_n) \\ &\qquad = p(e_1) \prod_{k=2}^{n} p(e_k \mid e_{k-1}) \prod_{k=1}^{n} p(f_k \mid e_k)\end{aligned} \tag{8}$$

The term translation probabilities $p(f \mid e)$ are estimated from a bilingual dictionary as follows:

$$\Pr(f \mid e) = \frac{\delta(f,e)}{\sum_{f'} \delta(f',e)} \qquad (9)$$

where $\delta(f,e) = 1$ if the term e is one of the translations of term f and $\delta(f,e) = 0$ otherwise. This flat distribution can be refined through the EM algorithm (Dempster et al., 1977) by exploiting a large corpus in the source language.

The target LM probabilities $p(e \mid e')$, are estimated on the target document collection, through an order-free bigram LM, which tries to compensate for different word positions induced by the source and target languages. Let

$$p(e \mid e') = \frac{p(e,e')}{\sum_{e''} p(e'',e')} \qquad (10)$$

where $p(e,e')$ is the probability of e co-occurring with e', regardless of the order, within a text window of fixed size. Smoothing of this probability is performed through absolute discounting and interpolation.

3. System architecture

As shown in Section 2, the ITC-irst multilingual IR system features several independent bilingual retrieval systems, which return collection-dependent rankings, and a module for merging these results into a global ranking with respect to the whole multilingual collection. Moreover, language-dependent text preprocessing modules have been implemented to process documents and queries. Figure 3. shows the architecture of the system.

Two merging criteria were developed. The first, we call `stat` method, implements the statistical model introduced in Section 2: for each language, language-dependent relevance scores of documents, computed by the bilingual IR systems are normalized in order to have language independent scores, and, hence, a global ranking is created.

The second criterion, we call `rank` method, exploits the document rank positions only, i.e. all the collection dependent rank lists are joined and documents are globally sorted according to the inverse of their original rank position.

Monolingual and bilingual versions of the system trivially follows by omitting the query-translation model and by limiting the collection to one language, respectively.

3.1. Preprocessing

In order to homogenize the preparation of data, and, hence, to reduce workload, a standard procedure was defined. More specifically, the following preprocessing steps were applied both to documents and queries in every language:

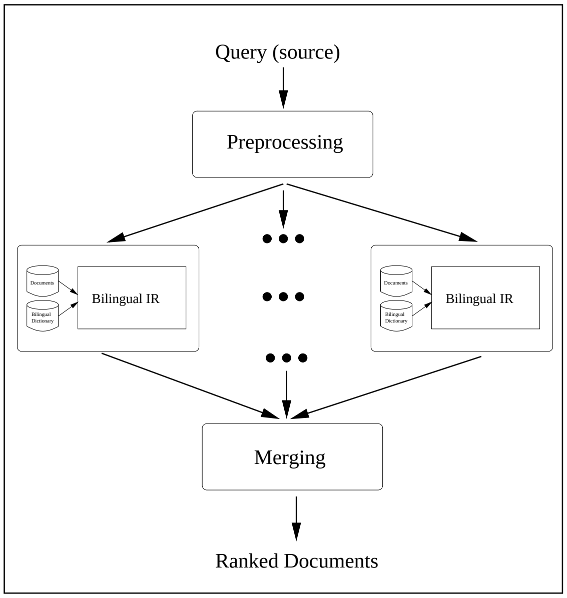

Figure 1: Architecture of the multilingual IR system.

- *Tokenization* was performed to separate words from punctuation marks, to recognize abbreviations and acronyms, correct possible word splits across lines, and discriminate between accents and quotation marks.

- *Stemming* was performed by using a language-dependent Porter-like algorithm (Frakes and Baeza-Yates, 1992), freely available at snowball.tartarus.org.

- *Stop-terms removal* was applied on the documents by removing terms included in a language-dependent public list (www.unine.ch/info/clef).

- *Proper names and numbers* in queries were recognized in order to improve coverage of the dictionary.

- *Out-of-dictionary terms* which have not been recognized as proper names or numbers were removed.

3.2. Blind Relevance Feedback

After document ranking, the following Blind Relevance Feedback (BRF) technique was applied. First, the documents matching the source query $\mathbf{e}$ are ranked, then the B best ranked documents are taken and the R most relevant terms in them are added to the query, and the retrieval phase is repeated. In the CLIR framework, R terms are added to each single translation of the N-best list and the retrieval algorithms is repeated once again. In this work, 15 new search terms are selected from the top 5 documents according to the Offer Weight proposed in (Johnson et al., 1999).

4. Experimental Evaluation

ITC-irst submitted 4 monolingual runs in French, German, Italian, and Spanish, 4 Italian-Spanish bilingual runs, 2 German-Italian bilingual runs, and 4 small multilingual runs using queries in English to search documents in English, French, German, and Spanish. Moreover, some unofficial experiments were performed for the sake of comparison.

4.1. Data

In Table 1, statistics about the target collections for the five considered languages are reported.

Language	#docs	#words
English	166,754	100,971,969
French	129,809	52,275,689
German	294,809	99,461,570
Italian	153,208	54,434,345
Spanish	454,045	171,971,487
Multi-4	1,045,417	424,680,715

Table 1: Statistics about target collections.

Table 2 reports statistics about the topics and corresponding relevant documents in each collection (topics with no relevant document are not considered).

Language	#queries	#rel.docs
English	54	1006
French	52	946
German	56	1825
Italian	51	809
Spanish	57	2368
Multi-4	60	6145

Table 2: Statistics about queries.

Dictionary	#entries	avg. # translations
English-French	44728	1.97
English-German	131429	1.88
English-Italian	44195	1.95
English-Spanish	47305	1.83
Italian-Spanish	66059	3.94
German-Italian	103618	3.91

Table 3: Statistics about dictionaries.

Bilingual dictionaries from English to the other languages were gathered from public available resources. Unfortunately, German-Italian and Italian-Spanish dictionaries were not available. Hence, the missing dictionaries were built from other available dictionaries using English as a pivot language. For example, an Italian-Spanish dictionary was derived by exploiting the Spanish-English and Italian-English dictionaries as follows: the translation alternatives of an Italian term are all Spanish translations of all English translations of that term. Table 2 reports some statistics of the bilingual dictionaries. It is worth noticing that for the generated dictionaries the average number of translation alternatives is about twice larger than that of original dictionaries. This would suggest that they contain two wrong translations per entry, on the average.

Moreover, all term translation probabilities, but the German-Italian ones, were estimated through the EM algorithm by using the corresponding document collections.

4.2. Results

Table 4 reports main settings and official `mAvPr` scores for each run. In particular, the number of N-best translations (1 vs. 10), the type of bilingual dictionary (flat vs. estimated through EM algorithm), and the merging policy (looking at the rank vs. the stat) are indicated. Source and target languages are indicated in the run name.

Monolingual results As shown in Table 4, our monolingual retrieval system achieves good results for all languages. More than 70% of queries have `mAvPr` greater than or equal to the median values. It is worth noticing that `mAvPrs` are pretty the same for all languages.

Bilingual results Italian-Spanish results show that the estimation of translation probabilities through the EM algorithm is quite effective, especially in combination with the 10-best translations.

Language	monolingual	bilingual from English
French	.5339	.4297
German	.5173	.4378
Italian	.5397	.4184
Spanish	.5375	.4298

Table 5: Comparison of monolingual and bilingual performance.

Table 5 reports `mAvPr` for monolingual and bilingual runs for every language; the 10-best translations were obtained with EM estimated translation probabilities. A relative degradation between 15% and 22% is always observed. This means that the translation process causes almost equal losses in performance for each language pair.

Multilingual results As shown in Table 4, about 60% of the queries have `mAvPr` greater than or equal

Official Run	Setting	mAvPr	<mdn	=mdn	>mdn	bst
IRSTfr_1		.5339	15	10	27	11
IRSTde_1		.5173	16	5	35	6
IRSTit_1		.5397	11	8	32	10
IRSTes_1		.5375	17	3	37	5
IRSTit2es_1	10-best, EM	.4262	31	1	25	2
IRSTit2es_2	10-best, flat	.4006	36	1	20	2
IRSTit2es_3	1-best, EM	.4053	33	1	23	2
IRSTit2es_4	1-best, flat	.4009	35	1	21	2
IRSTde2it_1	10-best, flat	.2291	38	0	18	0
IRSTde2it_2	1-best, flat	.2437	36	0	20	0
IRSTen2xx_1	10-best, EM, rank	.3147	23	1	36	0
IRSTen2xx_2	10-best, EM, stat	.3089	22	2	36	1
IRSTen2xx_3	10-best, flat, rank	.3084	25	2	33	0
IRSTen2xx_4	10-best, flat, stat	.3036	25	1	34	1

Table 4: Main settings and results of the official runs. Comparison against the median and best values.

to the median values. The merging method based on the rank is a little more effective, but differences are very low. Again, the EM estimation of term probabilities slightly improves performance.

The merging criteria were also applied to the monolingual runs, in order to obtain an upper bound for our multilingual retrieval system. The achieved mAvPrs for this virtual experiment were .3754 and .3667 for the "rank" and "stat" criteria, respectively. The relative degradation is very similar to that observer for bilingual experiments.

5. Cross-Language Spoken Document Retrieval

ITC-irst participated also in the Cross-Language Spoken Document Retrieval (CLSDR) track, which consists in searching for relevant stories within a collection of automatically transcribed English broadcast news. Topics correspond in 50 short queries manually translated from English into French, German, Italian, and Spanish. For the CLSDR track, the bilingual version of the ITC-irst IR system was applied, with little changes in the BRF expansion of queries. Moreover, German text were also processed for splitting compound words, by using a DP based algorithm.

5.1. Query expansion on parallel corpora

As the number of stories in the SDR target collection was quite small, a double query expansion policy was chosen. New terms are added which are extracted not only from the target collection, but also from a large corpus of written texts, consisting of newspapers and news wires.

As a parallel corpus for query expansion, newspaper articles of the North American News Text corpus were used (www.nist.gov/speech/tests/sdr). In particular, 313K documents are extracted from *Los Angeles Times*, *Washington Post*, *New York Times*, and *Associated Press Worldstream*, issued between September 1997 and April 1998. Unfortunately, the available texts do not entirely cover the test period. The following strategy was chosen: first query expansion was performed on parallel texts, and then on target collection.

5.2. Results

Table 6 reports the official submitted runs, and some unofficial runs (in italics), used for comparison.

Official run	Query	mAvPr
mono-brf	EN	.3944
mono-brf-brf	EN	.4244
fr-en-1bst-brf-bfr	FR	.2281
fr-en-sys-brf-bfr	FR	.3064
de-en-dec-1bst-brf-bfr	DE	.2676
de-en-1bst-brf-bfr	DE	.2523
de-en-sys-brf-bfr	DE	.2880
it-en-1bst-brf-bfr	IT	.2347
it-en-sys-brf-bfr	IT	.3218
es-en-1bst-brf-bfr	ES	.2746
es-en-sys-brf-bfr	ES	.3555

Table 6: mAvPr results of CLSDR track at CLEF 2003

The official English monolingual run was performed in order to evaluate the quality of the retrieval system. ITC-irst performance is about 10% above the other participants. For this experiment the query expansion on the parallel corpus was not applied. If not so, a relative improvement of 7% is observed. As the double query expansion policy is quite effective, was applied in all the other experiments.

In the bilingual experiments, query were translated either through our 1-best translation approach or by the Babelfish translation service, powered by Systran, which is available on the Internet (world.altavista.com). Run names are indicating with `1bst` and `sys`, respectively. Commercial translations outperforms our approach.
German word decompounding seems to be slightly effective, as shown by comparing the run without decompounding (*de-en-1bst-brf-bfr*) and the with (`de-en-dec-1bst-brf-bfr`).

6. Conclusion

This paper presented a multilingual IR system developed at ITC-irst. A complete statistical model was defined which combines several bilingual retrieval model. The system was evaluated in the CLEF2003 campaign in the monolingual, bilingual, and multilingual tracks. The basic monolingual IR model resulted very competitive on every languages. The multilingual IR systems also achieves higher performance than the median. Experiments in the Cross-Language Spoken Document Retrieval task, which uses very short queries, showed that significantly better results are still achieved by using translations produced by a commercial system.

7. References

Bertoldi, N. and M. Federico, 2001. ITC-irst at CLEF 2000: Italian monolingual track. In Carol Peters (ed.), *Cross-Language Information Retrieval and Evaluation*, volume 2069 of *Lecture Notes in Computer Science*. Heidelberg, Germany: Springer Verlag.

Bertoldi, N. and M. Federico, 2002. ITC-irst at CLEF 2001: Monolingual and bilingual tracks. In Carol Peters, Martin Braschler, Julio Gonzales, and Michael Kluck (eds.), *Cross-Language Information Retrieval and Evaluation*, volume 2406 of *Lecture Notes in Computer Science*. Heidelberg, Germany: Springer Verlag.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39:1–38.

Federico, Marcello and Nicola Bertoldi, 2002. Statistical cross-language information retrieval using n-best query translations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Tampere, Finland.

Frakes, William B. and Ricardo Baeza-Yates (eds.), 1992. *Information Retrieval: Data Structures and Algorithms*. Englewood Cliffs, NJ: Prentice Hall.

Johnson, S.E., P. Jourlin, K. Spark Jones, and P.C. Woodland, 1999. Spoken document retrieval for TREC-8 at Cambridge University. In *Proceedings of the 8th Text REtrieval Conference*. Gaithersburg, MD.

Rabiner, Lawrence R., 1990. A tutorial on hidden Markov models and selected applications in speech recognition. In Alex Weibel and Kay-Fu Lee (eds.), *Readings in Speech Recognition*. Los Altos, CA: Morgan Kaufmann, pages 267–296.