Italian Text Retrieval for CLEF 2000 at ITC-irst

Nicola Bertoldi and Marcello Federico

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

Abstract

This paper presents work on document retrieval for Italian carried out at ITC-irst. Two different approaches to information retrieval were investigated, one based on the Okapi weighting formula and one based on a statistical model. Development experiments were carried out using the Italian sample of the TREC-8 CLIR track. Performance evaluation was done on the Cross Language Evaluation Forum (CLEF) 2000 Italian monolingual track.

1. INTRODUCTION

This paper reports on Italian text retrieval research that has recently started at ITC-irst. Experimental evaluation was carried out in the framework of the Cross Language Evaluation Forum (CLEF), a text retrieval system evaluation activity coordinated in Europe from 2000, in collaboration with the US National Institute of Standards and Technology (NIST) and the TREC Conferences.

ITC-irst has recently started to develop monolingual text retrieval systems (Sparck Jones and Willett, 1997) for the main purpose of accessing broadcast news audio and video data (Federico, 2000). This paper presents two Italian monolingual text retrieval systems that have been submitted to CLEF 2000: a conventional Okapi derived model, and a statistical retrieval model. After the evaluation, a combined model was also developed that just integrates the scores of the two basic models. This simple and effective model shows a significant improvement over the two single models.

The paper is organized as follows. In Section 2, the text preprocessing of documents and queries is presented. Section 3 and 4 introduce the text retrieval models that were officially evaluated at CLEF and present experimental results. Section 5 discusses improvements on the basic models that were made after the CLEF evaluation. In particular, a combined retrieval model is introduced and evaluated on the CLEF test collection. Finally, Section 6 offers some conclusions regarding the research at ITC-irst in the field of text retrieval.

2. TEXT PREPROCESSING

Document and query preprocessing implies several stages: tokenization, morphological analysis of words, part-of-speech (POS) tagging of text, base form extraction, stemming, and stop-terms removal.

Tokenization. Tokenization of text 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.

Morphological analysis. 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.

POS tagging. POS tagging is based on a Viterbi decoder that computes the best text-POS alignment on the basis of a bigram POS language model and a discrete observation model (Merialdo, 1994). The employed tagger works with 57 tag classes and has an accuracy around 96%.

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

Stemming. Word stemming is applied at the level of tagged base forms. POS specific rules were developed that remove suffixes from verbs, nouns, and adjectives.

Stop-terms removal. Words in the collection that are considered non relevant for the purpose of information retrieval are discarded in order to save index space. Words are filtered out on the basis either of their POS or their inverted document frequency. In particular, punctuation is eliminated together with articles, determiners, quantifiers, auxiliary verbs, prepositions, conjunctions, interjections, and pronouns. Among the remaining terms, those with a low inverted document frequency, i.e. that occur in many different documents, are eliminated.

An example of text preprocessing is presented in Table 8.

$f_d(w)$	frequency of word w in document d
$f_q(w)$	frequency of w in query q
$f(w)$	frequency of w in the collection
f_d	length of document d
f	length of the collection
$\bar{l}$	mean document length
N	number of documents
N_w	number of documents containing w
V_d	vocabulary size of document d
$\bar{V}_d$	average document vocabulary size
V	vocabulary size of the collection

Table 1: Notation used in the information retrieval models.

Terms	Stop	$\bar{l}$	V	$\bar{V}_d$
text	no	225	160K	134
base forms	no	225	126K	129
stems	no	225	101K	126
base forms	yes	103	125K	80
stems	yes	103	100K	77

Table 2: Effect of text preprocessing steps on the mean document length, global vocabulary size, and mean document vocabulary size.

3. INFORMATION RETRIEVAL MODELS

3.1. Okapi Model

Okapi (Robertson et al., 1994) is the name of a retrieval system project that developed a family of weighting functions in order to evaluate the relevance of a document d versus a query q. In this work, the following Okapi weighting function was applied:

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

where:

$$c_d(w) = \frac{f_d(w)(k_1+1)}{k_1(1-b) + k_1 b \frac{f_d}{\bar{l}} + f_d(w)} \qquad (2)$$

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 w inside the collection. The model implies two parameters k_1 and b to be empirically estimated over a development sample. An explanation of the involved terms can be found in (Robertson et al., 1994) and other papers referred in it.

3.2. Statistical Model

A statistical retrieval model was developed based on previous work on statistical language modeling (Federico and De Mori, 1998).

The match between a query q and a document d can be expressed through the following conditional probability distribution:

$$P(d \mid q) = \frac{P(q \mid d) P(d)}{P(q)} \qquad (4)$$

where $P(q \mid d)$ represents the likelihood of q, given d, $P(d)$ represents the a-priori probability of d, and $P(q)$ is a normalization term. By assuming no a-priori knowledge about the documents, and disregarding the normalization factor, documents can be ranked, with respect to q, just by the likelihood term. If we interpret the likelihood function as the probability of d generating q and assume an order-free multinomial model, the following log-probability score can be derived:

$$\log P(q \mid d) = \sum_{w \in q} f_q(w)\ \log\ P(w \mid d) \qquad (5)$$

The probability that a term w is generated by d can be estimated by applying statistical language modeling techniques. 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. Hence, the following probability estimate is applied:

$$P(w \mid d) = \frac{f_d(w)}{f_d + V_d} + \frac{V_d}{f_d + V_d} P(w) \qquad (6)$$

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

$$P(w) = \frac{f(w)}{f + V} + \frac{V}{f + V} \frac{1}{V} \qquad (7)$$

3.3. 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

Data Set	# docs	Avg. # words/ doc
CLIR - *Swiss News Agency*	62,359	225
CLEF - *La Stampa*	58,051	552

Table 3: Development and test collection sizes.

	# of Words			
Data Set (topic #s')	Min	Max	Avg.	Total
CLIR (54-81)	41	107	70.4	1690
title	3	8	5.1	122
description	8	27	17.1	410
narrative	25	81	48.3	1158
CLEF (1-40)	31	96	60.8	2067
title	3	9	5.3	179
description	7	35	15.7	532
narrative	14	84	39.9	1356

Table 4: Topic statistics of development and test collections. For development and evaluation, queries were generated by using all the available topic fields.

	# of Relevant Docs			
Data Set (topic #'s)	Min	Max	Avg.	Total
CLIR (54-81)	2	15	7.1	170
CLEF (1-40)	1	42	9.9	338

Table 5: Document retrieval statistics of development and test collections.

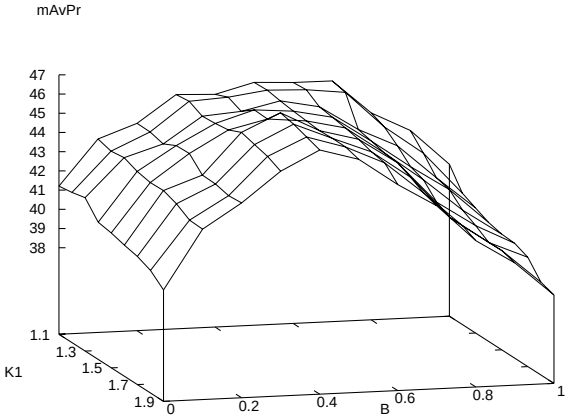

Figure 1: Mean Average Precision versus different settings of Okapi formula's parameters k_1 and b.

documents are taken and the T 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 frequency of word w inside the B top documents.

4. EXPERIMENTS

This section presents work done to develop and test the presented models. Development and testing were done on two different Italian document retrieval tasks. Performance was measured in terms of Average Precision (AvPr) and mean Average Precision (mAvPr). Given the document ranking provided against a given query q, let $r_1 \leq \ldots \leq r_k$ be the ranks of the retrieved relevant documents. The AvPr for q is defined as the average of the precision values achieved at all recall points, i.e.:

$$\text{AvPr} = 100 \times \frac{1}{k} \sum_{i=1}^{k} \frac{i}{r_i} \quad (9)$$

The mAvPr of a set of queries corresponds to the mean of the corresponding query AvPr values.

4.1. Development

For the purpose of parameter tuning, development material made available by CLEF was used. The collection consists of the test set used by the 1999 TREC-8 CLIR track and its relevance assessments. The CLIR collection contains topics and documents in four languages: English, German, French, and Italian. The Italian part consists of texts issued by the Swiss News Agency (*Schweizerische Depeschenagentur*) from 17-11-1989 until 12-31-1990, and 28 topics, four of which have no corresponding Italian relevant documents[1]. More details about the development collection are provided in Tables 3, 4, and 5.

4.2. Okapi Tuning

Tuning of the parameters in formula (2) was carried out on the development data. In Figure 1 a plot of the mAvPr versus different values of the parameters is shown. Finally, the values $k_1 = 1.5$ and $b = 0.4$ were chosen, because they provided consistently good results also with other evaluation measures. The achieved mAvPr is 46.07%.

4.3. Blind Relevance Feedback Tuning

Tuning of BRF parameters B and T was carried out just for the Okapi model. In Figure 2 a plot of the mAvPr versus different values of the parame-

[1]CLIR topics without Italian relevant documents are 60, 63, 76, and 80.

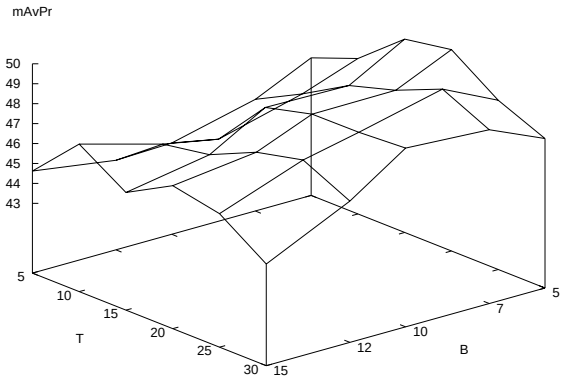

Figure 2: Mean Average Precision versus different settings of blind relevance feedback parameters B and T.

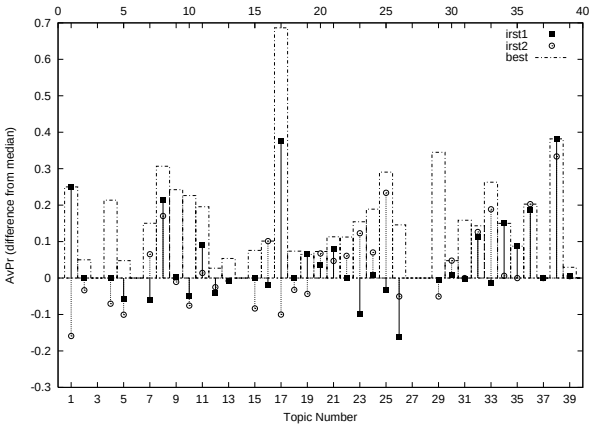

Figure 3: Difference (in mean average precision) from the median for each of the 34 topics in the CLEF 2000 Italian monolingual track. Moreover, the best AvPr reference is plotted for each topic.

ters is shown. Finally, the number of relevant documents $B = 5$ and the number of relevant terms $T = 15$ were chosen, whose combination gives a mAvPr of 49.2%, corresponding to a 6.8% improvement over the first step.

Further work was done to optimize the performance of the first retrieval step. Indeed, performance of the BRF procedure is determined by the precision achieved, by the first retrieval phase, on the very top ranking documents. In particular, an higher resolution for documents and queries was considered by using base forms instead of stems. In Table 6 mAvPr values are shown by considering different combinations of text preprocessing before and after BRF. In particular, we considered using base forms before and after BRF, using word stems before and after BRF, and using base forms before BRF and stems after BRF. The last combination achieved the largest improvement (8.6%) and was adopted for the final system.

		# of relevant terms T					
I	II	5	10	15	20	25	30
st	st	46.4	47.3	49.2	49.6	48.3	48.5
ba	ba	46.2	47.6	47.6	47.6	47.7	47.3
ba	st	46.7	48.7	50.0	48.5	48.6	48.6

Table 6: Mean Average Precision by using base forms (ba) or word stems (st) before (I) and after (II) blind relevance feedback (with B=5).

4.4. Official Evaluation

The two presented models were evaluated on the CLEF 2000 Italian monolingual track. The test collection consists of newspaper articles published by *La Stampa*, during 1994, and 40 topics. As six of the topics do not have corresponding documents in the collection they are not taken into account[2]. More details about the CLEF collection and topics are in Tables 3, 4, and 5.

Official results of the Okapi and statistical models are reported in Figure 3 with the names *irst1* and *irst2*, respectively. Figure 3 shows the difference in AvPr between each run and the median reference provided by the CLEF organization. As a further reference, performance differences between the best result of CLEF and the median are also plotted. The mAvPr of irst1 and irst2 are 49.0% and 47.5%, respectively. Both methods score above the median reference mAvPr, which is 44.5%. The mAvPr of the median reference was computed by taking the average over the median AvPr scores.

5. IMPROVEMENTS

By looking at Figure 3 it emerges that the Okapi and the statistical model have quite different behaviors. This would suggest that if the two methods rank documents independently, some information about the relevant documents could be gained by integrating the scores of both methods.

In order to compare the rankings of two models A and B, the Spearman's rank correlation can be applied. Given a query, let $r(A(d))$ and $r(B(d))$ represent the ranks of document d given by A and B, respectively. Hence, Spearman's rank correlation (Mood et al., 1974) is defined as:

$$S = 1 - \frac{6 \sum_{d} [r(A(d)) - r(B(d))]^2}{N(N^2 - 1)} \qquad (10)$$

[2]CLEF topics without Italian relevant documents are 3, 6, 14, 27, 28, and 40.

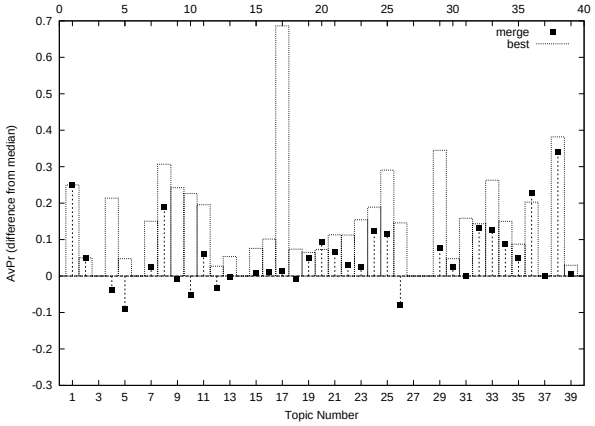

Figure 4: Difference (in mean average precision) from the median of the combined model and the best reference of CLEF 2000.

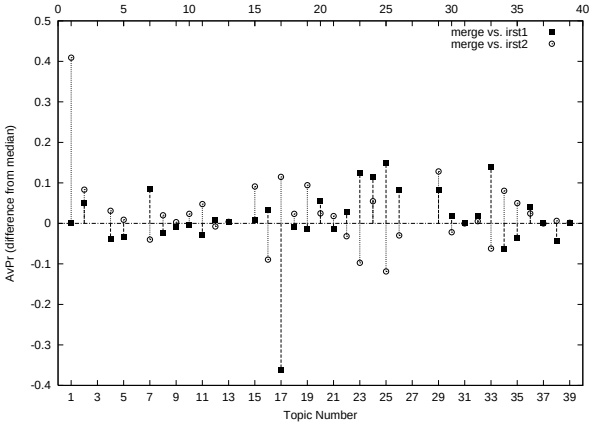

Figure 5: Difference (in mean average precision) of the combined model from each single model.

Retrieval Model	Official Run	mAvPr
Okapi	irst1	49.0
Statistical model	irst2	47.5
Combined model	-	50.0

Table 7: Performance of retrieval models on the CLEF 2000 Italian monolingual track.

Under the hypothesis of independence between A and B, S has mean 0 and variance $1/(N-1)$. On the contrary, in case of perfect correlation the S statistics has value 1.

By taking the average of S over all the queries [3], a rank correlation of 0.4 resulted between the irst1 and irst2 runs.

This results confirms some degree of independence between the two information retrieval models. Hence, a combination of the two models was implemented by just taking the sum of scores. Actually, in order to adjust scale differences, scores of each model were normalized in the range $[0, 1]$ before summation. By using the official relevance assessments of CLEF, a mAvPr of 50.0% was achieved by the combined model.

In Figure 4 and Figure 5 detailed results of the combined model (*merge*) are provided for each query, respectively, against the CLEF references and the irst1 and irst2 runs. It results that the combined model performs better than the median reference on 24 topics of 34, while irst1 and irst2 improved the median AvPr 16 e 17 times, respectively. Finally, the combined model improves the best reference on two topics (20 and 36).

[3] As an approximation, rankings were computed for the union of the 100 top documents retrieved by each model.

6. CONCLUSION

This paper presents preliminary research results by ITC-irst in the field of text retrieval. Nevertheless, participation to the CLEF evaluation has been considered important in order to gain experience and feedback about our progress. Future work will be done to improve the statistical retrieval model, develop a statistical blind relevance feedback method, and extend the text retrieval system to other languages, i.e. English and German.

7. References

Federico, Marcello, 2000. A system for the retrieval of italian broadcast news. *Speech Communication*, 33(1-2).

Federico, Marcello and Renato De Mori, 1998. Language modelling. In Renato De Mori (ed.), *Spoken Dialogues with Computers*, chapter 7. London, UK: Academy Press.

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.

Merialdo, Bernard, 1994. Tagging English text with a probabilistic model. *Computational Linguistics*, 20(2):155–172.

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

Mood, Alexander M., Franklin A. Graybill, and Duane C. Boes, 1974. *Introduction to the Theory of Statistics*. Singapore: McGraw-Hill.

Ng, Kenney, 1999. A maximum likelihood ratio information retrieval model. In *Proceedings of the 8th Text REtrieval Conference*. Gaithersburg, MD.

Text	POS	Base form	Stem	R
IL	RS	IL	IL	0
PRIMO	AS	PRIMO	PRIM	1
MINISTRO	SS	MINISTRO	MINISTR	1
LITUANO	AS	LITUANO	LITUAN	1
,	XPW	,	,	0
SIGNORA	SS	SIGNORA	SIGNOR	1
KAZIMIERA	SPN	KAZIMIERA	KAZIMIER	1
PRUNSKIENE	SPN	PRUNSKIENE	PRUNSKIEN	1
,	XPW	,	,	0
HA	#VI#	AVERE	AVERE	0
ANCORA	B	ANCORA	ANCORA	0
UNA	RS	UNA	UNA	0
VOLTA	SS	VOLTA	VOLT	1
SOLLECITATO	VSP	SOLLECITARE	SOLLECIT	1
OGGI	B	OGGI	OGGI	0
UN	RS	UN	UN	0
RAPIDO	#SS#	RAPIDO	RAPID	1
AVVIO	SS	AVVIO	AVVIO	1
DEI	EP	DEI	DEI	0
NEGOZIATI	SP	NEGOZIATO	NEG	1
CON	E	CON	CON	0
L’	RS	L’	L’	0
URSS	YA	URSS	URSS	1
,	XPW	,	,	0
RITENENDO	VG	RITENERE	RITEN	0
FAVOREVOLE	AS	FAVOREVOLE	FAVOR	1
L’	RS	L’	L’	0
ATTUALE	AS	ATTUALE	ATTUAL	1
SITUAZIONE	SS	SITUAZIONE	SIT	1
NEI	EP	NEI	NEI	0
RAPPORTI	SP	RAPPORTO	RAPPORT	1
FRA	E	FRA	FRA	0
MOSCA	SPN	MOSCA	MOSC	1
E	C	E	E	0
VILNIUS	SPN	VILNIUS	VILNIUS	1

Table 8: Example of text preprocessing. The flag in the last column indicates if the term survives or not after the stop-terms removal. The two POSs marked with # are wrong, nevertheless they permit to generate correct base forms and stems.

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

Sparck Jones, Karen and Peter Willett (eds.), 1997. *Readings in Information Retrieval*. San Francisco, CA: Morgan Kaufmann.

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.