Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Processing and Management 42 (2006) 1203–1217
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Relevance feedback and cross-language information retrieval

Viviane Moreira Orengo *, Christian Huyck

School of Computing Science, Middlesex University, The Burroughs, London NW4 4BT, UK

Received 24 August 2005; received in revised form 22 December 2005; accepted 22 December 2005
Available online 24 February 2006

Abstract

This paper presents a study of relevance feedback in a cross-language information retrieval environment. We have performed an experiment in which Portuguese speakers are asked to judge the relevance of English documents; documents hand-translated to Portuguese and documents automatically translated to Portuguese. The goals of the experiment were to answer two questions (i) how well can native Portuguese searchers recognise relevant documents written in English, compared to documents that are hand translated and automatically translated to Portuguese; and (ii) what is the impact of misjudged documents on the performance improvement that can be achieved by relevance feedback. Surprisingly, the results show that machine translation is as effective as hand translation in aiding users to assess relevance in the experiment. In addition, the impact of misjudged documents on the performance of RF is overall just moderate, and varies greatly for different query topics.

Keywords: Cross-language information retrieval; Relevance feedback

1. Introduction

The increasing availability of information in different languages and the growing number of people speaking different mother tongues who want to find information have been motivating research on cross-language information retrieval (CLIR). CLIR is responsible for receiving search requests expressed in one language and retrieving documents written in another language. The scope of CLIR typically involves mapping a concept from one language into another.

Some traditional information retrieval (IR) techniques, such as relevance feedback (RF) acquire a new dimension in this cross-linguistic environment: the user's ability to recognise relevant documents written in a foreign language or translated, by some means, into his language. The RF operation is an automatic process for the modification of search requests based on relevance assessments provided by the user population for previously retrieved documents (Salton, 1971). The idea behind it is that users are unlikely to produce perfect

* Corresponding author. Present address: Instituto de Informática—UFRGS, Av. Bento Gonçalves, 9500-Bloco IV, CEP 91501-970 Porto Alegre, RS, Brazil. Tel.: +55 51 3026 8263; fax: +55 51 3316 7308.
E-mail address: vmorengo@inf.ufrgs.br (V.M. Orengo).

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2005.12.003

queries, especially if given just one attempt. The typical process improves the query specification by choosing important terms attached to previously retrieved documents that have been identified as relevant by the user. Thus, RF usually involves query expansion and term reweighting. The method consists of asking the user to analyse an initial sample of documents retrieved in response to a query and judge them for relevance. The original query is then modified (by the IR system) and re-submitted. A new list of retrieved documents is generated. The assumption is that this new list will be better than the previous. RF is an interactive method that can be repeated several times, until the user is satisfied with the results.

RF can also be performed without user interference, in a technique known as pseudo-relevance feedback (PRF). In PRF, *n* top ranked documents are assumed relevant and used for the feedback runs. This technique typically achieves less improvement than original RF, however it has the advantage of being done automatically without any burden to the user. Xu and Croft (1996) proposed a related strategy called local context analysis (LCA) that combines PRF and global analysis. Global analysis extract concepts from text, based on word co-occurrences using statistical techniques.

There is yet another type of feedback, discussed by Efthimiadis and Robertson (1989), in which query reformulation is not trusted entirely to IR system, allowing the process to be controlled by the user. The idea is to offer related search terms for the user to choose from and expand the query. These related terms could be obtained from a thesaurus or from previous search results.

This paper focuses on the RF process, and the scenario considered is that of a Portuguese–English cross-language information retrieval (CLIR) system. The aims of this paper are twofold. The first aim is to find out how well native Portuguese searchers can recognise relevant documents written in English compared to documents that are hand translated and automatically translated to Portuguese. The second aim is to analyse the effect that misjudged documents have on the change in performance achieved through the RF process.

The remainder of this paper is organised as follows: Section 2 presents related work; Section 3 describes the CLIR system used; Section 4 details the experimental design; Section 5 evaluates the users ability in making relevance judgements; Section 6 assesses the impact that errors in judgement have on the performance of RF; Section 7 finalises the paper presenting the summary and conclusions.

2. Related work

The concept of RF was introduced in the mid-1960s. The first RF methods were designed to be used with vector queries. Some early experiments were performed by Rocchio (1971) on the SMART Retrieval System. Since then, the method has been applied to other IR approaches. In 1976, Robertson and Sparck Jones (1976) experimented with RF on a probabilistic model; and Dillon and Desper (1980) proposed an RF method for a Boolean Retrieval System. In addition, several different feedback procedures have been proposed. Some of the best known were developed by Ide (1971). More recently, new approaches include Neural Networks (Crestani, 2000); Genetic Algorithms (Biron & Kraft, 1995), and Machine Learning (Drucker, Shahary, & Gibbon, 2001).

Several experiments report on the performance improvement achieved by relevance feedback; Salton and Buckley (1997) experimented with several RF methods. The improvement achieved with the technique (measured using the residual collection approach) ranged between 47% (for the CISI collection) and 160% (for the Cranfield collection). Harman (1992b) reports improvements of 112% for the Cranfield collection when expanding the query with 20 terms chosen from the relevant documents.

Salton and Buckley (1997) also concluded that some types of collections may benefit more from the RF process. These are collections with short queries; collections with queries that perform relatively poorly in an initial search; and technical collections. Collections with short queries can benefit from the RF process as it will add context, making the query more complete. Collections with queries that have a weak performance on the initial run have more potential for improvement. In technical collections, it is possible that the set of relevant documents for a given query is concentrated in a small area of the document space.

McNamee and Mayfield (2003) report that RF does not always improve performance, and in some cases can even decrease it. They observed this problem especially when the initial queries are longer. However, those findings were based on PRF which normally performs worse than user RF.

Despite the great research interest on RF, very few experiments have been carried out using human searchers. A user-centered investigation has been made by Efthimiadis (2000); he observed 25 searchers querying the

INSPEC database. The main findings confirm the effectiveness of RF. The initial search produced on average three highly relevant documents, and the feedback run produced on average nine further highly relevant documents. Another study involving users was done by Spink (1994); she performed an experiment with forty users to assess how humans perform query expansion in an interactive environment. She concluded that users are able to select effective terms for query expansion. However, because the study used real user queries, it is not possible to calculate evaluation measures and thus quantify the gain obtained from reformulation.

RF has been widely applied to CLIR with good results. However, the vast majority of the experiments have used PRF (McNamee & Mayfield, 2002; Qu, Eilerman, Jin, & Evans, 2000; Yang, Carbonell, Brown, & Frederking, 1997) or LCA (Ballesteros & Croft, 1997). Neither of these methods employ users to assess relevance. Users' assessments of relevance are especially important for CLIR since the feedback process involves the subjects' ability to assess the relevance of documents written in foreign languages or automatically translated into the user's language.

This lack of user experiments in the CLIR environment has been addressed in part by the Interactive Track at the cross-language evaluation forum (iCLEF) (Oard & Gonzalo, 2001, 2003a, 2003b), which provides a common framework for participant groups to evaluate several aspects related to the formulation of queries, translation of queries, and assessment of relevance.

A related study developed by Karlgren and Hansen (2003) for iCLEF compared the performance of users assessing documents in their native language (Swedish) with their performance in assessing documents in a language they know well (English). As expected, they found that users take longer and make more mistakes when judging documents in a foreign language even assuming a good knowledge of that language.

The use of Machine Translation (MT) in aiding relevance assessments was analysed by iCLEF in three studies:

- Wang and Oard (2001) compared the performance of full MT and term-for-term gloss translations obtained from bilingual term lists found on the web. Subjects had little or no knowledge of the language of the documents. The results show that searchers were able to make relevance judgements with either approach. However, MT achieved slightly better results.
- Bathie and Sanderson (2002) compared users' ability in judging native language documents and documents originally written in a foreign language and automatically translated into the user's language. The documents were articles from the LA Times in their original language and from Le Monde automatically translated into English. The study concluded that users were able to make judgements with the same accuracy for both types of documents.
- López-Ostenero, Gonzalo, Peñas, and Verdejo (2001) compared the performance of MT and a phrase translation based algorithm developed with the use of comparable corpora. Searchers had low or no proficiency on the language of the documents. The results show that precision was similar for both systems, but recall was better when using phrasal translations.

Though these experiments demonstrate that MT can facilitate relevance assessment in a CLIR environment, no study has yet examined the extent to which judging relevance may be better when using human translations of foreign language documents rather than MT. Further, the reviewed literature does not present any research on how the errors in judgement affect the change in performance achieved by the RF process. Those aspects are addressed by the experiment described in this paper.

3. The CLIR system

The CLIR system used in this paper[1] was implemented using Latent Semantic Indexing (LSI), a method proposed by Deerwester, Dumais, Furnas, Landauer, and Harshman (1990), and extensively tested by Dumais (1991, 1995) and Dumais and Nielsen, 1992. The main goal of using LSI for CLIR is to provide means for

[1] A more detailed description of the CLIR system can be found in Orengo and Huyck (2003).

matching text segments in one language with text segments of similar meaning in another language without needing to translate either, by creating a language-independent representation of the words.

LSI was first applied to CLIR by Landauer and Littman (1990). The method used here is essentially the same as theirs. However, since there is no parallel corpus containing Portuguese and English, SYSTRAN 3.0 Professional was used to translate a sample of documents (approximately 20%) from the collection described in Section 4.2 to simulate a parallel corpus. The reason for choosing SYSTRAN is that it is a widely used translator in CLIR literature, especially for CLEF experiments (Braschler, 2003).

LSI is applied to a matrix of terms by documents. Therefore, the first step is to build such a matrix based on a set of dual-language documents.[2] The matrix contains the number of occurrences (or weights) of each term in each document. In an ideal situation the pattern of occurrence of a term in language A should be identical to the pattern of occurrence of its match in language B. The resulting matrix tends to be very sparse, since most terms do not occur in every document.

This matrix is then factorised by singular value decomposition[3] (SVD). SVD reduces the number of dimensions, throwing away the small sources of variability in term usage. The k most important dimensions are kept. Roughly speaking, these dimensions (or factors) may be thought of as artificial concepts; they represent extracted common meaning components of many different words and documents. Each term or document is then characterised by a vector of weights indicating its strength of association with each of these underlying concepts. Since the number of factors or dimensions is much smaller than the number of unique terms, words will not be independent. For example, if two terms are used in similar documents, they will have similar vectors in the reduced-dimension representation.

LSI implements the vector-space model, in which terms, documents and queries are represented as vectors in a k-dimensional semantic space. After deriving the semantic space with an initial sample of dual-language documents, new documents can be added. Those new documents will be placed at a location calculated by averaging the vectors of the terms that it contains. This process is known as ''folding in''. Queries are treated as pseudo-documents and placed at the weighted sum of its component term vectors. The similarity between query and documents is measured using the cosine between their vectors.

SVD causes synonyms to be represented by similar vectors (since they would have many co-occurrences), which allows relevant documents to be retrieved even if they do not share any terms with the query. This is what makes LSI suitable for CLIR, given that a term in one language will be treated as a synonym to its match in the other language. The main advantages of using LSI for CLIR are:

- There is no traditional-style translation. All terms and documents are transformed to a language-independent representation.
- New languages can be added easily, provided you have training data.
- There is no need for expensive resources such as dictionaries, thesauri or machine translation systems.

Furthermore, Yang et al. (1997) performed tests with several CLIR approaches, including query translation and statistical methods, in all tests, LSI's performance was among the best. In addition, the loss in performance between monolingual and bilingual executions was small, about 15%. The bilingual version of our system achieved 81% of the monolingual performance (with the LA Times test collection, described in Section 4.2). This also shows MT is a feasible alternative for simulating a parallel corpus.

An important aspect of LSI which made us choose this method for the experiments described in the next section is that the same RF strategy used in a monolingual LSI system can be directly applied to a CLIR system. That represents an important advantage, since most RF methods cannot be directly applied to CLIR as the words from the documents will not match the words from the queries. The literature contains some cases in which RF methods have been adapted for CLIR. Yang et al. (1997) proposed a method for PRF on a bilingual

[2] Dual-language documents are composed by the document in the original language together with its translation in another language.
[3] Mathematics are presented in detail in Deerwester et al. (1990).

collection which consists in doing an initial retrieval on the collection that has the same language of the query; finding the translation mates for the top ranked documents; and then using those documents to create a query in the target language. Ballesteros and Croft (1998) proposed a method for using PRF with dictionary methods. Three alternatives are tested: pre-translation query modification, post-translation query modification and a combination of both. Qu et al. (2000), suggested a similar method for PRF in MT-based systems. All methods reported above mention performance improvements.

4. Experimental design

The design of the experiment aims at answering two main questions:

(i) How well can native Portuguese searchers recognise relevant documents written in English, compared to documents that are hand translated and automatically translated to Portuguese?
(ii) What is the impact of misjudged documents on the performance improvement that can be achieved by RF?

It is worth pointing out that in order to answer the second question, we could have employed simulated user judgements. However, real users provide a better insight on the type and frequency of judgement errors that are made in an operational setting, since their choices are not random. Considering that user judgements were vital for answering the first question, we opted for using them for the second aspect as well.

The next subsections describe the design of the experiment. Characteristics of the searcher, document collection, query topics and procedure are detailed.

4.1. Searcher

The aim was to obtain subjects that would be likely users of a CLIR system. In this case: Portuguese speakers who have basic or no knowledge of English, that are not able to express their queries in English and that are familiar with computer searching. The searchers were recruited among students and lecturers from UCPel (Universidade Católica de Pelotas—http://www.ucpel.tche.br), in the south of Brazil. A total of 27 participants were obtained. The average age was 29.

Language skills are hard to measure accurately; what may be considered ''intermediate'' to one person, might be considered ''advanced'' by another. Ideally, the searchers would have taken a standard English language test such as TOEFL, enabling a more exact categorisation of their knowledge. However, that was not possible. The approach taken was to ask the searchers to rate their ability in writing and reading in English (in two separate questions). There were 5 levels of ability ranging from ''none'' to ''proficient''. Fig. 1 presents a sample question. Most of the answers (13) fell into category 4. The remainder fell into categories 3 (8) and 5 (6).

4.2. Document collection

The collection used in the experiments consists of over 113 000 news articles from the Los Angeles Times amounting to 450 Mb. The documents, which were published in 1994, deal with a broad variety of subjects

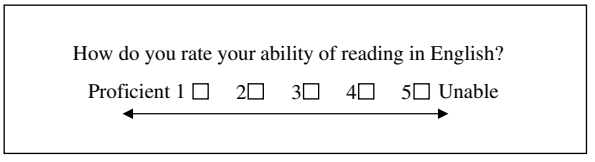

Fig. 1. English knowledge question.

such as politics, business, sports, culture and entertainment. This collection has been provided by the cross-language evaluation forum (CLEF).[4]

4.3. Query topics

Six query topics were extracted from CLEF 2002, which had a total of 50 queries. The Portuguese version of the topics was used. Below we present the criteria for topic selection, which were defined based on initial search results:

- Select topics that had more than 10 relevant documents. This criteria prevents the situation in which all relevant documents are presented to the user for feedback.
- Select topics that have relevant documents among top ten retrieved. Since the RF method used only positive feedback, this criteria was used to prevent the situation in which the user does not judge any document as being relevant.

Seventeen of the fifty topics satisfied the above conditions. Six of them were then randomly selected. The English version of the selected topics is presented below:

Topic 1
⟨num⟩ C092 ⟨/num⟩
⟨EN-title⟩ UN sanctions against Iraq ⟨/EN-title⟩
⟨EN-desc⟩ What measures has Iraq taken to effect the lifting of the UN economic embargo and political sanctions imposed after its invasion of Kuwait in 1990? ⟨/EN-desc⟩
⟨EN-narr⟩ Documents must include ways in which Iraq has attempted to get the sanctions lifted. Mere descriptions of the sanctions or rhetoric against the sanctions are not relevant. Expressions of regret for invading Kuwait by Iraqi officials are relevant. ⟨/EN-narr⟩

Topic 2
⟨num⟩ C094 ⟨/num⟩
⟨EN-title⟩ Return of Solzhenitsyn ⟨/EN-title⟩
⟨EN-desc⟩ Find documents which report about the return of the Nobel prize winner for literature Solzhenitsyn to Russia. ⟨/EN-desc⟩
⟨EN-narr⟩ Relevant documents report the reasons and the time of the return of Solzhenitsyn to Russia. They may also talk about the reasons for his emigration to the US. ⟨/EN-narr⟩

Topic 3
⟨num⟩ C107 ⟨/num⟩
⟨EN-title⟩ Genetic Engineering ⟨/EN-title⟩
⟨EN-desc⟩ How does genetic engineering affect the human food chain? ⟨/EN-desc⟩
⟨EN-narr⟩ Articles must directly address the introduction of genetic engineering, and its effects on the human food chain. They will discuss both pros and cons. Reports on tobacco bioengineering and human gene engineering are not relevant. ⟨/EN-narr⟩

Topic 4
⟨num⟩ C123 ⟨/num⟩
⟨EN-title⟩ Marriage Jackson–Presley ⟨/EN-title⟩
⟨EN-desc⟩ Find documents that report on the presumed marriage of Michael Jackson with Lisa Marie Presley or on their separation. ⟨/EN-desc⟩

[4] http://www.clef-campaign.org/.

⟨EN-narr⟩ In May 1994, the famous pop star, Michael Jackson, was reported to have married Lisa Marie Presley, the daughter of the king of rock and roll. Relevant documents must either contain some details regarding the wedding, such as where or when it was held, or must discuss the later separation of the couple. ⟨/EN-narr⟩

Topic 5
⟨num⟩ C130 ⟨/num⟩
⟨EN-title⟩ Death of Nirvana leader ⟨/EN-title⟩
⟨EN-desc⟩ How did the lead singer of the American rock and grunge group, Nirvana, die? ⟨/EN-desc⟩
⟨EN-narr⟩ Kurt Cobain, lead singer of Nirvana, the famous popular music group, died in April 1994. Documents that report the death of Cobain without mentioning the cause are not relevant. ⟨/EN-narr⟩

Topic 6
⟨num⟩ C140 ⟨/num⟩
⟨EN-title⟩ Mobile phones ⟨/EN-title⟩
⟨EN-desc⟩ Prospects for the use of cellular phones. ⟨/EN-desc⟩
⟨EN-narr⟩ Relevant documents report on the prospects for the use of cellular phones and the development of the mobile phone industry. ⟨/EN-narr⟩

4.4. Procedure

Fig. 2 shows how the experiment was performed. Searchers, represented by the man in the left of the figure, were presented with query topics (written in Portuguese) and a ranked list of 10 documents returned in response to an initial query. This ranked list was produced by presenting all terms from the "title" and "description"[5] fields to the CLIR-LSI system, described in Section 3, as queries. The documents whose vectors had the highest cosine with the query vector were ranked as best matches. The participants were asked to classify each document in relation to the topic in one of three categories: "relevant", "not relevant" or "not sure". Similar to what is done by iCLEF (Oard & Gonzalo, 2003b), the participants were given a definition of relevance. They were told to picture the situation in which they had to write a report on the query topic. They should consider relevant any document that contains information on the topic. Documents with only a part (or portion) related to the topic should also be considered relevant. Additionally, each document should be judged independently of other documents, even if they contain the same information.

Each participant read 6 queries and 10 documents for each query, amounting to 60 relevance judgements per participant and 1620 in total. The users saw the full text of the documents, which was presented in one of the three formats presented below:

- the original English text, as returned from the CLIR-LSI system (System 1),
- a machine translation produced using SYSTRAN 3.0 Professional (System 2),
- a human translation, produced by the first author (System 3).

The number of relevant documents per query varies. Similarly, the number of relevant documents ranked in the top ten and presented to the user varied (see Table 1).The order of the queries was varied systematically in a Latin square design, which controlled for learning effect and tiredness of the searchers. The order in which the different systems were presented has also been varied. Table 2 shows a 9-subject matrix. As there were 27 participants, the same matrix was used three times. Participant 1 saw the documents for topics 1 and 2 in the original language (English), then the documents for topics 3 and 4 automatically translated into Portuguese and finally, documents for topics 5 and 6 manually translated to Portuguese. Participants 1, 2 and 3 had

[5] Braschler (2003) reports that the vast majority (80%) of the runs submitted for CLEF used the fields 'title' and 'description', discarding the narrative. For our experiments, this combination produces the best results.

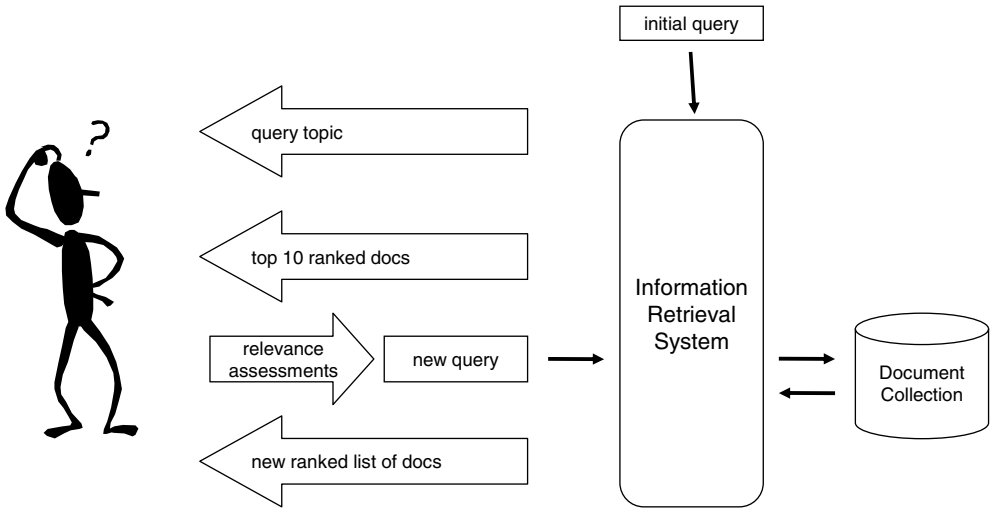

Fig. 2. Experiment procedure.

Table 1
Number of relevant documents per topic

Topic	Number of relevant documents ranked in top 10	Number of relevant documents in the collection
1	6	27
2	8	12
3	5	32
4	2	19
5	7	63
6	5	70

Table 2
Subject matrix showing topic-system combination and the presentation order

Participant	First batch			Second batch			Third batch		
1	S1	1	2	S2	3	4	S3	5	6
2	S2	3	4	S3	5	6	S1	1	2
3	S3	5	6	S1	1	2	S2	3	4
4	S1	3	4	S2	5	6	S3	1	2
5	S2	5	6	S3	1	2	S1	3	4
6	S3	1	2	S1	3	4	S2	5	6
7	S1	5	6	S2	1	2	S3	3	4
8	S2	1	2	S3	3	4	S1	5	6
9	S3	3	4	S1	5	6	S2	1	2

the same topic-system combination, however the order in which the query topics were presented was different for each subject. The average time taken to judge all sixty documents was one hour.

Besides providing relevance judgements, the users were asked some questions related to their language skills, experience in computer searching, confidence in the judgements made, prior knowledge of the query topics, difficulty of the judgements, and if they preferred to view the documents in their original language or translated into Portuguese.

After gathering the judgements for all 27 searchers, the queries were re-formulated, resubmitted and re-evaluated for recall and precision. RF was performed by replacing the original query with the vector average of the documents the user selected as relevant, as described in Dumais (1991).

5. Users ability in making relevance judgements

As reported in the previous section, a three-point relevance scale was used. However, to be compatible with CLEF assessments and the evaluation software, all "not sure" were forced to "irrelevant". Analysis of the data concentrated mainly on the following aspects:

1. The number of mistakes made by the searcher.
2. The level of agreement between the CLEF judgements and the judgements of each user.
3. Confidence of the judgements, the difficulty of the task, and prior knowledge of the topics.

The data collected for most variables is not perfectly normally distributed. The statistical test chosen to compare the results for different groups was an ANOVA as it is robust in dealing with data that depart from the normality assumption. For all tests reported, α was set to 0.05.

5.1. Number of mistakes

The relevance judgements provided by CLEF were considered as "correct answers". Each judgement collected from the participants was compared against them. Two types of mistakes were analysed: (i) false alarm, if the searcher judged an irrelevant document as relevant; and (ii) relevant missed, if the user judged a relevant document as irrelevant.

A total of 1620 judgements were made, 540 for each system (see Section 4.4). Each query topic had a variable number of relevant documents (see Table 1), so the figures for relevant missed and false alarm had to be properly weighted to allow for fair comparisons between topics.

Table 3 shows the numbers for missed relevant, false alarms, and correct judgements. It also displays how the judgements spread across the 3 possible categories: relevant, not relevant and not sure.

The number of relevant missed was virtually the same for the machine translated texts and the hand translated texts. The number of relevant missed for the original texts was much higher (43%). That happened because most judgements (63%) for this system fell into the "unsure category". The number of false alarm was very small in the original texts for the same reason, and was the largest for the hand translated texts because people made more positive judgements in that system. An ANOVA test on missed relevant and false alarm has shown no significant difference between judgements made using hand and machine translated texts (p-values 0.95 and 0.12, respectively).

5.2. Overlap

Overlap has been defined by Lesk and Salton (1968) as the intersection of the relevant documents divided by the union of the relevant document sets. In the context of this experiment, overlap measures how accurate the participant's judgements were, as it tells how similar each participant's judgements were compared to the relevance judgements provided by CLEF. Only the documents presented to the users were considered when calculating the overlap.

Table 3
Summary of judgements

	Hand translated	Machine translated	Original
Missed relevant	142 (26%)	141 (26%)	234 (43%)
False alarm	94 (17%)	80 (15%)	43 (8%)
Correct	304 (56%)	319 (59%)	263 (48%)
Relevant	249 (46%)	236 (43%)	106 (19%)
Not relevant	243 (45%)	241 (45%)	96 (18%)
Not sure	48 (9%)	63 (12%)	338 (63%)
Overlap	0.40	0.41	0.16

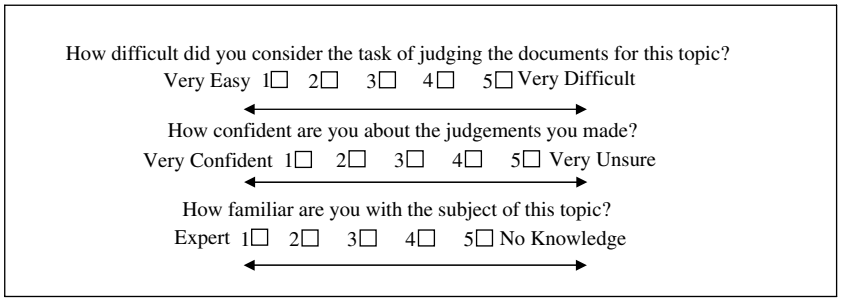

Fig. 3. Questions on difficulty, confidence and familiarity with the topic's subject.

The average overlap between the CLEF assessors and the participants of the experiment is shown at the bottom of Table 3. These results confirm the ones from the previous section, as the figures for the hand translated and machine translated texts are very similar, and much better than the scores achieved when assessing the original texts. An ANOVA test has shown no significant difference (p-value = 0.58) between the overlap of judgements made on hand and machine translation texts.

5.3. Difficulty, confidence and knowledge

The participants were asked to rate each topic in terms of difficulty of the task and the confidence they had in the judgements made. They were also asked how familiar they were with the subject of the query topics. For all three questions there were five levels. An English version of the questions asked is shown in Fig. 3.

Participants found judging English documents considerably more difficult than judging MT or hand translated documents. They also had less confidence in the judgements made. Confidence and difficulty were similar for MT and hand translated documents.

Those three measures have also been evaluated separately for each topic. Difficulty and confidence had very similar distributions along the five categories. No significant difference was found among the level of difficulty of the topics. Likewise, no significant difference was found among the degree of confidence in the judgements made for each topic. An ANOVA test confirms that (p-values are 0.65 for difficulty of topics, and 0.84 for the level of confidence in the judgements).

5.4. Testing other user groups

The experiment was repeated with a smaller sample of 6 bilingual participants, all native Portuguese speakers with very good English skills. Four participants rated their ability in reading and writing in English on category 1 (proficient) and the remaining two rated their ability on category number two.

The results are summarised in Table 4. Bilingual participants were able to make judgements with the same accuracy using hand translated, machine translated and original texts. An ANOVA test using the data from false alarm, missed relevant and the overlap showed no significant difference among the three means.

Table 4
Results for bilingual users

	Hand translated	Machine translated	Original
Missed relevant	21 (18%)	25 (20%)	30 (25%)
False alarm	11 (9%)	19 (16%)	21 (18%)
Correct	88 (73%)	76 (64%)	69 (57%)
Relevant	66 (55%)	67 (56%)	67 (56%)
Not relevant	51 (43%)	34 (28%)	45 (38%)
Not sure	3 (2%)	19 (16%)	8 (6%)
Overlap	0.55	0.49	0.43

Table 5
Results native speakers

Missed relevant	36 (30%)
False alarm	3 (3%)
Correct	81 (67%)
Relevant	33 (28%)
Not relevant	78 (65%)
Not sure	9 (7%)
Overlap	0.46

The experiment has also been repeated with native English speakers. The aim was to establish the expected degree of agreement between a participant and CLEF judges, when the participant fully understood the language of the documents. The 6 participants recruited saw only the English documents, and each judged 2 query topics (20 documents). In total, data for 12 queries were analysed. Table 5 shows the results for this group.

The results shown here are comparable to the ones obtained by Voorhees (1998) in an experiment using three groups of relevance assessors, all native English speakers from a similar background judging English documents. She found that the overlap between pairs ranged from 0.42 to 0.49.

The performance of the Portuguese participants from both groups (poor English skills and bilingual) judging hand and automatically translated documents is equivalent to the performance of native English speakers judging English documents. That confirms the conclusions of the main experiment and indicates that MT is as effective as hand translation in aiding users to assess relevance.

5.5. Discussion

The superior performance of machine translated texts in comparison to the judgements made on the original texts happened despite the many translation errors and awkward grammar, which led to several complaints from the participants. Those results imply that there is no advantage in judging relevance on hand translated documents despite the extra time and cost incurred in generating them.

It seems logical that the presence of proper names from the queries such as Nirvana or Solzhenitsyn in the documents would provide the user with great help for the judgements. This was expected to aid the assessment of the English documents in particular. However, the results do not confirm that assumption, as even with the presence of such keywords, most users were not able to accurately judge documents. A possible reason is that the participants did not rely on the presence of such clues alone, when they could not understand the context in which such proper names were used.

6. Errors in judgement and change in performance

The precision of the initial query (baseline) was compared with the precision attained after the RF process for each user. Recall that the RF process involved replacing the initial query with the vector average of the documents the user judged relevant. The performance was evaluated using the "residual collection" method, whereby the documents that have been judged by the searcher are excluded from the collection and from the relevance assessments. The second ranked list will only contain documents that have not been judged. This method provides an unbiased evaluation of RF and is a de facto standard used by most RF research (Harman, 1992a).

Table 6 shows average precision for the baseline run and different feedback runs using different sets of relevance judgements. The feedback runs compared are:

- CLEF—using the official judgements provided by CLEF.
- Best—using the judgements provided by the participant who achieved the biggest overall improvement.
- Worst—using the judgements provided by the participant who got the largest overall decline.

Table 6
Average precision figures for initial and feedback runs

Topic	Baseline	CLEF	Best	Worst	Optimal	Average
1	0.4633	0.7176	0.7287	0.4633	0.7393	0.6517
2	0.2665	0.0704	0.0722	0.0273	0.2665	0.1287
3	0.1064	0.0557	0.4773	0.0039	0.6012	0.1256
4	0.0599	0.1813	0.1813	0.0570	0.2617	0.1366
5	0.1661	0.2171	0.2631	0.1748	0.2779	0.2021
6	0.4694	0.4831	0.6457	0.4652	0.6928	0.4792
Average	0.2553	0.2875	0.3947	0.1986	0.4732	0.2873
Change	–	+12.64%	+54.62%	−22.20%	+85.38%	+12.53%

Table 7
Correlation between change in precision and accuracy in judgement

	Change in precision
Missed	−0.24
False alarm	−0.27
M + F	−0.51
Overlap	0.33

- Optimal—selecting the set of judgements that yielded the best result for each topic.
- Average—combining the results from all participants for each topic.

The change in performance varied greatly from one user to another, ranging from a deterioration of 22% to an improvement of 54%. Averaging the results for all users resulted in an improvement of 12.53% on average precision (change from 0.2553 to 0.2873). For the user, this means that for each query there were on average two new relevant documents identified amongst the top ten. The relevance assessments provided by 6 participants yielded a decrease in performance. An important observation is that the official CLEF judgements, which are the gold standard, did not produce the biggest overall improvement. In fact, the official run was outperformed by the runs of 10 participants of the main experiment, which implies that judgement errors may sometimes help the RF process. This fact has also been observed by Shen, Tan, and Zhai (2005) in an implicit feedback experiment. Their system achieved performance improvement even when users clicked on non-relevant documents.

In order to establish what affects the performance of RF, it is necessary to analyse the effect that each variable had in the change in performance achieved. This can be done by calculating correlation coefficients between the change in performance and the other variables analysed. The correlation coefficients obtained when analysing the results for all 162 queries are shown in Table 7.

The overall results show a moderate negative correlation between the change in performance and the missed relevant. A similar correlation is found between the change in performance and false alarm. That is somewhat surprising as it seems logical that selecting an irrelevant document as relevant is a more serious mistake than failing to recognise a relevant document, however this trend seems to be very small. Similarly, there is a moderate positive correlation between the change in performance and the overlap. In summary, it can be said that 7% of the change in performance can be explained by the false alarms, 6% by the missed relevant and 11% by the overlap or accuracy in judgement.

6.1. Evaluation by system

With the purpose of understanding if the different systems had an effect on the performance change, the next step is to compare the results for different systems. Table 8 shows correlation coefficients between the improvement in precision with the misjudged documents (relevant missed and false alarm), and with the overlap for each system.

Table 8
Correlation between change in precision and accuracy in judgement (by system)

	Change in precision hand translation	Change in precision machine translation	Change in precision original
Missed	−0.12	−0.25	−0.38
False alarm	−0.37	−0.27	−0.22
M + F	−0.53	−0.47	−0.54
Overlap	0.25	0.31	0.49

Table 9
Correlation between change in precision and accuracy in judgement (by topic)

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
Missed	−0.83	0.52	−0.05	−0.59	−0.56	−0.21
False alarm	0.67	−0.61	−0.48	−0.26	0.02	0.10
M + F	−0.59	0.37	−0.47	−0.51	−0.61	−0.16
Overlap	0.86	−0.52	0.24	0.70	0.61	0.25

The sum of mistakes has very similar coefficients for all three systems. Differences are found for false alarm, as it is higher for hand translation. That can be attributed to this type of mistake being more frequent for that system than for the other two. Likewise, missed relevant had a higher coefficient in the original texts because that system suffered more from that type of mistake. However, the differences among systems are not large enough to justify assigning the effect of the changes in performance to system variations.

6.2. Evaluation by topic

The next step then is to evaluate the correlation coefficients topic by topic. Table 9 presents correlation coefficients between change in precision and the mistakes for each of the six query topics.

The figures presented in Table 9 show big differences from one topic to another. Some topics show a strange behaviour, for example, missed relevant has a positive correlation with change in performance for topic 2; false alarm presents a positive correlation with the change in performance for topics 1 and 6, and no correlation for topic 5.

The different behaviour of topics in response to the judgement mistakes indicates that the factors that affect the performance of RF vary from one topic to another. This fact seems analogous to the fact that different queries are better solved by different IR systems. In other words, if a system achieved a high precision in one query, that does not determine it will achieve a good result with another topic. Mandl and Womser-Hacker (2003) have also shown this when evaluating several CLEF runs. They observed a high standard deviation for the performance of the topics and a high standard deviation for the performance of each run. They concluded that no run performed well in all topics. Presently, there are no means for determining which type of topics will do better with which type of IR system.

In conclusion, this study was able to establish that the effects of misjudged documents are different for each topic. We have shown that the main source of impact on the change in performance produced by RF is the topics, and not the users or the systems. However, the characteristics of the topics that determine the relationship between change in performance and the misjudged documents remain unclear. An analysis of the characteristics of the topics would require a larger number of queries than used in this experiment.

7. Summary and conclusion

This paper reported experiments to evaluate RF in a CLIR system. Portuguese speakers were asked to judge the relevance of some documents returned in response to an initial query. The 27 participants recruited have assessed English documents, documents hand-translated to Portuguese, and documents automatically translated to Portuguese. The accuracy of such judgements was evaluated by comparing them to the official

relevance assessments provided by (CLEF). In addition the relationship between accuracy in judgement and the performance of RF was studied. The main findings are summarised below:

- Less than half (44%) of the participants were able to assess English documents.
- Machine Translation can indeed aid searchers in making relevance assessments, despite producing documents that are awkward to read. Participants judged machine translated documents with the same accuracy they judged hand-translated documents.
- There is a moderate negative correlation between the number of misjudged documents and the improvement that RF can provide.
- The factors that impact the change in performance vary greatly from one topic to another. Each topic responded differently to judgement errors. However, the characteristics of the topics that determine the relationship between change in performance and errors in judgement remain unclear.
- No relationship was found between the change in performance and the difficulty of the topics or the confidence in the assessments or the knowledge of the subject.
- Most participants consider a CLIR system very useful and would like the results translated into their native language.

Possibilities for future work include repeating the experiment using a different CLIR system, language-pair, and topics to assess whether the same results are achieved in different conditions.

Acknowledgements

We thank Páraic Sheridan for the helpful comments on this paper.
This work was partially supported by a CAPES-PRODOC Grant (Brazilian Government).

References

Ballesteros, L., & Croft, W. B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In *AAAI spring symposium on text and speech retrieval*. Stanford, CA: AAAI Press.

Ballesteros, L., & Croft, W. B. (1998). Statistical methods for cross-language information retrieval. In G. Grefenstette (Ed.), *Cross-language information retrieval* (pp. 23–40). Boston: Kluwer Academic Publishers.

Bathie, Z., & Sanderson, M. (2002). iCLEF at Sheffield. In C. Peters, M. Braschler, J. Gonzalo, & M. Kluck (Eds.), *Evaluation of cross-language information retrieval systems. Second workshop of the cross-language evaluation forum, CLEF 2001. Lecture notes in computing science* (vol. 2406, pp. 332–335). Rome, Italy: Springer.

Biron, P. V., & Kraft, D. H. (1995). New methods for relevance feedback: improving information retrieval performance. In: *ACM Symposium on Applied Computing*, Nashville, Tennessee.

Braschler, M. (2003). CLEF 2002—Overview of results. In C. Peters, M. Braschler, J. Gonzalo, & M. Kluck (Eds.), *Advances in cross-language information retrieval. Third workshop of the cross-language evaluation forum, CLEF 2002. Lecture Notes in Computing Science* (vol. 2785). Rome, Italy: Springer.

CLEF. Cross-language evaluation forum (2005). <http://www.clef-campaign.org> Accessed 28.10.2005.

Crestani, F. (2000). Neural relevance feedback for information retrieval. In B. Bouchon-Meunier, R. R. Yager, & L. A. Zadeh (Eds.), *Uncertainty in intelligent in information systems*. Singapore: World Scientific.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 1–13.

Dillon, M., & Desper, J. (1980). Automatic relevance feedback in boolean retrieval system. *Journal of Documentation, 36*, 197–208.

Drucker, H., Shahary, B., & Gibbon, D. (2001). Relevance feedback using support vector machines. In *18th International conference on machine learning (ICML)*.

Dumais, S. (1991). Improving retrieval of information from external sources. *Behaviour Research Methods, Instruments and Computers, 23*(2), 229–269.

Dumais, S. (1995). Using LSI for information filtering: TREC-3 experiments. In D. Harman (Ed.), *Third text retrieval conference (TREC-3): NIST*.

Dumais, S., & Nielsen, J. (1992). Automating the assignment of submitted manuscripts to reviewers. In *15th annual international ACM SIGIR conference on research and development in information retrieval*, Denmark.

Efthimiadis, E. (2000). Interactive query expansion: A user-based evaluation in a relevance feedback environment. *Journal of the American Society for Information Science, 51*(11), 989–1003.

Efthimiadis, E., & Robertson, S. (1989). Feedback and interaction in information retrieval. In C. Oppenheim (Ed.), *Perspectives in information management* (pp. 257–272). Butterworths.

Harman, D. (1992a). Relevance feedback and other query modification techniques. In W. B. Frakes & R. Baeza-Yates (Eds.), *Information retrieval data structures and algorithms* (pp. 241–263). Englewood Cliffs, NJ: Prentice-Hall.

Harman, D. (1992b). Relevance feedback revisited. In *15th annual international ACM SIGIR conference on research and development in information retrieval*, Denmark.

Ide, E. (1971). New experiments in relevance feedback. In G. Salton (Ed.), *The SMART retrieval system. Experiments in automatic document processing* (pp. 337–354). Englewood Cliffs, NJ: Prentice-Hall.

Karlgren, J., & Hansen, P. (2003). Cross-language relevance assessment and task context. In C. Peters, M. Braschler, J. Gonzalo, & M. Kluck (Eds.), *Advances in cross-language information retrieval. Third workshop of the cross-language evaluation forum, CLEF 2002. Lecture Notes in Computing Science* (vol. 2785, pp. 255–259). Rome, Italy: Springer.

Landauer, T., & Littman, M. (1990). Automatic cross-language document retrieval using latent semantic indexing. In *Sixth annual conference of the UW centre for the new oxford English dictionary and text research* (pp. 31–38). Waterloo, Canada.

Lesk, M., & Salton, G. (1968). Relevance assessments and retrieval system evaluation. *Information Storage and Retrieval, 4*, 343–359.

López-Ostenero, F., Gonzalo, J., Peñas, A., & Verdejo, F. (2001). Noun phrase translations for cross-language document selection. In C. Peters (Ed.), *Evaluation of cross-language information retrieval systems. Second workshop of the cross-language evaluation forum, CLEF 2001. Lecture Notes in Computing Science* (vol. 2406). Rome, Italy: Springer.

Mandl, T., & Womser-Hacker, C. (2003). Linguistic and statistical analysis of the CLEF topics. In C. Peters (Ed.), *Advances in cross-language information retrieval. Third workshop of the cross-language evaluation forum, CLEF 2002. Lecture Notes in Computing Science* (vol. 2785). Rome, Italy: Springer.

McNamee, P., & Mayfield, J. (2002). JHU/APL Experiments at CLEF: Translation resources and score normalization. In C. Peters, M. Braschler, J. Gonzalo, & M. Kluck (Eds.), *Evaluation of cross-languane information retrieval systems. Second workshop of the cross-language evaluation forum, CLEF 2001. Lecture Notes in Computer Science* (vol. 2406). Rome, Italy: Springer.

McNamee, P., & Mayfield, J. (2003). Scalable multilingual information access. In C. Peters, M. Braschler, J. Gonzalo, & M. Kluck (Eds.), *Advances in cross-language information retrieval. Third workshop of the cross-language evaluation forum, CLEF 2002. Lecture Notes in Computing Science* (vol. 2785). Rome, Italy: Springer.

Oard, D. W., & Gonzalo, J. (2001). The CLEF 2001 interactive track. In C. Peters, M. Braschler, J. Gonzalo, & M. Kluck (Eds.), *Second workshop of the cross-language evaluation forum. Lecture Notes in Computer Science* (vol. 2406). Rome, Italy: Springer.

Oard, D. W., & Gonzalo, J. (2003a). The CLEF 2002 interactive track. In C. Peters (Ed.), *Advances in cross-language information retrieval. Third workshop of the cross-language evaluation forum, CLEF 2002. Lecture Notes in Computing Science* (vol. 2785, pp. 245–254). Rome, Italy: Springer.

Oard, D. W., & Gonzalo, J. (2003b). The CLEF 2003 interactive track. In C. Peters (Ed.), *Working notes for the CLEF 2003 workshop*. Trondheim, Norway.

Orengo, V. M., & Huyck, C. R. (2003). Portuguese–English cross-language information retrieval using latent semantic indexing. In C. Peters, M. Braschler, J. Gonzalo, & M. Kluck (Eds.), *Advances in cross-language information retrieval. Third workshop of the cross-language evaluation forum, CLEF 2002. Lecture Notes in Computing Science* (vol. 2785). Rome, Italy: Springer.

Qu, Y., Eilerman, A. N., Jin, H., & Evans, D. A. (2000). The effect of pseudo relevance feedback on MT-based CLIR. In *Proceedings of content-based multimedia information access: Recherche d'informations assistee par ordinateur (RIAO 2000)*. Paris, France.

Robertson, S., & Sparck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science, 27*(3), 129–146.

Rocchio, J. J. (1971). Relevance feedback in information retrieval. In G. Salton (Ed.), *The SMART retrieval system. Experiments in automatic document processing* (pp. 313–323). Englewood Cliffs, NJ: Prentice-Hall.

Salton, G. (1971). Relevance feedback and the optimisation of retrieval effectiveness. In G. Salton (Ed.), *The SMART retrieval system. Experiments in automatic document processing* (pp. 324–336). Englewood Cliffs, NJ: Prentice-Hall.

Salton, G., & Buckley, C. (1997). Improving retrieval performance by relevance feedback. In K. Sparck Jones & P. Willet (Eds.), *Readings in information retrieval* (pp. 355–364). San Francisco, CA: Morgan Kauffmann.

Shen, X., Tan, B., & Zhai, C. (2005). Context-sensitive information retrieval using implicit feedback. In G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates, & N. Ziviani (Eds.), *28th annual international ACM SIGIR conference on research and development in information retrieval*. Salvador, Brazil: ACM.

Spink, A. (1994). Term relevance feedback and query expansion: Relation to design. In *17th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 81–90). Dublin, Ireland.

SYSTRAN. <http://www.systransoft.com/> Accessed 21.09.2005.

Voorhees, E. (1998). Variations in relevance judgements and the measurement of retrieval effectiveness. In W. B. Croft (Ed.), *21st annual international ACM-SIGIR conference on research and development in information retrieval* (pp. 315–323). ACM Press.

Wang, J., & Oard, D. W. (2001). iCLEF 2001 at Maryland. In C. Peters, M. Braschler, J. Gonzalo, & M. Kluck (Eds.), *Evaluation of cross-language information retrieval systems. Second workshop of the cross-language evaluation forum, CLEF 2001. Lecture Notes in Computing Science* (vol. 2406). Rome, Italy: Springer.

Xu, J., & Croft, W. B. (1996). Query expansion using local and global document analysis. In *19th annual international ACM SIGIR conference and research and development in information retrieval* (pp. 4–11).

Yang, Y., Carbonell, J., Brown, R. D., & Frederking, R. (1997). Translingual information retrieval. *Paper presented at the 15th international joint conference on artificial intelligence (IJCAI)*. Nagoya, Japan.