Information Retrieval, 7, 53–72, 2004

Statistical Models for Monolingual and Bilingual Information Retrieval

NICOLA BERTOLDI bertoldi@itc.it
MARCELLO FEDERICO federico@itc.it
ITC-irst, Centro per la Ricerca Scientifica e Tecnologica, I-38050, Povo, Italy

Received December 5, 2002; Revised May 15, 2003; Accepted May 16, 2003

Abstract. This work reviews information retrieval systems developed at ITC-irst which were evaluated through several tracks of CLEF, during the last three years. The presentation tries to follow the progress made over time in developing new statistical models first for monolingual information retrieval, then for cross-language information retrieval. Besides describing the underlying theory, performance of monolingual and bilingual information retrieval models are reported, respectively, on Italian monolingual tracks and Italian-English bilingual tracks of CLEF. Monolingual systems by ITC-irst performed consistently well in all the official evaluations, while the bilingual system ranked in CLEF 2002 just behind competitors using commercial machine translation engines. However, by experimentally comparing our statistical topic translation model against a state-of-the-art commercial system, no statistically significant difference in retrieval performance could be measured on a larger set of queries.

Keywords: monolingual information retrieval, cross-language information retrieval, statistical models, machine translation

1. Introduction

Information Retrieval (IR) is the task of finding documents, inside a known collection, which are relevant to a given topic or query. If topics and documents are written in the same language, e.g. English, we have so called monolingual IR, otherwise cross-language IR (CLIR) occurs. In particular, if only two languages are involved, e.g. French for queries and English for documents, IR is called bilingual; if the collection contains documents in more than one language, e.g. English, French and Italian, IR is instead called multilingual. As habit in the IR literature and with no risk of confusion, henceforth, the term CLIR will be used only to mean bilingual IR.

This paper presents on IR system developed at ITC-irst, during the last three years, to tackle monolingual and bilingual IR tracks at CLEF. The monolingual IR system features three different models for matching topics against documents: a statistical language model, an Okapi model, and a combination of the two approaches. The bilingual IR system features a statistical framework which couples two basic components: a translation model, based on hidden Markov models, and a retrieval model which works as in the monolingual case. The two models can be either put in cascade or tightly coupled. The latter case results in a probability score computed by integrating over a set of possible translations of the query. Training data for the translation model consists of a bilingual dictionary and the target document collection.

ITC-irst participated with these systems in Italian monolingual tracks of all CLEF evaluation campaigns, and in Italian-English bilingual tracks at CLEF 2001 and CLEF 2002.

This paper is organized as follows. Section 2 presents work on monolingual IR. In particular, it introduces three query-document matching models and describes topic/document preprocessing and query expansion. Section 3 presents the statistical query-translation model used for CLIR. Section 4 introduces two bilingual IR models, which combine in different ways the translation model and the query-document matching models, presented beforehand. Performance of the systems in the CLEF tracks are shown in Section 5. Sections 6 concludes the paper with a discussion about interesting issues which emerged from our participation in CLEF.

2. Monolingual IR

Formally, monolingual IR can be approached as follows: given a query $\mathbf{q} = q_1, \ldots, q_n$, rank all documents d in a collection $\mathcal{D}$ according to a probability or a scoring function $\mathcal{S}(\mathbf{q}, d)$, which measures the relevance of d with respect to $\mathbf{q}$. In the following, three query-document matching criteria are introduced. The first is based on a statistical language model (LM), the second is derived from the Okapi framework, and the last is a combination of the first two. (Main notation used in the following is summarized in Table 1.)

2.1. *Language model*

The relevance of a document d with respect to a query $\mathbf{q}$ can be expressed through a joint probability, which can be decomposed as follows:

$$\Pr(\mathbf{q}, d) = \Pr(\mathbf{q} \mid d) \Pr(d) \tag{1}$$

where $\Pr(\mathbf{q} \mid d)$ represents the likelihood of $\mathbf{q}$ given d, and $\Pr(d)$ represents the a-priori probability of d. By assuming no a-priori knowledge about the documents and an order-free multinomial model for the likelihood, the following probability score can be derived:

$$\Pr(\mathbf{q}, d) \propto \prod_{i=1}^{n} \Pr(q_i \mid d) \tag{2}$$

Table 1. List of often used symbols.

$\mathbf{q}, \mathbf{f}, \mathbf{e}, d$	Generic query, query in French, query in English, and document
q, f, e	Generic term, term in French, term in English
$\mathcal{D}$	Collection of documents
$\mathcal{V}$,$\mathcal{V}(d)$	Number of different terms in $\mathcal{D}$, and in document d
N, $N(d)$	Number of term occurrences in $\mathcal{D}$, and in document d
$N(q)$, $N(d, q)$, $N(\mathbf{q}, q)$	Frequency of term q in $\mathcal{D}$, in document d, and in query $\mathbf{q}$
$\bar{l}$	Average length of documents in $\mathcal{D}$
N_q	Number of documents in $\mathcal{D}$ which contain term q

By taking the logarithm, we can define the following scoring function:

$$lm(\mathbf{q}, d) = \sum_{q \in \mathbf{q}} N(\mathbf{q}, q) \ \log \Pr(q \mid d) \tag{3}$$

where the sum is over the set of terms in the query $\mathbf{q}$.

The probability $\Pr(q \mid d)$ that a term q is generated by d can be estimated by applying statistical language modeling techniques (Federico and De Mori 1998). Previous work (Miller et al. 1998, Ng 1999) proposed to interpolate relative frequencies of each document with those of the whole collection, with interpolation weights estimated by maximum likelihood on the documents. Here, the same interpolation scheme is applied but weights are estimated according to the smoothing method by Witten and Bell (1991). In particular, word frequencies of a document are smoothed linearly and the amount of probability assigned to never observed terms is made proportional to the number of different words contained in the document. Hence, the following probability estimate results:

$$\Pr(q \mid d) = \frac{N(d, q)}{N(d) + \mathcal{V}(d)} + \frac{\mathcal{V}(d)}{N(d) + \mathcal{V}(d)} \ \Pr(q) \tag{4}$$

where $\Pr(q)$, the word probability over the collection, is estimated by interpolating the smoothed relative frequency with the uniform distribution over the collection's vocabulary V:

$$\Pr(q) = \frac{N(q)}{N + \mathcal{V}} + \frac{\mathcal{V}}{N + \mathcal{V}} \frac{1}{\mathcal{V}}. \tag{5}$$

2.2. Okapi model

Okapi (Robertson et al. 1994) is the name of a retrieval system project that developed a family of scoring functions. According to the Okapi framework, every term in the query is weighted according to its relevance within a document and within the whole collection. In our IR system the following function was used:

$$\text{okapi}(\mathbf{q}, d) = \sum_{q \in \mathbf{q}} N(\mathbf{q}, q) \, W_d(q) \, \log W_{\mathcal{D}}(q) \tag{6}$$

where

$$W_d(q) = \frac{N(d, q)(k_1 + 1)}{k_1(1 - b) + k_1 b \frac{N(d)}{l} + N(d, q)} \tag{7}$$

weighs the relevance of the term q inside the document d, and:

$$W_{\mathcal{D}}(q) = \frac{N - N_q + 0.5}{N_q + 0.5} \tag{8}$$

is the term inverted document frequency, which weighs the relevance of term q inside the whole collection $\mathcal{D}$.

Parameter values $k_1 = 1.5$ and $b = 0.4$ were empirically estimated (Bertoldi and Federico 2001) on some development data. It is worth noticing that our scoring function corresponds to the well known BM25(k_1, k_2, k_3, b) model (Robertson et al. 1994), with the setting $k_2 = 0$, $k_3 = \infty$, $k_1 = 1.5$ and $b = 0.4$.

The Okapi and the language model scoring functions present some analogy. In particular, formula (6) can be put in a probabilistic form which maintains the original ranking, thanks to the monotonicity of the exponential function. Hence, a joint probability distribution can be defined which, disregarding a normalization constant factor, is:

$$\Pr(\mathbf{q}, d) \propto \prod_{i=1}^{n} W_{\mathcal{D}}(q_i)^{W_d(q_i)} \tag{9}$$

Henceforth, query-document relevance models will be indicated by the joint probability $\Pr(\mathbf{q}, d)$, regardless of the used model, unless differently specified.

2.3. *Combined method*

By looking at the Italian monolingual runs of our first participation in CLEF 2000 (Bertoldi and Federico 2001), it emerged that the LM and the Okapi model have quite different behaviors. This suggested that if the two methods rank documents independently, more information about the relevant documents could be gained by integrating the scores of the two methods.

In order to compare the rankings of two models, the Spearman's rank correlation (Mood et al. 1974) was applied, which confirmed some degree of independence between the two information retrieval models. Hence, a combination of the two models (Bertoldi and Federico 2001) was implemented by just taking the sum of scoring functions, namely $lm(\mathbf{q}, d)$ and okapi($\mathbf{q}$, d). Actually, in order to adjust scale differences, single scores were re-scaled in the range [0, 1] before summation. Normalization was computed over union of the 300 top ranking documents of each method. It can be shown that summation of the normalized scores corresponds to a multiplication of probabilities, according to the above defined joint probabilities.

2.4. *Document/Query preprocessing*

In the following, a brief description of the modules used to preprocess Italian (Federico 2000) and English documents/queries is given. Tables 2 and 3 show, respectively, an English topic and the various preprocessing steps.

Tokenization. Words are isolated from punctuation marks, abbreviations and acronyms are recognized, possible word splits across lines are corrected, and accents are distinguished from quotation marks.

Table 2. English topic 44.

<top>
<num> C044 </num>
<EN-title> Indurain Wins Tour </EN-title>
<EN-desc> Reactions to the fourth Tour de France won by Miguel Indurain. </EN-desc>
<EN-narr> Relevant documents comment on the reactions to the fourth consecutive victory of Miguel Indurain in the Tour de France. Also relevant are documents discussing the importance of Indurain in world cycling after this victory. </EN-narr>
</top>

Table 3. Processing of English short topic 44.

a	Title	Indurain Wins Tour.
	Desc.	Reactions to the fourth Tour de France won by Miguel Indurain.
b	Title	indurain win tour.
	Desc.	reaction to the fourth tour de franc won by miguel indurain.
c		indurain win tour reaction fourth tour franc won miguel indurain
d	−55.66	vincere tour reazione quarto tour francia
	−56.07	vincere tour reazione quarto giro francia
	−56.27	vincere tour reazione quarto tournee francia
	...	
e	−55.66	indurain vincere tour reazione quarto tour francia miguel indurain
	−56.07	indurain vincere tour reazione quarto giro francia miguel indurain
	−56.27	indurain vincere tour reazione quarto tournee francia miguel indurain
	...	
f		indurain vincere tour reazione quarto tour de france vincere miguel indurain
g		giro vittoria indurain reazione quarto giro de francia vincere miguel indurain

(a) Title and Descriptive fields of the original topic; (b) tokenized and stemmed query; (c) query after stop term removal; (d) N-best translations into Italian with logarithm of probabilities; (e) translated queries after adding proper names; (f) corresponding human translated query; (g) corresponding translated query by Systran.

Morpho-syntactic analysis. Base forms of Italian words are obtained by combining morpho-syntactic analysis and statistical parts-of-speech tagging.

Stemming. Word stemming is performed on English texts by using the Porter algorithm (Porter 1980).

Stop-terms removal. Non relevant words are filtered out on the basis of their POS (only for Italian) and their inverted document frequency.

2.5. Blind relevance feedback

Blind relevance feedback (BRF) is a well known technique that permits to improve retrieval performance. The basic idea is to perform retrieval in two steps. First, documents matching

the original query **q** are ranked, then the R top ranked documents are taken and the T most relevant terms in them are added to the query. Hence, retrieval is repeated with the augmented query. In this work, new search terms are extracted from the R top documents according to (Johnson et al. 1999):

$$r_q \log \frac{(r_q + 0.5)(N - N_q - R + r_q + 0.5)}{(N_q - r_q + 0.5)(R - r_q + 0.5)} \tag{10}$$

where r_q is the number of documents, among the top R, which contain term q.

BRF parameters R and T were estimated just for the Okapi model on some development data (Bertoldi and Federico 2001). The best settings resulted $T = 15$ and $R = 5$.

3. Query translation model

Query translation for CLIR is based on a hidden Markov model (HMM) (Rabiner 1990), in which the observable part is the query **f** in the source language, e.g. French, and the hidden part is a corresponding query **e** in the target language, e.g. English. The model only assumes that the two queries have the same length. The joint probability of a pair (**f**, **e**) is computed as follows:

$$\Pr(\mathbf{f} = f_1, \ldots, f_n, \mathbf{e} = e_1, \ldots, e_n) = \prod_{k=1}^{n} \Pr(f_k \mid e_k) \Pr(e_k \mid e_{k-1}) \tag{11}$$

Formula (11) puts in evidence two different conditional probabilities: the term translation probabilities $p(f \mid e)$ and the target LM probabilities $p(e \mid e')$.

Probabilities $Pr(f \mid e)$ are estimated from a translation dictionary as follows:

$$\Pr(f \mid e) = \frac{\delta(f, e)}{\sum_{f'} \delta(f', e)} \tag{12}$$

where $\delta(f, e) = 1$ if the English term e is one of the translations of the French term f and $\delta(f, e) = 0$ otherwise.

Probabilities $\Pr(e \mid e')$ are estimated on the target document collection, through the following bigram LM, that tries to compensate for different word orderings induced by the source and target languages:

$$\Pr(e \mid e') = \frac{\Pr(e, e')}{\sum_{e''} \Pr(e, e'')} \tag{13}$$

where $\Pr(e, e')$ is the probability of e co-occurring with e', regardless of the order, within a text window of fixed size. Smoothing of the probability is performed through absolute discounting and interpolation (Federico and De Mori 1998) as follows:

$$\Pr(e, e') = \max\left\{\frac{C(e, e') - \beta}{N}, 0\right\} + \beta \Pr(e) \Pr(e') \tag{14}$$

$C(e, e')$ is the number of co-occurrences appearing in the corpus, $\Pr(e)$ is estimated according to Eq. (5), and the absolute discounting term β is equal to the estimate proposed in Ney et al. (1994):

$$\beta = \frac{n_1}{n_1 + 2n_2} \tag{15}$$

with n_k representing the number of term pairs occurring exactly k times in the corpus.

Given a query-translation model and a query $\mathbf{f}$, the most probable translation $\mathbf{e}^*$ can be computed through the well known Viterbi search algorithm (Rabiner 1990). Moreover, intermediate results of the Viterbi algorithm can be used by an A^* search algorithm (Nilsson 1982) to efficiently compute the N most probable, or N-best, translations of $\mathbf{f}$. A detailed explanation of this procedure, which is a simplified version of the so called tree-trellis based algorithm (Soong and Huang 1991), can be found in Federico and Bertoldi (2002).

4. Bilingual IR

From a statistical perspective, bilingual IR can be formulated as follows. Given a query $\mathbf{f}$, in the source language, one would like measure the relevance of a documents d, in the target language, by a joint probability $\Pr(\mathbf{f}, d)$. To fill the gap of language between query and documents, the hidden variable $\mathbf{e}$ is introduced, which represents a term-by-term translation of $\mathbf{f}$ in the target language. Hence, the following decomposition is derived:

$$\begin{aligned}
\Pr(\mathbf{f}, d) &= \sum_{\mathbf{e}} \Pr(\mathbf{f}, \mathbf{e}, d) \\
&\approx \sum_{\mathbf{e}} \Pr(\mathbf{f}, \mathbf{e}) \Pr(d \mid \mathbf{e}) \\
&= \sum_{\mathbf{e}} \Pr(\mathbf{f}, \mathbf{e}) \frac{\Pr(\mathbf{e}, d)}{\sum_{d'} \Pr(\mathbf{e}, d')}
\end{aligned} \tag{16}$$

In deriving formula (16), one makes the assumption (or approximation) that the probability of document d given query $\mathbf{f}$ and translation $\mathbf{e}$, does not depend on $\mathbf{f}$. Formula (16) contains probabilities $\Pr(\mathbf{e}, d)$ and $\Pr(\mathbf{f}, \mathbf{e})$, which correspond, respectively, to the query-document and query-translation models described in the previous sections. In figure 1 a scheme of the resulting CLIR architecture is depicted. In particular, also the required data to train the two models are shown.

In principle, the probability $\Pr(\mathbf{f}, d)$ results very expensive to compute. In fact, the main summation in (16) is taken over the set of possible translations of $\mathbf{f}$. As terms of $\mathbf{f}$ may typically admit more than one translation, the size of this set can grow exponentially with the length of $\mathbf{f}$. For instance, the Italian-English dictionary, used for our experiments, returns on average 1.84 English words for each Italian entry. Hence, the number of possible translations for a 40 word long query is in the order of 10^{10}! Finally, the denominator in formula (16) requires summing over all document in $\mathcal{D}$ and should be computed for every possible translation $\mathbf{e}$.

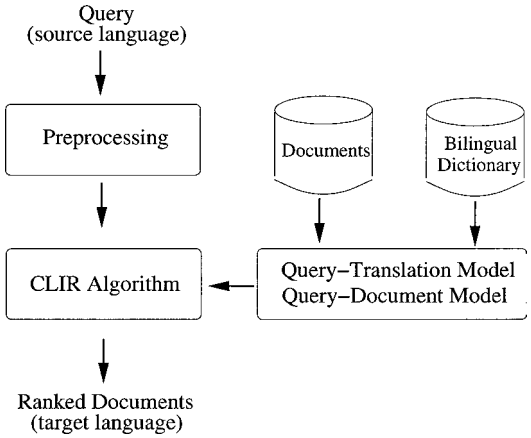

Figure 1. CLIR system architecture.

The derivation of formula (16) is of course not unique. Other types of statistical models for CLIR have been derived in the literature (Hiemstra and de Jong 1999, Xu et al. 2001, Berger and Lafferty 1999). A comparative discussion of these models with respect to the one presented here can be found in Federico and Bertoldi (2002). Non statistical models for CLIR, dealing with multiple translations, are instead discussed in Pirkola (1998) and Ballesteros and Croft (1998).

Now, two algorithms are introduced which approximate, with increasing accuracy, the computation of formula (16).

4.1. Cascade approach

A method to cope with the complexity of (16), is to apply the following maximum approximation:

$$\begin{aligned}
\Pr(\mathbf{f}, d) &= \sum_{\mathbf{e}} \Pr(\mathbf{f}, \mathbf{e}) \frac{\Pr(\mathbf{e}, d)}{\sum_{d'} \Pr(\mathbf{e}, d')} \\
&\approx \max_{\mathbf{e}} \left\{ \Pr(\mathbf{f}, \mathbf{e}) \frac{\Pr(\mathbf{e}, d)}{\sum_{d'} \Pr(\mathbf{e}, d')} \right\} \\
&\approx \Pr(\mathbf{f}, \mathbf{e}^*) \frac{\Pr(\mathbf{e}^*, d)}{\sum_{d'} \Pr(\mathbf{e}^*, d')} : \mathbf{e}^* = \arg\max_{\mathbf{e}} \Pr(\mathbf{f}, \mathbf{e}) \\
&\propto \Pr(\mathbf{f}, \mathbf{e}^*)\ \Pr(\mathbf{e}^*, d) \qquad (17)
\end{aligned}$$

This approximation permits to decouple the translation and retrieval phases. Given a query $\mathbf{f}$, the Viterbi decoding algorithm is applied to compute the most probable translation $\mathbf{e}^*$, as explained in Section 3. Then, the document collection is searched by applying any

Table 4. Algorithm of the CLIR cascade approach.

1. Input $\mathbf{f}$
2. Compute the best translation of $\mathbf{f}$: $\mathbf{e}^* = \arg\max_{\mathbf{e}} Pr[\mathbf{f}, \mathbf{e}]$
3. Order documents according to $P[\mathbf{e}^*, d]$

monolingual IR model, explained in Section 2, with the query $\mathbf{e}^*$. Table 4 shows the algorithm for the cascade approach.

4.2. Integrated approach

A more refined algorithm is now presented that relies on two approximations in order to limit the set of possible translations and documents to be taken into account in formula (16).

Approximation 1. The first approximation redefines the query-translation probability by limiting its support set to just the N-best translations of $\mathbf{f}$, indicated by $\mathcal{T}_N(\mathbf{f})$. Hence,

$$\Pr'(\mathbf{f}, \mathbf{e}) = \begin{cases} \dfrac{\Pr(\mathbf{f}, \mathbf{e})}{K_1(\mathbf{f})} & \text{if } \mathbf{e} \in \mathcal{T}_N(\mathbf{f}) \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

$K_1(\mathbf{f})$ is a normalization term which can be disregarded in formula (16) for the sake of document ordering, as being constant with respect to the ranking variable d.

Approximation 2. A second approximation is introduced to reduce the computational burden of the denominator on Eq. (16). Hence, the support set of the query-document model is limited to only documents which contain at least one term of the query. Given a translation $\mathbf{e}$, let $\mathcal{I}(\mathbf{e})$ indicate the set of documents containing terms of $\mathbf{e}$. This set is easy to compute when the collection is accessed through an inverted index (Frakes and Baeza-Yates 1992). Hence,

$$\Pr'(\mathbf{e}, d) = \begin{cases} \dfrac{\Pr(\mathbf{e}, d)}{K_2(\mathbf{e})} & \text{if } d \in \mathcal{I}(\mathbf{e}) \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

where $K_2(\mathbf{e})$ is a normalization term that occurs both in the numerator and denominator of the fraction in (16), and is therefore deleted. Thanks to this approximation, computation of the denominator in formula (16) can be performed by summing up the scores of just the documents accessed through the inverted index.

The CLIR algorithm applying the two approximations is shown in Table 5. Briefly, given an input query $\mathbf{f}$, the N-best translations are computed first. Then, for each translation $\mathbf{e}$, the addenda in formula (16) are computed only for documents containing at least one term of $\mathbf{e}$. This requires one additional loop over the documents in order to compute the normalization term. The complexity of the algorithm can be estimated as follows:

Table 5. Algorithm of the CLIR integrated approach.

1.	Input $\mathbf{f}$
2.	Compute $\mathcal{T}_N(\mathbf{f})$ with scores $P[\mathbf{f}, \mathbf{e}]$
3.	For each $\mathbf{e} \in \mathcal{T}_N(\mathbf{f})$
4.	$N = 0$
5.	For each $d \in \mathcal{I}(\mathbf{e})$
6.	Compute $P[\mathbf{e}, d]$
7.	Update $N = N + P[\mathbf{e}, d]$
8.	For each $d \in \mathcal{I}(\mathbf{e})$
9.	Update $P[\mathbf{f}, d] = P[\mathbf{f}, d] + P[\mathbf{e}, d] * P[\mathbf{f}, \mathbf{e}]/N$
10.	Order documents according to $P[\mathbf{f}, d]$

- $\mathcal{O}(n\ \bar{\mathcal{E}}^2 + n\ N^2\ \bar{\mathcal{E}})$ for step 2 in the average case (Federico and Bertoldi 2002)
- $\mathcal{O}(n\ N\ \bar{\mathcal{I}})$ for steps 3–9 in the average case
- $\mathcal{O}(n\ \bar{\mathcal{E}}\ \bar{\mathcal{I}}\ \log(n\ \bar{\mathcal{E}}\ \bar{\mathcal{I}}))$ for step 10, in the worst case, i.e. N-best translations use all the available terms,

where n denotes the length of the query, N the number of generated translations, $\bar{\mathcal{E}}$ is the average number of translations of a term, and $\bar{\mathcal{I}}$ is the average number of documents spanned by the inverted file index. The latter number is somehow controlled by the stop-term removal phase applied during document indexing. Generally, terms occurring in many documents are not considered significant for IR and are removed from the index. For instance, in the performed Italian-English experiments we had $\bar{\mathcal{I}} \approx 110$ with $|\mathcal{D}| = 110,282$, and $\bar{\mathcal{E}}$=1.84.

Remark. It is worth noticing that the cascade approach is a special case of the integrated approach, which results by taking $N = 1$. The cascade method permits indeed to eliminate the normalization term in Eq. (16).

4.3. Query/Document preprocessing: Multi-words

When translating, phrasal verbs, noun phrases, compounds, have to be recognized and correctly transferred to the target language. As dictionaries typically contain many multi-word entries, these were included in the statistical translation model. Besides including them into the lexicon probabilities, co-occurrences of multi-words were also collected in the target LM. Multi-words were indeed not considered for the sake of indexing and retrieval. Hence, after the translation step they were split into single words.

4.4. Out-of-dictionary words

The query-translation model relies on commercial Italian-English dictionary of about 51 K translation pairs. On the average, each Italian term is translated by the dictionary into 1.84

English words, and vice-versa each English term has 1.68 Italian translations. The estimated coverage of the dictionary with respect to the query terms is 89.7% for Italian and 90.2% for English, including numbers which are translated verbatim.

Translation coverage was artificially augmented by applying proper name recognition on the original query and by forcing verbatim translation of proper names (with English version stemmed) which do not occur in the dictionary. The addition of proper name recognition increased translation coverage to 94.6% for Italian words and 96.1% for English words. However, for proper names, guarantee about the correctness of the translation is lost; in fact, names of people are usually written with the same transliteration both in Italian and in English, but names of locations and organizations often differ. An example is the name *Chechnya*, which in Italian is *Cecenia*. Finally, it is worth mentioning that proper names and numbers are excluded from the computation of the N-best translations, but are just added to them afterwards.

4.5. Blind relevance feedback

In CLIR, BRF is not used to expand the original query but its translations. In order to save computation time, BRF is performed on the N-best translation as a whole, as Eq. (16) suggests. Hence, relevant terms of the top ranking documents are added to all N-best translations, without modifying their probabilities. Parameter setting of BRF was the same as for Italian monolingual IR.

5. Experimental evaluation

This section reports performance of the presented IR systems on the CLEF tracks. In particular, monolingual IR experiments were carried out on Italian, while CLIR was performed from Italian to English and vice versa. Reported performance are in terms of mean average precision `mAvPr`, as done in CLEF. Statistical significance in `mAvPr` differences, over the same set of queries, are computed with the paired sign test (Johnson and Wichern 1992). In particular, the test is applied to paired average precision measures of single queries, by testing for a median difference of zero.

5.1. CLEF benchmark

CLEF tracks consists of collections of document and sets of topics, in different languages, and relevance assessments for every pair of document collection and set of topics (see Brachler and Peters, this volume). Three document collections of CLEF have been used here:

- an English collection (EC) consisting of 110,282 documents, from *Los Angeles Times* and issued in 1994;
- an Italian collection (IC1) including 58,051 documents, from *La Stampa* and issued in 1994 (used in CLEF 2000);

Table 6. Upper table: total numbers of documents and running words for each considered collection. Lower table: statistics about each pair of topic set and collection: i.e. number of topics with relevant documents in the collection, and total number of relevant documents in the collection.

	Collection		
	EC	IC1	IC2
Documents	110,282	58,051	108,578
Size	425 MB	193 MB	278 MB

	Collection					
	EC		IC1		IC2	
	Topics	rel.docs	Topics	rel.docs	Topics	rel.docs
Q1	33	579	34	338		
Q2	47	856			47	1246
Q3					49	1072

- an Italian collection (IC2) including IC1 and other 50,527 documents from the *Swiss News Agency*, issued in 1994, for a total of 108,578 documents (used in CLEF 2001 and CLEF 2002).

Topics consist in three fields (Title, Descriptive, Narrative), as shown in Table 2. Available topics are 40 for CLEF 2000, 50 for CLEF 2001, and 50 for CLEF 2002, for a total of 140. Table 6 reports statistics about document collections and topics. Notice that topics which do not have relevant documents in a given collection were removed from the corresponding track.

In the following, topics used for CLEF 2000, CLEF 2001, and CLEF 2002 will be referred to by Q1, Q2 and Q3, respectively. The language of each collection is indicated by the prefix of its name (I for Italian, E for English), while that of the queries depends on the considered track. Experiments were carried out using both short (TD) and long (TDN) topics.

5.2. Monolingual IR results

Table 7 reports performance achieved on each set of Italian topics and their union as well.

It can be noticed that performance on long topics is significantly better than on short ones. This is mainly due to the different number of content words which is available to search documents. Moreover, figures show that query expansion is very effective. Relative improvements due to BRF are between 8% and 22% in the case of TD topics and between 5% and 10% for TDN topics. More precisely, performance improvements on the whole set of topics (Q1-2-3, both TD and TDN) result significant at level $p \geq 0.986$. It is worth noticing that BRF results more effective with the statistical LM approach than with the Okapi one.

Table 7. Mean average precision results for monolingual IR, with different sets of Italian topics, topic types (TD vs. TDN), document collections (IC1, IC2, and both), and three retrieval models, each either with or without query expansion.

Topics									
Set	Type	Lang	Coll.	Statistical	+BRF	Okapi	+BRF	Combined	+BRF
Q1	TD	IT	IC1	.3671	.4481	.4215	.4551	.4110	**.4556**
Q1	TDN	IT	IC1	.4447	.4941	.4920	**.5198**	.4722	.5152
Q2	TD	IT	IC2	.4141	.4662	.4449	.4815	.4379	**.4883**
Q2	TDN	IT	IC2	.4372	.4847	.4664	.4939	.4625	**.5041**
Q3	TD	IT	IC2	.3862	.4656	.4058	.4703	.4042	**.4920**
Q3	TDN	IT	IC2	.4453	.5271	.4432	.5028	.4516	**.5304**
Q1-2-3	TD	IT	IC1-2	.3913	.4612	.4240	.4704	.4182	**.4811**
Q1-2-3	TDN	IT	IC1-2	.4422	.5031	.4644	.5040	.4609	**.5169**

A direct comparison between LM and Okapi shows that the latter performs slightly better, but differences become smaller after BRF. The last two columns of Table 7 report `mAvPr` results of the combined scoring model. Respectively, the columns correspond to the combination of scores taken before and after BRF on the single models. Figures show that after query expansion, the combined model, but in one case, improves over the best of the two single methods. Over the complete set of queries, relative improvements in mean-average precision over the best performing model are of 2.3% ($p \geq 0.984$) for short topics, and 2.6% ($p \geq 0.986$) for long topics.

5.3. Bilingual IR results

CLIR experiments were performed from Italian to English and in the opposite direction. In the CLEF evaluation campaigns, Italian to English tracks used topics Q1 and Q2, whereas English to Italian tracks used all sets of topics. Results of these runs are reported for each language direction in Tables 8 and 9, respectively. In all tracks, the combined method was used for the query-document model. Results are provided both for short and long topics, for each set of topics, and for their union. It is worth noticing that the Italian target collection changed between the first and the second CLEF campaign.

By looking at the results corresponding to different numbers of employed translations, it seems, at least on the average, that using more than one translation slightly improves performance. However, this conclusion is not confirmed from a statistical point of view. Only for the Italian to English task (after BRF, TDN topics), a significant difference in `mAvPr` between 5-best translations and 1-best translations was observed at level $p \geq .998$.

Considerations about query expansion, as stated for monolingual IR, are fully confirmed by the CLIR experiments.

Further experiments were carried out to evaluate the query-translation model. In particular, CLIR experiments were performed by using query translations computed by the Viterbi

Table 8. Mean average precision results of Italian-English CLIR with the combined model. Experiments consider different sets of topics, topic types, always in Italian, one English target collection, and different numbers of N-best translations (1,5, and 10). Retrieval performance are reported either with or without blind relevance feedback.

Topics									
Set	Type	Lang	Coll.	1-best	+ BRF	5-best	+ BRF	10-best	+ BRF
Q1	TD	IT	EC	.3287	**.3463**	.3271	.3366	.3277	.3307
Q1	TDN	IT	EC	.3917	.4096	.3864	**.4391**	.3863	.4188
Q2	TD	IT	EC	.4593	.5035	.4537	**.5196**	.4532	.5128
Q2	TDN	IT	EC	.4934	.5132	.4977	**.5255**	.4737	.5226
Q1-2	TD	IT	EC	.4054	.4387	.4014	**.4441**	.4014	.4379
Q1-2	TDN	IT	EC	.4514	.4705	.4518	**.4899**	.4376	.4798

Table 9. Mean average precision results of English-Italian CLIR with the combined model. Experiments consider different sets of topics, topic types, always in English, different Italian target collections (IC1, IC2, and both), and numbers of N-best translations (1,5, and 10). Retrieval performance are reported either with or without blind relevance feedback.

Topics									
Set	Type	Lang	Coll.	1-best	+ BRF	5-best	+ BRF	10-best	+ BRF
Q1	TD	EN	IC1	.3125	**.3382**	.3192	.3339	.3068	.3180
Q1	TDN	EN	IC1	.3604	.3922	.3748	.4042	.3727	**.4135**
Q2	TD	EN	IC2	.3829	.4624	.3828	.4544	.3881	**.4691**
Q2	TDN	EN	IC2	.4156	.4851	.4184	.4849	.4235	**.4855**
Q3	TD	EN	IC2	.2993	.3444	.3086	.3531	.3161	**.3552**
Q3	TDN	EN	IC2	.3646	.4286	.3712	**.4410**	.3757	.4247
Q1-2-3	TD	EN	IC1-2	.3330	.3854	.3382	.3847	.3397	**.3866**
Q1-2-3	TDN	EN	IC1-2	.3819	.4395	.3892	**.4472**	.3922	.4438

search algorithm, by a commercial state-of-the-art machine translation system, and, finally, by a human. In the second case, the Babelfish translation service, powered by Systran[1] was used. As Systran is supposed to work on fluent texts, preprocessing and translation steps were inverted in this case. As human translations, the topics in the documents' language were used, as provided by CLEF.

Given all topic translations, the CLIR algorithm for the 1-best case was applied. Results for Italian-English and English-Italian IR are reported in Tables 10 and 11, respectively.

Remarkably, the statistical query-translation method outperforms the Systran translation system on the union sets of topics. Significant differences between the two translation methods could only be measured on the English to Italian retrieval task. Differences were significant at level $p \geq 0.96$ on short topics, and at level $p \geq 0.76$ on long topics.

From both Tables 10 and 11 it is evident that IR results with 1-best translations shows more oscillations around the global `mAvPr` value computed over the union sets of topics.

Table 10. Mean average precision results of Italian-English CLIR with the combined model including query expansion. Experiments consider different sets of Italian topics, topic types, one English collection, and different kinds of translations: computed by Systran, the 1-best statistical model, and human made.

Topics				Translation		
Set	Type	Lang	Coll.	Systran	1-best	Human
Q1	TD	IT	EC	**.4007**	.3463	.4866
Q1	TDN	IT	EC	**.4565**	.4096	.5029
Q2	TD	IT	EC	.3900	**.5035**	.5559
Q2	TDN	IT	EC	.4786	**.5132**	.5703
Q1-2	TD	IT	EC	.3944	**.4387**	.5273
Q1-2	TDN	IT	EC	.4695	**.4705**	.5425

Table 11. Mean average precision results of English-Italian CLIR with the combined model including query expansion. Experiments consider different sets of English topics, topic types, different document collections, and different kinds of translations: computed by Systran, the 1-best statistical model, and human made.

Topics				Translation		
Set	Type	Lang	Coll.	Systran	1-best	Human
Q1	TD	EN	IC1	.3378	**.3382**	.4556
Q1	TDN	EN	IC1	.3781	**.3922**	.5152
Q2	TD	EN	IC2	.3637	**.4624**	.4883
Q2	TDN	EN	IC2	.3872	**.4851**	.5041
Q3	TD	EN	IC2	**.4037**	.3444	.4920
Q3	TDN	EN	IC2	**.4412**	.4286	.5304
Q1-2-3	TD	EN	IC1-2	.3720	**.3854**	.4811
Q1-2-3	TDN	EN	IC1-2	.4052	**.4395**	.5196

To investigate this issue, standard deviations of the average precision were computed over the whole set of topics, for each experimental condition. On the Italian to English track, standard deviations with TD topics were .314 and .298, respectively, for 1-best and Systran translations. On TDN topics, standard deviations were exactly the same, .301 for both translation methods. On the English to Italian track, 1-best translations seem to cause even less variability than the Systran ones: on TD topics, standard deviations of .323 and .331 were respectively measured, while on TDN topics the corresponding standard deviations were .316 and .329.

Unfortunately, these measurements confirm the difficulty of finding some statistically meaningful explanation of the different `mAvPr` behavior of the tested systems over the single sets of topics.

A problem in translating topics is that some random noise is introduced in the retrieval process. Erroneous translations of content words may indeed severely affect retrieval

performance and, in general, the loss in performance is not strictly related to the number of translation errors.

An indication about the noise introduced by the translation process comes from the lower standard deviations which can be measures on the retrieval results with human translations: .287 for TD topics and .279 for TDN topics, in the Italian-to-English track, and .300 for TD topics and .289 on TDN topics, in the English-to-Italian track. Hence, in general, automatic translation increases uncertainty in `mAvPr`, which can be quantified in 4%–10% relative increase of standard deviation.

In our statistical model, the chance of correctly translating a content word, in a given context, depends on several nested events: the dictionary contains the word, the right translation is among the ones available for that word, and, finally, the correct one is selected. In the following, a qualitative analysis of translation errors is presented.

5.4. Qualitative evaluation

A qualitative analysis of results was carried out to better understand possible weak points of the statistical query translation method. Differences in average precision (`AvPr`) achieved on each single topic were computed. The resulting plots are shown for both translation directions in figures 2 and 3, respectively. More specifically, results refer to the combined model, using short topics and no BRF.

It results that the two translation approaches achieve similar performance for most of the topics, in fact, only 10% of them show `AvPr` differences higher than 0.4. Hence, a

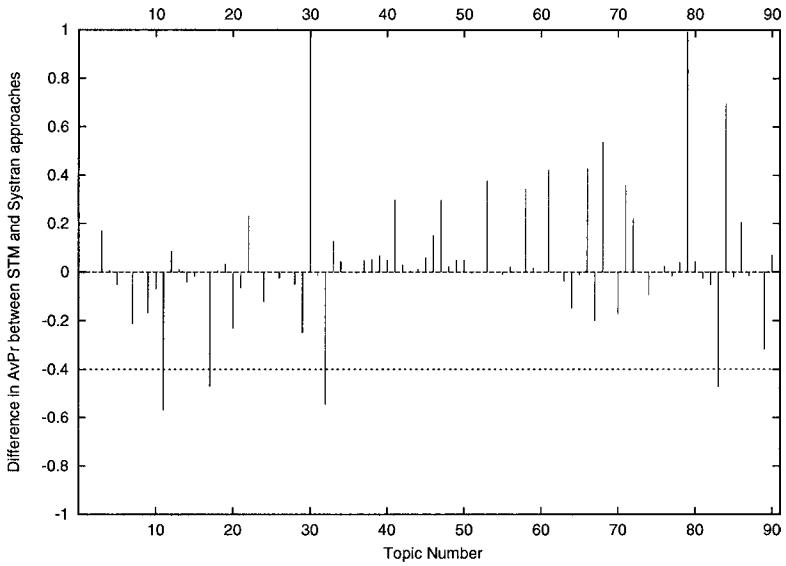

Figure 2. Differences in average precision corresponding to the same CLIR system using translations computed either by the statistical translation model (STM) or by a commercial machine translation system. Topics are in Italian, documents in English.

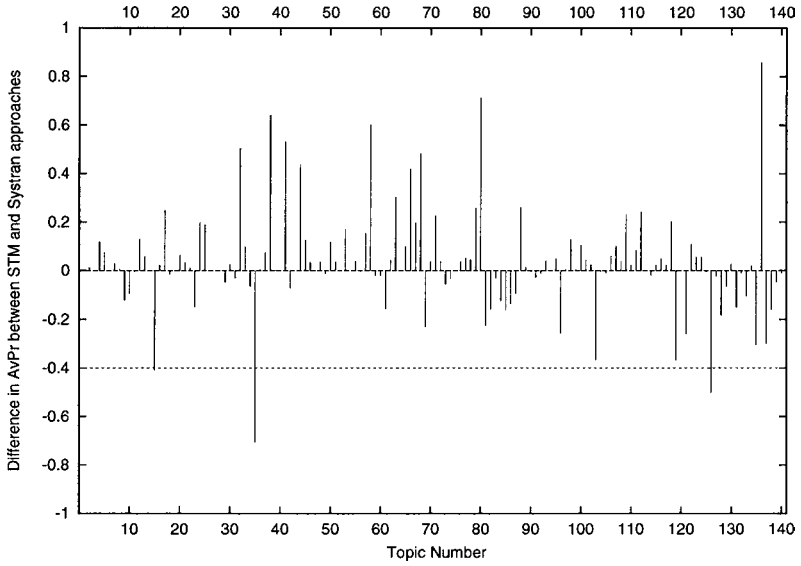

Figure 3. Differences in average precision corresponding to the same CLIR system using translations computed either by the statistical translation model (STM) or by a commercial machine translation system. Topics are in English, documents in Italian.

more detailed analysis was made on the subset of topics on which Systran translations performed significantly better. For Italian to English CLIR, the topics 11, 17, 32, and 83 were considered, while for English to Italian CLIR the topics 15, 35, and 126 were analyzed.

Poor performance generally resulted from translation errors of content words. Translation errors were either caused by a wrong analysis at the POS tagging or word stemming levels, or by coverage failures of the bilingual dictionaries. For instance, the Italian word *preti* (*priests*) was not correctly transformed into its singular form *prete* (*priest*); similarly, the English word *wolves* was wrongly stemmed as *wolv*, instead of *wolf*. Hence, in both cases the corresponding entries were not found in the bilingual dictionary. Bad retrieval also occurred because correct translations appeared with a low rank; e.g. the best Italian translation of *fur* (*pelliccia*) appears for the first time in the translation at rank 22.

6. Discussion and conclusion

Interesting issues emerged from our participation in the CLEF campaigns, which are briefly discussed in the following.

- The statistical LM approach well compares with the Okapi model, and results very competitive on long topics or after query expansion. Moreover, consistent improvement

in performance over the two methods was achieved by combining the Okapi and LM scores after some normalization.

- Comparing CLIR models results quite difficult. As a matter of fact, retrieval performance seems very sensitive to the translation quality, which mainly depends on the coverage of the available dictionaries and on the generation of correct word stems and base-forms. Retrieval performance measured by using our translation model and a commercial translation system showed, over a set of 140 queries, many large fluctuations. In fact, such high variability did not permit to rank the approaches in a statistically significant way, at least on available sets of queries. From our point of view, this also means that, for the sake of IR, our statistical translation model, which is quite simple to implement, did not perform worse than a state-of-the-art commercial translation engine, which was developed over several decades.
- Qualitative analysis of results, suggest that improvements in CLIR should be pursued in two main directions: by developing better statistical CLIR models (see below), and by augmenting coverage of bilingual dictionaries. On the other hand, recent experiments showed that text preprocessing based on morpho-syntactic analysis is not superior than basic word stemming. This widens the applicability of the proposed CLIR approach to other language pairs for which bilingual dictionaries are available. Preliminary experiments (Bertoldi and Federico 2003) were carried out on a cross-language spoken document retrieval track, with spoken documents in English and topics in French, German, Italian, and Spanish. Promising results were achieved by only using publicly available dictionaries and stemming algorithms.
- As concerns with future work, there are some ideas about how to improve the here presented statistical CLIR model.

  First, word translation probabilities, $\Pr(f \mid e)$, could be improved by replacing the currently used uniform probabilities with others estimated from data. Better estimates could be computed on a large text corpus in the target language by means of the EM algorithm, as suggested in Koehn and Knight (2000).

  Other improvements could be achieved by applying more sophisticated statistical machine translation models. In particular, models which introduce word reordering are currently considered. The rationale is that highly correlated word pairs in the query could be found and translated chain-wise. In this way translation, ambiguity might be significantly reduced by exploiting word dependencies modeled by the target language model.
- Finally, it is well known that blind relevance feedback is determinant step which boosts retrieval performance, especially for short queries. However, up to now, not enough effort has been devoted to embed BRF into a statistically sound framework. Besides this theoretical issue, we will also investigate how BRF could be specifically devised for CLIR, for instance, to improve quality of translations.

Acknowledgments

This work was carried out at ITC-irst under the project WebFAQ, which is partially funded by the FDR-PAT program of the Province of Trento.

Note

1. Babelfish translation service is available at http://world.altavista.com

References

Ballesteros L and Croft WB (1998) Resolving ambiguity for cross-language retrieval. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 64–71.

Berger A and Lafferty JD (1999) Information retrieval as statistical translation. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 222–229.

Bertoldi N and Federico M (2001) ITC-irst at CLEF 2000: Italian monolingual track. In: Peters C, Ed. Cross-Language Information Retrieval and Evaluation, vol. 2069 of Lecture Notes in Computer Science, Heidelberg, Germany, Springer Verlag, pp. 261–272.

Bertoldi N and Federico M (2003) Cross-language spoken document retrieval on the TREC SDR collection. In: Peters C et al., Eds. Cross-Language Information Retrieval and Evaluation, Lecture Notes in Computer Science, Heidelberg, Germany, Springer Verlag (to appear).

Federico M (2000) A system for the retrieval of Italian broadcast news. Speech Communication, 32(1/2):37–47.

Federico M and Bertoldi N (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, pp. 167–174.

Federico M and De Mori R (1998) Language modelling. In: Mori RD, Ed. Spoken Dialogues with Computers, chapter 7. Academy Press, London, UK.

Frakes WB and Baeza-Yates R, editors (1992). Information Retrieval: Data Structures and Algorithms. Prentice Hall, Englewood Cliffs, NJ.

Hiemstra D and de Jong F (1999). Disambiguation strategies for cross-language information retrieval. In: Proceedings of the 3rd European Conference on Research and Advanced Technology for Digital Libraries, pp. 274–293.

Johnson RA and Wichern DW, Eds. (1992) Applied Multivariate Statistical Analysis. Prentice Hall, Englewood Cliffs, NJ.

Johnson S, Jourlin P, Jones KS and Woodland P (1999) Spoken document retrieval for TREC-8 at Cambridge University. In: Proceedings of the 8th Text REtrieval Conference, Gaithersburg, MD, pp. 197–206.

Koehn P and Knight K (2000) Estimating word translation probabilities from unrelated monolingual corpora using the EM algorithm. In: *AAAI/IAAI*, pp. 711–715.

Miller, DRH, Leek T and Schwartz RM (1998) BBN at TREC-7: Using hidden Markov models for information retrieval. In: Proceedings of the 7th Text REtrieval Conference, Gaithersburg, MD, pp. 133–142.

Mood AM, Graybill FA and Boes DC (1974) Introduction to the Theory of Statistics. McGraw-Hill, Singapore.

Ney H, Essen U and Kneser R (1994) On structuring probabilistic dependences in stochastic language modelling. Computer Speech and Language, 8:1–38.

Ng K (1999) A maximum likelihood ratio information retrieval model. In: Proceedings of the 8th Text REtrieval Conference, Gaithersburg, MD, pp. 483–492.

Nilsson NJ (1982) Principles of Artificial Intelligence. Springer Verlag, Berlin, Germany.

Pirkola A (1998) The effect of query structure and dictionary setups in dictionary-based cross-language information retrieval. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 55–63.

Porter MF (1980) An algorithm for suffix stripping. Program, 14(3):130–137.

Rabiner LR (1990) A tutorial on hidden Markov models and selected applications in speech recognition. In: Weibel A and Lee K, Eds. Readings in Speech Recognition, Morgan Kaufmann, Los Altos, CA, pp. 267–296.

Robertson SE, Walker S, Jones S, Hancock-Beaulieu MM and Gatford M (1994) Okapi at TREC-3. In: Proceedings of the 3rd Text REtrieval Conference, Gaithersburg, MD, pp. 109–126.

Soong FK and Huang EF (1991) A tree-trellis based fast search for finding the n-best sentence hypotheses in continuos speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, Toronto, Canada, pp. 705–708.

Witten IH and Bell TC (1991) The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. IEEE Transaction on Information Theory, IT-37(4):1085–1094.

Xu J, Weischedel R and Nguyen C (2001) Evaluating a probabilistic model for cross-lingual information retrieval. In: Proceedings of the 24h Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 105–110.