ITC-irst at CLEF 2001: Monolingual and Bilingual Tracks

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 (CLEF) of 2001. ITC-irst has taken part to two tracks: the monolingual retrieval task, and the bilingual retrieval task. In both cases, Italian was chosen as the query language, while English was chosen as the document language of the bilingual task. The employed retrieval engine combines scores computed by an Okapi model and a statistical language model. The cross language system employes a statistical query translation model, which is estimated on the target document collection and on a translation dictionary.

1. Introduction

This paper reports on the participation of ITC-irst in two Information Retrieval (IR) tracks of the Cross Language Evaluation Forum (CLEF) 2001: the monolingual retrieval task, and the bilingual retrieval task. The language for the queries was always Italian, and English documents were searched for in the bilingual task. With respect to the 2000 CLEF evaluation (Bertoldi and Federico, 2000), the monolingual IR system was just slightly refined, while most of the effort was dedicated to develop an original cross-language IR system.

The basic IR engine, used for both evaluations, combines scores of a standard Okapi model and of a statistical language model. For cross-language IR, a light-weight statistical model for translating queries was developed, which does not need any parallel or comparable corpora to be trained, but just the target document collection and a bilingual dictionary.

This paper is organized as follows. In Section 2, the employed text pre-processing modules are presented. Section 3 describes the employed IR models, Section 4 introduces the cross-language specific models, namely the query translation model and the retrieval model. Section 5 presents the official evaluation results. Finally, Section 6 gives some conclusions.

2. Text Pre-Processing

Text pre-processing is performed in several stages, which may differ according to the task and language. In the following a list of modules used to pre-process documents and queries is given, by also specifying to which languages they apply.

2.1. Tokenization - IT+EN

Text tokenization is performed in order to isolate words from punctuation marks, recognize abbreviations and acronyms, correct possible word splits across lines, and discriminate between accents and quotation marks.

2.2. Morphological analysis - IT

A morphological analyzer decomposes each Italian inflected word into its morphemes, and suggests all possible POSs and base forms of each valid decomposition. By base forms we mean the usual not inflected entries of a dictionary.

2.3. POS tagging - IT

Words in a text are tagged with parts-of-speech (POS) by computing the best text-POS alignment through a statistical model. The employed tagger works with 57 tag classes and has an accuracy around 96%.

2.4. Base form extraction - IT

Once the POS and the morphological analysis of each word in the text is computed, a base form can be assigned to each word.

2.5. Stemming - EN

Word stemming is just performed on English words by using the Porter's algorithm.

2.6. Stop-terms removal - IT+EN

Words that are not considered relevant for IR are discarded in order to save index space. Words are filtered out on the basis either of their POS (if available) or their inverted document frequency.

2.7. Multi-word recognition - EN

Multi-words are just used for the sake of the query translation. Hence, the statistics used by the translation models do contain multi-words. After translation, multi-words are split into single words.

Q, T, D	random variables of query, translation, and document
q, t, d	instances of query, query translation, and document
w, i, e	generic term, Italian term, English term
$\mathcal{D}$	collection of documents
$\mathcal{V}$,$\mathcal{V}(d)$	set of terms occurring in $\mathcal{D}$, and in document d
$N, N(d)$	number of term occurrences in $\mathcal{D}$, and in a document d
$N(w), N(d,w), N(q,w)$	frequency of term w in $\mathcal{D}$, in document d, and in query q
N_w	number of documents in $\mathcal{D}$ which contain term w
$\lvert \cdot \rvert$	size of a set

Table 1: List of often used symbols.

3. Information Retrieval Models

3.1. Okapi Model

To score the relevance of a document d versus a query q, the following Okapi weighting function is applied:

$$s(d) = \sum_{w \in q \cap d} N(q,w)\, W_d(w)\, idf(w) \qquad (1)$$

where:

$$W_d(w) = \frac{N(d,w)(k_1+1)}{k_1(1-b) + k_1 b \frac{N(d)}{l} + N(d,w)} \qquad (2)$$

scores the relevance of w in d, and the inverted document frequency:

$$idf(w) = \log \frac{N - N_w + 0.5}{N_w + 0.5} \qquad (3)$$

evaluates the relevance of term w inside the collection. The model implies two parameters k_1 and b to be empirically estimated over a development sample. As in previous work, the setting $k_1 = 1.5$ and $b = 0.4$ were used. An explanation of the involved terms can be found in (Robertson et al., 1994) and other papers referred in it.

3.2. Language Model

According to this model, the match between a query random variable Q and a document random variable D is expressed through the following conditional probability distribution:

$$Pr(D \mid Q) = \frac{Pr(Q \mid D) Pr(D)}{Pr(Q)} \qquad (4)$$

where $Pr(Q \mid D)$ represents the likelihood of Q, given D, $Pr(D)$ represents the a-priori probability of D, and $Pr(Q)$ is a normalization term. By assuming a uniform a-priori probability distribution about the documents, and disregarding the normalization factor, documents can be ranked, with respect to Q, just by the likelihood term $Pr(Q \mid D)$. If we assume an order-free multinomial model, the likelihood is:

$$Pr(Q = w_1, \ldots, w_n \mid D = d) = \prod_{k=1}^{n} Pr(w_k \mid d) \qquad (5)$$

The probability that a term w is generated by d can be estimated by a statistical language model (LM). Previous work on statistical information retrieval (Miller et al., 1998; Ng, 1999) proposed to interpolate relative frequencies of each document with those of the whole collection, with interpolation weights empirically estimated from the data.
In this work we use an interpolation formula which applies the smoothing method proposed by (Witten and Bell, 1991). This method linearly smoothes word frequencies of a document, and the amount of probability assigned to never observed terms is proportional to the number of different words contained in the document, i.e.:

$$Pr(w \mid d) = \frac{N(d,w)}{N(d) + |\mathcal{V}(d)|} + \frac{|\mathcal{V}(d)|}{N(d) + |\mathcal{V}(d)|} P(w) \qquad (6)$$

where $Pr(w)$, the word probability over the collection, is estimated by interpolating the smoothed relative frequency with the uniform distribution over the vocabulary $\mathcal{V}$:

$$Pr(w) = \frac{N(w)}{N + |\mathcal{V}|} + \frac{|\mathcal{V}|}{N + |\mathcal{V}|} \frac{1}{|\mathcal{V}|} \qquad (7)$$

3.3. Combined model

Previous work (Bertoldi and Federico, 2000) showed that Okapi and the statistical model 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. Actually, in order to adjust scale differences, scores of each model are normalized in the range $[0, 1]$ before summation.

3.4. Blind Relevance Feedback

Blind relevance feedback (BRF) is a well known technique that allows to improve retrieval performance. The basic idea is to perform retrieval in two steps. First, the documents matching the original query q are ranked, then the B best ranked documents are taken and the R most relevant terms in them are added to the query. Hence, the retrieval phase is repeated with the augmented query. In this work, new

search terms are extracted by sorting all the terms of the B top documents according to (Johnson et al., 1999):

$$r_w \frac{(r_w + 0.5)(N - N_w - B + r_w + 0.5)}{(N_w - r_w + 0.5)(B - r_w + 0.5)} \quad (8)$$

where r_w is the number of documents, among the top B documents, which contain word w. In all the performed experiments the values $B = 5$ and $R = 15$ were used.

4. Cross-language IR Model

4.1. Query Translation Model

Query translation is based on a *hidden Markov model* (HMM) (Rabiner, 1990), in which the observable part is the query Q in the source language (Italian), and the hidden part is the corresponding query T in the target language (English). Hence, the joint probability of a pair Q, T can be decomposed as follows:

$$Pr(Q = i_1, \ldots, i_n, T = e_1, \ldots, e_n) = \prod_{k=1}^{n} Pr(i_k \mid e_k) Pr(e_k \mid e_{k-1}) \quad (9)$$

Given a query $Q = i_1, \ldots, i_n$ and estimates of the discrete distributions in the right side of equation (9), the most probable translation $T^* = e_1^*, \ldots, e_n^*$ can be determined through the well known Viterbi algorithm (Rabiner, 1990). Probabilities $Pr(i \mid e)$ are estimated from a translation dictionary as follows:

$$Pr(i \mid e) = \frac{\delta(i, e)}{\sum_{i'} \delta(i', e)} \quad (10)$$

where $\delta(i, e) = 1$ if the English term e is one of the translations of Italian term i and $\delta(i, e) = 0$ otherwise. For the CLEF evaluation an Italian-English dictionary of about 45K entries was used.
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:

$$Pr(e \mid e') = \frac{Pr(e, e')}{\sum_{e''} Pr(e, e'')} \quad (11)$$

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 as follows:

$$Pr(e, e') = \max\{\frac{C(e, e') - \beta}{N}, 0\} + \beta Pr(e) Pr(e') \quad (12)$$

$C(e, e')$ is the number of co-occurrences appearing in the corpus, $Pr(e)$ is estimated according to equation (7), and the absolute discounting term β is equal to the estimate proposed in (Ney et al., 1994):

$$\beta = \frac{n_1}{n_1 + 2n_2} \quad (13)$$

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

4.2. Cross-Language IR Model

As a first method to perform cross-language retrieval, a simple plug-in method was devised, which decouples the translation and retrieval phases. Hence, given a query Q in the source language, the Viterbi decoding algorithm is applied to compute the most probable translation T^* in the target language, according to the statistical query translation model explained above. Then, the document collection is searched by applying a conventional monolingual IR method.

1.	Find the best translation of query Q: $T^* = \arg\max_T Pr(Q, T)$
2.	Order documents by using the translation T^*

Table 2: Plug-in method for cross-language IR.

5. Evaluation

5.1. Monolingual Track

Two monolingual runs were submitted to the Italian monolingual track. The first run used all the information available for the topics, while the second one just the title and description parts. The track consisted of 47 topics, for a total of 1,246 documents to be retrieved inside a collection of 108,578 documents.
A detailed description of the used system follows now:

- Document/query pre-processing: tokenization, POS tagging, base form extraction, stop-term removal.
- Retrieval step 1: separate Okapi and LM runs.
- BRF: performed on each model output.
- Retrieval step 2: same as step 1 with the expanded query.
- Final rank: sum of Okapi and LM normalized scores.

Results of the submitted runs are given in Table 3.

5.2. Bilingual IR Evaluation

Two runs were submitted to the Italian-to-English bilingual track, with the same modalities of the monolingual track. The bilingual track consisted of 47 topics, for a total of 856 documents to be retrieved inside a collection of 110,282 documents. A detailed description of the used system follows now:

Retrieval Mode	Official Run	mAvPr
title+desc+narr	IRSTit1	48.59
title+desc	IRSTit2	46.44

Table 3: Results on the Italian monolingual tracks.

Retrieval Mode	Official Runs	mAvPr
title+desc+narr	IRSTit2en1	42.51
title+desc	IRSTit2en2	34.11

Retrieval Mode	Non Official Runs	mAvPr
title+desc+narr	Babelfish	44.53
title+desc	Babelfish	37.99

Table 4: Results on the Italian-English bilingual tracks.

- Document pre-processing: tokenization, stemming, stop-term removal.
- Query pre-processing: tokenization, POS tagging, base form extraction, stop term removal, translation, multi-words split, stemming.
- Retrieval step 1: separate Okapi and LM runs.
- BRF: performed on each model output.
- Retrieval step 2: same as step 1 with the expanded query.
- Final rank: sum of Okapi and LM normalized scores.

An important issue concerns with the use of multi-words. Multi-words were only used for the target language, i.e. English, and just for the translation process. After translation, multi-words in the query are split again into single words.
As a term of comparison, our statistical query translation model was replaced with the Babelfish text translation service powered by Systran and available on the Internet[1]. Cross-language retrieval performance was measured by keeping all the other components of the system fixed. Results obtained by the submitted runs and by the Babelfish translator are shown in Table 4. The mean average precision achieved with the commercial translation system shows to be about 5%-10% better, depending to the retrieval mode. Detailed results of the experiments are shown in Table 4.

6. Conclusion

In this work we presented the monolingual and cross-language information retrieval systems developed at ITC-irst and evaluated at the CLEF 2001. In particular, the cross-language system uses a statistical query translation algorithm that requires minimal language resources: a bilingual dictionary and the target document collection. Results on the CLEF 2001 evaluation data show that satisfactory performance can be achieved with this simple translation model. However, experience gained from the many performed experiments suggest that a fair comparison between different systems would require a much larger amount of queries. The retrieval performance shows in fact to be very sensitive to the translation step.
Current work is in the direction of further developing the here proposed statistical model for cross-language IR. In particular, significant improvements have been achieved by closely integrating the translation and retrieval models.

[1] http://world.altavista.com

Acknowledgements

The authors would like to thank their colleagues at ITC-irst Bernardo Magnini and Emanuele Pianta for putting at disposal an electronic Italian-English dictionary.

7. References

Bertoldi, N. and M. Federico, 2000. Italian text retrieval for CLEF 2000 at ITC-irst. In *Working notes of CLEF 2000*. Lisbon, Portugal.

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

Miller, David R. H., Tim Leek, and Richard M. Schwartz, 1998. BBN at TREC-7: Using hidden Markov models for information retrieval. In *Proc. of 7th TREC*. Gaithersburg, MD.

Ney, Herman, Ute Essen, and Reinhard Kneser, 1994. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech and Language*, 8:1–38.

Ng, Kenney, 1999. A maximum likelihood ratio information retrieval model. In *Proc. of 8th TREC*. 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.

Robertson, S. E., S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford, 1994. Okapi at TREC-3. In *Proc. of 3rd TREC*. Gaithersburg, MD.

Witten, Ian H. and Timothy C. Bell, 1991. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Trans. Inform. Theory*, IT-37(4):1085–1094.