The Effect of Pseudo Relevance Feedback on MT-Based CLIR

Yan Qu, Alla N. Eilerman, Hongming Jin, David A. Evans
CLARITECH Corporation
5301 Fifth Avenue
Pittsburgh, PA 15232, USA
yqu@claritech.com

Abstract

In this paper, we identify factors that affect machine translation (MT) of a source query for cross-language information retrieval (CLIR) and empirically evaluate the effect of pseudo relevance feedback on cross-language retrieval performance. Our experiments demonstrate that, by using pseudo relevance feedback, we can significantly improve cross-language retrieval performance and achieve the level of monolingual retrieval.

1. Introduction

The goal of cross-language information retrieval (CLIR) is to enable a user to query in one language but perform retrieval across multiple languages. Many resources have been exploited for crossing the language boundary between the query language and the document language, e.g., machine translation (Gachot et al., 1998; Gey et al., 1999; Oard, 1999; Oard and Hackett, 1998), machine-readable bilingual dictionaries (Hull and Grefenstette, 1996; Ballesteros and Croft, 1998; Davis and Ogden, 1997), parallel or comparable corpora (Landauer and Litmann, 1990; Carbonell et al., 1997; Sheridan and Ballerini, 1996), and controlled languages (Deikema et al., 1999). We are interested in finding techniques that improve cross-language retrieval performance using the best commercially or publicly available resources.

In this paper, we discuss a particular approach to CLIR based on machine translation. Machine translation is an area of research that could address some of the issues of multilingual environments. Machine translation systems (e.g., SYSTRAN) are good resources for translating texts in several major world languages, and have been gaining increasing importance in the information space, such as the Internet. We apply the SYSTRAN machine translation system as our approach to cross the language barriers between the user’s query language and the document languages, and we empirically evaluate the effectiveness of a specific IR technique—pseudo relevance feedback—on retrieval performance. In the following sections, we first discuss related work in Section 2 and describe the MT-based query translation and query expansion methods in Section 3. Then, we describe our experiments in Section 4. In Section 5, we identify the factors that affect MT-based query translation, and determine the effectiveness of pseudo relevance feedback on cross-language retrieval. Finally, we summarize our work in Section 6.

2. Related Work

Machine translation has been used to bridge the language gap between the source language and the target languages in CLIR (Gachot et al., 1998; Gey et al., 1999; Oard, 1999; Oard and Hackett, 1998). Translation of a document collection from the document language to the query language involves translation of the complete document collection and storage of the translations. Because of the amount of translation required and the memory required for storing translations, this approach does not scale well for large document collections. The other approach is to translate a query from the query language to the document language. This approach is fast and straightforward, and has demonstrated retrieval performance comparable to the document translation approach (Oard and Hackett, 1998).

Relevance feedback (RF) is an approach to query expansion by which a query is modified using information from documents whose relevance to the query has been judged (Salton and Buckley, 1990). Typically, terms found in relevant documents are added to the query. Pseudo relevance feedback (PRF) differs from RF in that the former assumes the top retrieved documents to be relevant without using human judgments. Both RF and PRF have been demonstrated to improve performance in monolingual retrieval compared with not using such expansion techniques (Evans and Lefferts, 1994; Milic-Frayling et al., 1998).

Pseudo relevance feedback has been adapted to cross-language retrieval tasks. For example, Carbonell and colleagues (1997) applied PRF to CLIR based on a bilingual corpus. They first found the top-ranking documents for a query in the source language, then substituted the corresponding documents in the target language using a parallel bilingual corpus, and used these documents to form the corresponding query in the target language. Ballesteros and Croft (1998) adapted PRF to dictionary-based CLIR. They proposed three feedback methods for cross-language retrieval: before query translation, after query translation, and at both places. They have demonstrated that applying feedback prior to translation creates a stronger base query by adding terms that emphasize query concepts; applying feedback after translation reduces the effects of irrelevant query terms by adding more context specific terms; and using feedback at both places has the advantages of both methods. Pseudo relevance feedback has been shown to produce significant improvement in cross-language retrieval performance.

In our work, we adopt the query translation approach using machine translation. Although pseudo relevance feedback has been shown to improve retrieval performance both in monolingual retrieval and in cross-language retrieval using bilingual dictionaries, we are not aware of reports of its effectiveness in MT-based CLIR. Our hypothesis is that query expansion via PRF will improve the retrieval effectiveness compared to simple MT-based query translation. In particular, we hope to identify:

- Factors that affect the quality of MT-based query translation
- Effectiveness of pre-translation, post-translation, and combined (pre- and post-translation) feedback methods for query expansion

3. System Description

We adopted the CLARIT retrieval system for cross-language retrieval. The CLARIT retrieval system offers advanced information management functionalities, such as automatic indexing, retrieval, thesaurus extraction, natural language processing, and clustering (Evans and Lefferts, 1994; Milic-Frayling et al., 1998; Evans et al., 1999; Evans et al., 2000). For our work in CLIR, we added new components for managing multilingual resources and integrated machine translation into the retrieval system.

Figure 1 illustrates the simple MT-based query translation procedure and three pseudo relevance feedback methods for bilingual retrieval. Figure 1(a) illustrates simple query translation without expansion. In this configuration, the queries in a source language (SL) are translated using the MT engine into texts in the designated target language (TL), which are then used for retrieval from a target language database. Figure 1(b) illustrates query expansion prior to translation. Here each query in a source language is first augmented with N thesaurus terms extracted from the top M subdocuments retrieved from a source language database. The top M subdocuments are assumed to be relevant to the query. The original query text and the additional thesaurus terms in the source language are then sent to the MT engine. The resulting query, including the translated query text and thesaurus terms in the target language, is used for retrieval from a target language database. In post-translation query expansion illustrated in Figure 1(c), the original query text is first translated via the MT engine, then the translated query text is expanded using the pseudo feedback process. The combined feedback method unites the feedback process prior to translation illustrated in Figure 1(b) and the feedback process after translation illustrated in Figure 1(c).

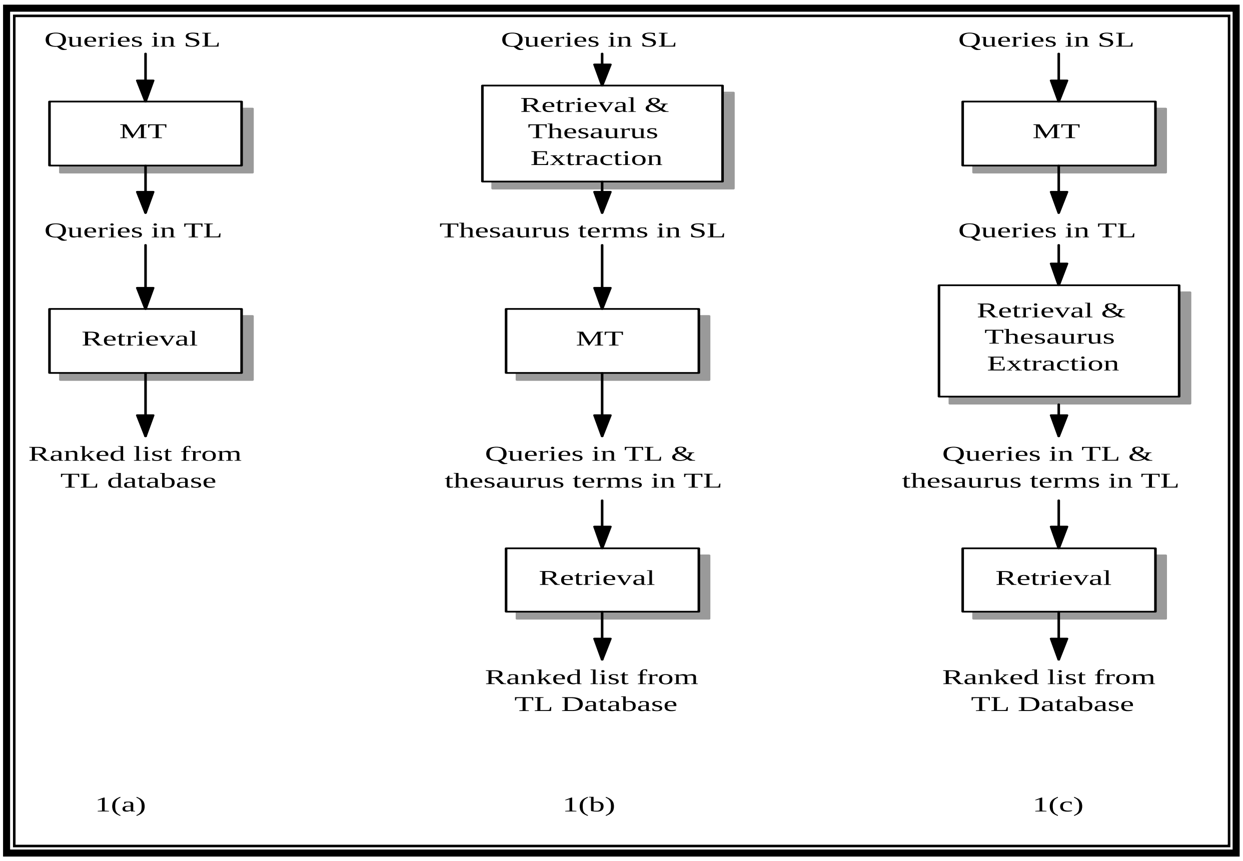

Queries in SL
MT
Queries in TL
Retrieval
Ranked list from TL database
1(a)
Queries in SL
Retrieval & Thesaurus Extraction
Thesaurus terms in SL
MT
Queries in TL & thesaurus terms in TL
Retrieval
Ranked list from TL Database
1(b)
Queries in SL
MT
Queries in TL
Retrieval & Thesaurus Extraction
Queries in TL & thesaurus terms in TL
Retrieval
Ranked list from TL Database
1(c)

Figure 1: CLIR with MT-based query translation and pseudo relevance feedback.

We use the SYSTRAN Enterprise software system for translating queries. The client-server configuration of this software allows us to integrate SYSTRAN's translation capability into our evaluation environment by calling the client API. The client API takes as input the source language query (plus feedback terms if feedback is used) stored in a file and the specific language pair for translation, and returns a file with the translation of the source text to the application program. Query translation process is a black box for the application program.

4. Experiments

We limit our report to experiments in English-to-French cross-language retrieval and baseline experiments in French monolingual retrieval. All experiments in this study were performed using the CLARIT CLIR evaluation environment presented in greater detail in Qu et al. (2000).

For processing the English corpus and queries, we used the CLARIT English NLP module, which consists of a parser and a morphological analyzer that use an English lexicon and grammar to identify linguistic structures in texts (Milic-Frayling et al., 1998). The CLARIT English NLP module supports discovery of various types of linguistic structures, such as simplex and complex noun phrases, verbs, and other selected constituents. For automatic query processing, we manually constructed a very short stop word lexicon to be used with the English core lexicon to filter out otherwise substantive words from the English query set (based on TREC topics) that are extraneous to the topics (e.g., *document*, *information*, *reference*, *relevant*, *optional*, *mention*, *report*). In the stop word lexicon, the words and their inflected forms were tagged with a special part of speech tag, which is discarded by the CLARIT parser. During the compilation step, the stop words overrode their original part of speech in the core lexicon, effectively excluding them from the set of indexing terms.

For processing the French corpus and queries, we have manually developed French language resources that are sufficient to achieve mostly correct phrase segmentation. These resources include a lexicon of closed-class categories with 1081 entries and a stop word list of nouns, adjectives and verbs from the French query set with 525 entries. Like the English stop word lexicon, the French stop word lexicon contains words and their inflected forms that are extraneous to the French topics (e.g., *document*, *information*, *pertinent*, *mention*, *mentionner*, *rapport*, *rapporter*, *recherche*, *rechercher*), and is used similarly to filter out stop words from the index terms. The English grammar was adapted to accommodate French categories. No French morphological normalization was done for the current experiments.

Query translation from English to French was done using the SYSTRAN Enterprise translation server, with English to French as the designated translation pair. No special or additional resources (e.g., customized dictionaries) were used to supplement the SYSTRAN default translation resources.

We conducted English-to-French cross-language retrieval experiments using the query topics and data collections from the TREC-6 CLIR track provided by the National Institute of Standards and Technology (NIST) (Voorhees and Harman, 1998). We took 22 English topics (1001-1007, 1009-1021, 1023, and 1024) as our experiment topics for English-to-French cross-language runs, and the corresponding French topics (2001-2007, 2009-2021, 2023, and 2024) for French monolingual runs.[1] The final English and French topic sets included only 22 topics each. All topics (both English and French) were composed of the title, description, and narrative fields of the TREC-6 English or French topics, and were processed automatically. The topics were equivalent across the languages, although some queries had a slightly different formulation in English and French. The average numbers of words (including stop words) in the source English topics and the ideal French topics were 43 and 51.3, respectively. Examples of English and French topics can be found in the Appendix. Term extraction for pseudo relevance feedback was based on an adaptation of the Rocchio formula (Milic-Frayling et al., 1998).

Evaluation was performed on the 250 MB collection of French SDA news (1988-1990) from the Schweizerische Depeschenagentur (Swiss News Agency). Training data for the pre-translation feedback experiments consisted of the 750 MB collection of the English AP news from the Associated Press covering the same period (1988-1990). Both the French evaluation corpus and the English training corpus were indexed on simplex noun phrases and all attested subterms. The English queries (and their French translations) and the ideal French queries were processed similarly into simplex noun phrases and decomposed into all attested subterms. The average number of English terms in a source English query is 17.3, while the average number of French terms in an ideal French query is 15.4.[2] For each TREC-6 topic, a set of French documents from the French SDA news was pre-judged by NIST for its relevance to the query. We used eleven-point average precision (N=1000 documents) as the basis of evaluation for all experiments. We also report precision at low recall levels (10, 20 and 100 documents), because these measures are more meaningful in an interactive setting.

[1] The English Topic 1008 and the corresponding French Topic 2008 were excluded because the English topic had no relevant English documents. The English Topic 1022 and the corresponding French Topic 2022 were excluded because the French topic had no relevant French documents. Topics 1025 and 2025 were excluded because they had no relevance judgments.

[2] It is worth noting the statistics for the English and French queries. Even though, on average, a French query contains more words than an English query (51.3 vs. 43), a French query generates fewer terms for retrieval by the system (15.4 vs. 17.3). On the one hand, the number of words in French queries may be higher because of the more frequent use of functional words (articles, prepositions, etc.) in French. On the other hand, English queries have a higher number of terms, which could result from phrasal terms. In English, phrasal terms are usually preserved as simplex noun phrases (e.g., *drug traffic),* while in French they are often broken down into individual words by prepositions (e.g., *trafic de stupéfiants*) and, therefore, are treated as complex noun phrases rather than simplex noun phrases by our current system.

We report the following experiment runs:

(1) F-nf — Automatic French monolingual run with no feedback, using the 22 ideal French topics.

(2) F-prf — Automatic French monolingual run with pseudo relevance feedback, using the 22 ideal French topics. The top 50 thesaurus terms selected from the top 75 subdocuments were used for pseudo relevance feedback.

(3) EF-nf — Automatic English-to-French cross-language run with no feedback using the 22 English topics.

(4) EF-prf-pre — Automatic English-to-French cross-language run with pre-translation pseudo relevance feedback using the 22 English topics. The top 50 thesaurus terms selected from the top 50 subdocuments were used for pseudo relevance feedback in the source language (English).

(5) EF-prf-post — Automatic English-to-French cross-language run with post-translation feedback using the 22 English topics. The top 50 thesaurus terms selected from the top 50 subdocuments were used for pseudo relevance feedback in the target language (French).

(6) EF-prf-comb — Automatic English-to-French cross-language run with both pre- and post-translation feedback using the 22 English topics. The top 50 thesaurus terms taken from the top 50 subdocuments were selected for pseudo relevance feedback in both the source language (English) and the target language (French). [3]

Runs F-nf and F-prf were designed as baselines for evaluating the effectiveness of the MT-based cross-language retrieval from English to French. The run F-prf represents the best retrieval performance of the automatic French monolingual retrieval by our current system. The run EF-nf was designed as a baseline run for evaluating the effectiveness of pseudo relevance feedback in the English-to-French cross-language retrieval.

5. Result Analysis

In this section, we present the analysis of our experimental results. First we report the result of cross-language retrieval with simple MT-based query translation and analyze the factors that affect MT-based query translation. We then determine the effectiveness of using pseudo relevance feedback for query expansion before query translation, after query translation, and at both places.

5. 1. Retrieval Using MT-Based Query Translation

We tested the effectiveness of bilingual retrieval using simple machine translation of a source English query with no feedback and identified factors that affected retrieval performance of the translated French query. The French monolingual experiment with ideal French queries and no feedback served as a baseline here. Table 1 compares the results of the English-to-French cross-language retrieval using MT-based query translation with no feedback (EF-nf) and the results of the French monolingual retrieval with no feedback (F-nf).

The comparison shows that the average precision achieved in the English-to-French cross-language retrieval with no feedback reached 73% of the French monolingual performance level. The recall reached 84% of the monolingual level. The precision at 10, 20, and 100 document cut-off points reached 84%, 83%, and 85% of the respective monolingual levels, while the exact precision reached 70% of the monolingual level.

[3] Note that in our current implementation of pseudo relevance feedback, we used the same settings of subdocument and term cutoff parameters for both the source language and the target language. In our future work, we plan to use different settings of feedback control parameters for the source language and the target language.

English-to-French Cross-Language Retrieval with No Feedback vs. French Monolingual Retrieval with No Feedback

	F-nf (baseline)	EF-nf	Percentage of baseline
RelRetDocCount	1006	845	84%
Recall	0.7306	0.6137	84%
Average Precision	**0.2548**	**0.1862**	**73%**
Precision at 10 Docs	0.3727	0.3136	84%
Precision at 20 Docs	0.3386	0.2818	83%
Precision at 100 Docs	0.1909	0.1618	85%
Exact Precision	0.3143	0.2213	70%

Table 1: English-to-French cross-language retrieval with no feedback
compared to French monolingual retrieval with no feedback.

5. 2. Translation Error Analysis

To determine the factors that affect the retrieval performance with MT-based query translation, we conducted an error analysis of the SYSTRAN translation outputs of the 22 English queries, focusing on the key terms in the queries and their translations. We identified the following seven types of translation errors:

E1: missing translation (i.e., an English term is not translated into French);
E2: unnecessary translation (i.e., a borrowed English term is translated literally into an equivalent French phrase);
E3: wrong sense disambiguation (i.e., a wrong translation equivalent is selected);
E4: wrong disambiguation caused by removed capitalization (i.e., without capitalization the term has a different meaning);
E5: word-by-word translation of a multiword (idiomatic) term;
E6: wrong phrase (i.e., combining words that do not belong to the same noun phrase);
E7: broken phrase (i.e., a simplex noun phrase is broken up by a preposition or another function word into single word terms, sometimes with an inappropriate part of speech choice).

Table 2 gives examples of each error type. The second column shows the number of occurrences of a particular error type in the 22 translated French queries. Note that several error types can occur simultaneously in one query translation. The third column shows the original English query term. The fourth column shows the corresponding term of an ideal French query. The fifth column displays the incorrect French translation of an English term produced by the SYSTRAN machine translation software.

Our analysis shows that the most prevalent error type in machine translation is the wrong disambiguation of a polysemous term (E3), which leads to an inappropriate choice of a translation equivalent. This suggests that even though machine translation represents the high-end knowledge-intensive approach to translation disambiguation, the quality of such disambiguation is still far from satisfactory. The other three frequent sources of errors are literal translation of an idiomatic phrase (E5), wrong combination of words (E6), and breaking up a simplex noun phrase into single word terms by inserting a preposition, often accompanied by an inappropriate part of speech choice for a modifying term (E7). Error types E5, E6, and E7 all involve multiword terms. Missing translation, unnecessary translation, and wrong disambiguation due to removed capitalization are rare for this particular set of queries. However, these error types may be more frequent in other domains, especially in texts with many proper nouns and abbreviations.

Error Type	Frequency	English Term	Ideal French Translation	SYSTRAN output
E1	1	agencies’	(des) agences	(d’)agencies
E2	1	fast food	fast food	aliments de préparation rapide “food of fast preparation”
E3	23	logging farming (to) stem “to stop or check”	déforestation “deforestation” culture “cultivation” contrôler “to control”	notation “notation” affermage “leasing, renting” tige “stem, stalk (of a plant)”
E4	2	aids (AIDS)	sida (SIDA) “AIDS”	aides “assistants; subsidies”
E5	7	death penalty solar powered cars third-world (countries)	la peine de mort voitures solaires le tiers monde	la pénalité de la mort voitures actionnées solaires le troisième-monde
E6	5	**austrian** president kurt waldheim’s **participation** **international** agencies’ **efforts**	la **participation** du président **autrichien** kurt waldheim les **mesures** des agences **internationales**	la **participation autrichienne** de waldheim de kurt de président des **efforts internationaux** d’agencies
E7	6	**austrian president kurt waldheim’s** participation nazi crimes sex education	la participation **du président autrichien kurt waldheim** crimes nazis (Adj) éducation sexuelle (Adj)	la participation **autrichienne de waldheim de kurt de président** crimes de nazi (N) éducation de sexe (N)

Table 2: Examples of translation errors produced by the SYSTRAN machine translation software.

It is difficult to determine precisely the effect of each error type on retrieval performance of a particular query, since different error types may co-occur in the same query. The performance often depends on whether the error affects a key term, and how much the contextual terms can compensate for the loss of meaning due to the translation error. In some cases, translation errors caused significant performance degradation. For example, in Topic 1023 (“Fast Food in Europe”), an E2 error affected the most important key term for this topic (*fast food*). In addition, there was an E3 error (*franchises* translated as *concessions*) in the same topic. These two errors combined caused a dramatic (99.9%) loss in average precision. In other cases, the negative effect of translation errors was insignificant or non-existent. For example, in Topic 1003 (“Drugs”), two terms (*to stem* and *to control*) were translated incorrectly (as *tige* “stem, stalk (of a plant)” and *commander* “to order, to command”), due to wrong disambiguation. However, the negative effect of these errors was negligible (1% loss in average precision), because the key terms (*drugs*, *drug traffic*, *arrests*, *seisures*, and others) were translated correctly (as *drogues*, *trafic de stupéfiants*, *arrestations*, *saisies*, etc.). Retrieval performance can also be affected by the choice of synonyms (*foi* “faith” vs. *religion* “religion”) and alternative spellings (*acuponcture* vs. *acupuncture*).

5. 3. Experiments with Pseudo Relevance Feedback

In order to determine the effectiveness of pseudo relevance feedback for enhancing queries in cross-language retrieval, we first examined the effectiveness of pseudo relevance feedback in monolingual retrieval performance. Specifically, we applied pseudo relevance feedback to monolingual retrieval using the set of ideal French queries against the French document collection. This experiment demonstrated a beneficial effect of pseudo relevance feedback on the French monolingual retrieval. Table 3 compares the results of French monolingual experiments with and without feedback. It shows that pseudo relevance feedback improved retrieval performance on all measures. Recall increased by 14% and average precision increased by 16% over the baseline.

French Monolingual Retrieval with Pseudo Relevance Feedback
vs. French Monolingual Retrieval with No Feedback

	F-nf (baseline)	**F-prf**	**Increase**
RelRetDocCount	1006	1147	14%
Recall	0.7306	0.8330	14%
Average Precision	**0.2548**	**0.2968**	**16%**
Precision at 10 Docs	0.3727	0.4273	15%
Precision at 20 Docs	0.3386	0.3523	4%
Precision at 100 Docs	0.1909	0.2236	17%
Exact Precision	0.3143	0.334	6%

Table 3: French monolingual retrieval with pseudo relevance feedback
compared to French monolingual retrieval with no feedback.

Then we conducted three experiments to test the effectiveness of pseudo relevance feedback for enhancing queries in cross-language retrieval. We applied pseudo relevance feedback to support query expansion prior to query translation (pre-translation), after query translation (post-translation), and at both points (combined). Table 4 compares the results of the cross-language retrieval experiments with different feedback methods (EF-prf-pre, EF-prf-post, and EF-prf-comb) and the baseline cross-language experiment with no feedback (EF-nf).

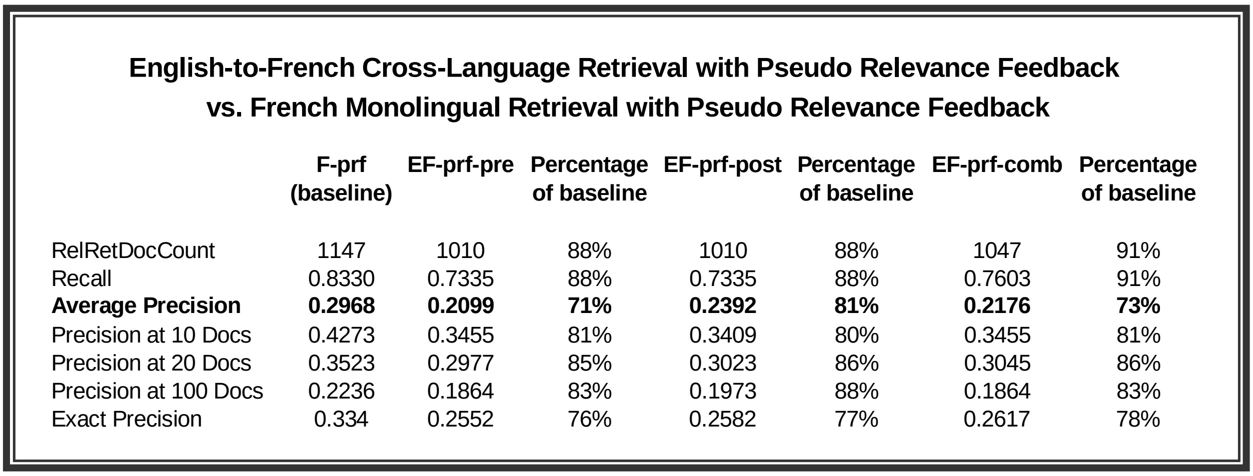
English-to-French Cross-Language Retrieval with Pseudo Relevance Feedback
vs. English-to-French Cross-Language Retrieval with No Feedback

	EF-nf (baseline)	**EF-pf-pre**	**Increase**	**EF-pf-post**	**Increase**	**EF-pf-comb**	**Increase**
RelRetDocCount	845	1010	19.5%	1010	19.5%	1047	23.9%
Recall	0.6137	0.7335	19.5%	0.7335	19.5%	0.7603	23.9%
Average Precision	**0.1862**	**0.2099**	**12.7%**	**0.2392**	**28.5%**	**0.2176**	**16.9%**
Precision at 10 Docs	0.3136	0.3455	10.2%	0.3409	8.7%	0.3455	10.2%
Precision at 20 Docs	0.2818	0.2977	5.6%	0.3023	7.3%	0.3045	8.1%
Precision at 100 Docs	0.1618	0.1864	15.2%	0.1973	21.9%	0.1864	15.2%
Exact Precision	0.2213	0.2552	15.3%	0.2582	16.7%	0.2617	18.3%

Table 4: English-to-French cross-language retrieval with different feedback methods
compared to English-to-French cross-language retrieval with no feedback.

All three feedback methods improved retrieval performance compared to the cross-language retrieval with no feedback. The highest average precision of 0.2392 (28.5% improvement) was achieved in the experiment with post-translation feedback. Combined feedback resulted in the second best result of 0.2176 (17% improvement). With pre-translation feedback, average precision was 0.2099 (13% improvement). The highest recall (0.7603) and the highest exact precision (0.2617) were observed in the experiment with combined feedback. Combined feedback also resulted in the best precision at 10 documents (0.3455, in a tie with the pre-translation method) and at 20 documents (0.3045), while post-translation feedback demonstrated the highest precision at 100 documents (0.1973).

Table 5 compares the results of the English-to-French cross-language retrieval using the three feedback methods (EF-prf-pre, EF-prf-post and EF-prf-comb) and the results of the French monolingual retrieval with pseudo relevance feedback (F-prf).



English-to-French Cross-Language Retrieval with Pseudo Relevance Feedback vs. French Monolingual Retrieval with Pseudo Relevance Feedback

	F-prf (baseline)	**EF-prf-pre**	**Percentage of baseline**	**EF-prf-post**	**Percentage of baseline**	**EF-prf-comb**	**Percentage of baseline**
RelRetDocCount	1147	1010	88%	1010	88%	1047	91%
Recall	0.8330	0.7335	88%	0.7335	88%	0.7603	91%
Average Precision	**0.2968**	**0.2099**	**71%**	**0.2392**	**81%**	**0.2176**	**73%**
Precision at 10 Docs	0.4273	0.3455	81%	0.3409	80%	0.3455	81%
Precision at 20 Docs	0.3523	0.2977	85%	0.3023	86%	0.3045	86%
Precision at 100 Docs	0.2236	0.1864	83%	0.1973	88%	0.1864	83%
Exact Precision	0.334	0.2552	76%	0.2582	77%	0.2617	78%

Table 5: English-to-French cross-language retrieval with different feedback methods compared to French monolingual retrieval with pseudo relevance feedback.

The comparison shows that the average precision achieved in the cross-language retrieval with pseudo relevance feedback reached 71%-81% of the French monolingual performance level, with the highest result observed in the experiment with post-translation feedback. The recall reached 88%-91% of the monolingual level, while the exact precision reached 76%-78% of the monolingual level. The combined feedback achieved both the highest recall and highest exact precision. The precision at 10, 20, and 100 document cut-off points reached 80%-81%, 85%-86%, and 83%-88% of their respective monolingual levels.

It is interesting to compare the results in Table 1 and Table 5, where the cross-language retrieval experiments are compared with their respective baseline monolingual runs. The monolingual baseline in Table 5 is higher than the baseline in Table 1 across all measures due to the positive effect of pseudo relevance feedback on monolingual retrieval. Even with a higher baseline, we observe that the cross-language runs with pseudo relevance feedback demonstrated less degradation of recall and exact precision than the cross-language runs without feedback (9-12% vs. 16% loss in recall and 22-24% vs. 30% loss in exact precision, in comparison with the monolingual performance levels). The average precision also suffered less with post-translation feedback than without feedback (19% vs. 27% loss, in comparison with the monolingual levels). The loss in average precision was a little higher with pre-translation feedback than without feedback (29% vs. 27% of the monolingual performance), but remained at the same level (27% of the monolingual performance) with the combined feedback.

When we compare all the French monolingual retrieval experiments and the English-to-French bilingual retrieval experiments as illustrated in Chart 1, we observe that, by using pseudo relevance feedback for cross-language retrieval, we can achieve a performance level that is very close to the monolingual retrieval performance with no feedback. While the best average precision and the best exact precision achieved in cross-language experiments are a little (by 6% and 16.7%, respectively) below the level of the monolingual retrieval with no feedback, the best cross-language retrieval precision at 100 documents exceeds the monolingual no-feedback level by a small (3.4%) margin. The recall in cross-language experiments with all feedback methods is 4-4.1% higher than the recall in the monolingual experiment with no feedback.

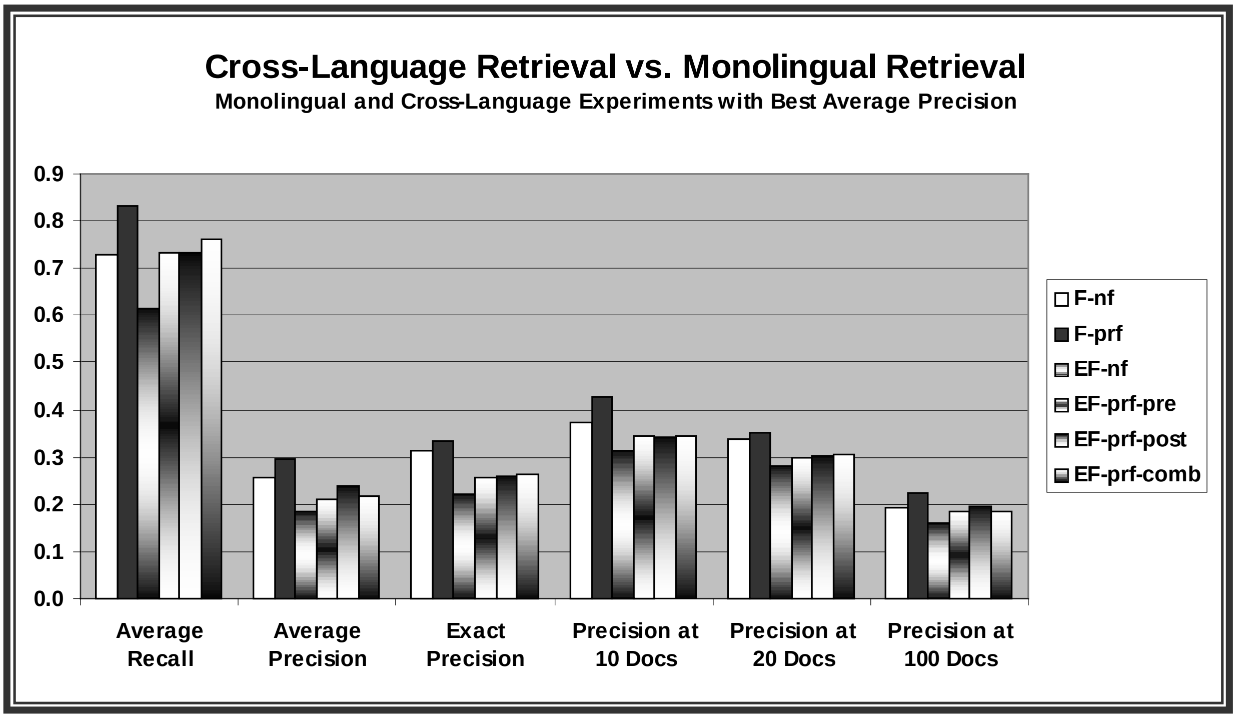


Chart 1: Comparison of retrieval performance in the French monolingual experiments and the English-to-French cross-language experiments.

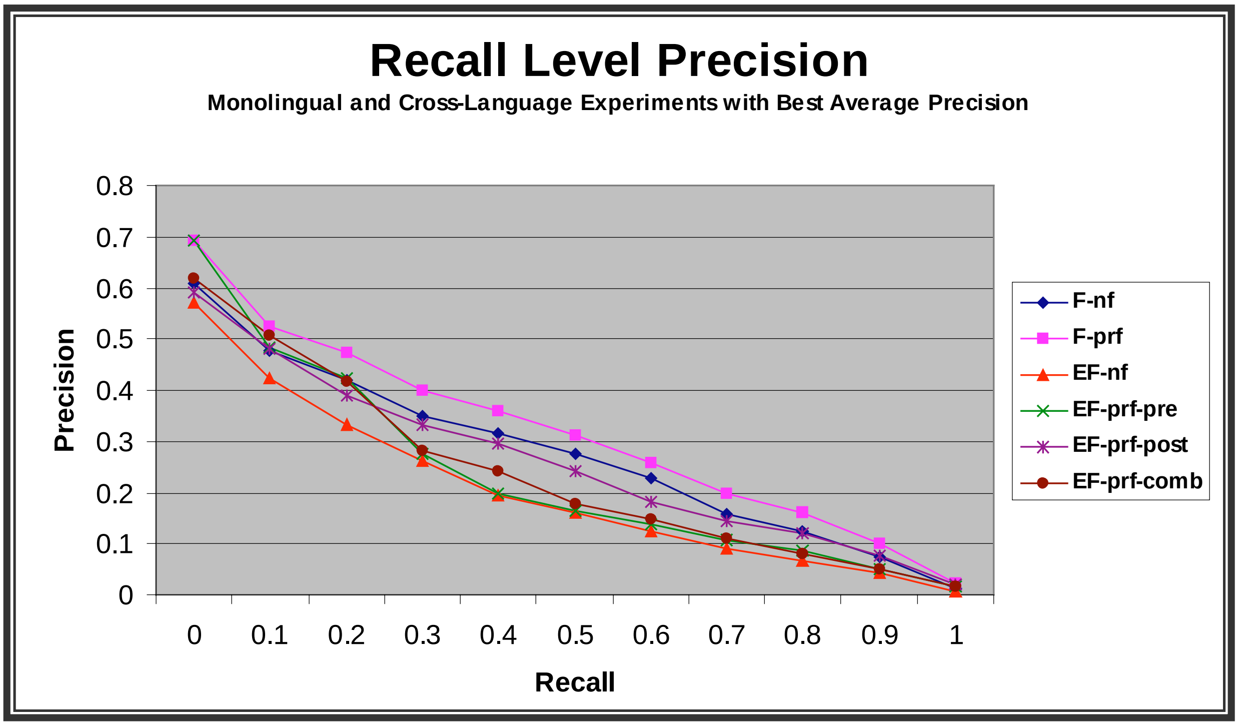


Chart 2: Comparison of the precision-recall curves for the French monolingual experiments and the English-to-French cross-language experiments.

Chart 2 compares the eleven-point recall precision curves for our monolingual and cross-language experiments. The top performance was achieved in the French monolingual retrieval with pseudo relevance feedback. At middle recall levels (0.3-0.8), the second best precision was observed in the French monolingual retrieval with no feedback. At low and high recall levels, the cross-language retrieval precision with feedback was a little higher than the monolingual no-feedback precision (with all three feedback methods at the 0.1 and 1 recall levels, with pre-translation feedback at the 0.2 recall level, and with post-translation feedback at the 0.9 recall level).

Comparison of the recall-precision curves for the cross-language experiments with different feedback methods shows that post-translation feedback outperforms the other two methods at middle and high recall levels (0.3–1), but pre-translation and combined methods perform better at low recall levels (0–0.2). Combined feedback is better than pre-translation feedback at the 0.3–0.7 recall points, but pre-translation feedback is a little better at the 0.2 and 0.8–1 recall points. All three feedback methods demonstrate higher precision than the cross-language retrieval with no feedback at all recall levels.

Overall, the experimental results suggest that pseudo relevance feedback is as effective in cross-language retrieval as in monolingual retrieval, and that, in most cases, it compensates for the negative effects of translation errors.

5. 4. The Effect of Different Feedback Methods on Cross-Language Retrieval for Individual Queries

We compared the effect of different feedback methods on cross-language retrieval for individual queries. Table 6 shows for how many queries each feedback method increased or decreased average precision over the baseline performance with no feedback. Post-translation feedback improved average precision for 18 queries, while each of the other two feedback methods improved average precision for 12 queries. In 12 out of 22 queries, post-translation feedback outperformed the other two feedback methods.

Pre-Translation Feedback	Post-Translation Feedback	Combined Feedback	Number of Queries
+	+	+	9
+	–	+	2
+	–	–	1
–	+	+	1
–	+	–	8
–	–	–	1

Table 6: Performance of different feedback methods for individual queries
("+" for average precision increase, "–" for average precision decrease).

Nine queries demonstrated improvement with all three methods. Eight of these queries had translation errors. We observed that the performance of different feedback methods tends to be very close when the important key terms are translated correctly (as in Topic 1003 "Drugs"). If, however, one or more key terms are translated incorrectly and the loss of meaning cannot be compensated for by the contextual terms, the improvement level largely depends on the place of feedback relative to the query translation. For example, in Topic 1009 ("Effects of Logging"), the key concept "logging" (="cutting down trees") was lost in the translated query, because of the wrong disambiguation ("logging" = "notation"). Feedback prior to the query translation neutralized the negative effect of the translation error by introducing useful thesaurus terms, which emphasized the central concepts of this topic (deforestation and its negative effects on the environment and climate). This resulted in a huge performance improvement (688% increase in average precision). Feedback after the query translation returned many useful terms, but also introduced a lot of noise caused by the wrong translation of the key term *logging* as *notation*. With this method, the increase in average precision was much lower (only 29%). Combined pre- and post-translation feedback eliminated the noise by creating a stronger base query prior to translation and further expanding it with appropriate terms after translation. The combined method achieved an improvement similar to pre-translation feedback (621% increase in average precision).

For two queries, both pre-translation and combined feedback improved average precision, while post-translation feedback lowered average precision. In Topic 1015, the key term *death penalty* was translated word by word as *la pénalité de la mort*, instead of the idiomatic phrase *la peine de mort*. Feedback prior to translation introduced many relevant English terms, most of which were translated correctly. The second feedback after translation further improved the query by finding relevant French words and phrases (*condamnation* “conviction”, *condamné* “convicted prisoner”, *exécutions* “execution”, *exécuté* “executed”, *peine capitale* “capital punishment”, *chaise électrique* “electric chair”, and others). Post-translation feedback alone did not introduce any terms specifically related to the topic of death penalty, because the key term (*peine de mort*) was missing from the translated query.

For one query (Topic 1002 “Marriages”, with a special focus on interfaith and international marriages), pre-translation feedback improved average precision, while post-translation and combined feedback lowered average precision. Although there were no translation errors, an important term, *religion*, was missing from the target query, because the English term *faith* was translated with a synonymous (although ambiguous) term *foi* “faith”. Pre-translation feedback found the term *religion* and a number of terms related to religion, while post-translation feedback found more terms related to nationality rather than religion. The term vector was improved by the combined feedback, but it was not as good as with pre-translation feedback alone.

Sometimes, the benefits of the pre-translation and combined feedback are reduced because both methods return many English person and place names which may be extraneous to the topic (as in Topic 1021 “Child Abuse”). Proper names are usually not translated by the machine translation software. In some cases, however, they are interpreted as common nouns, because capitalization is removed. For example, the first name *Jack* was incorrectly translated as *cric* “(car) jack”, and the last name *Barber* was incorrectly translated as *coiffeur* “hairdresser, barber”. Inappropriate and unnecessary translations of English proper names returned by the pre-translation feedback create noise in the final query vector.

For one query (Topic 1001 “Waldheim Affair”), both post-translation and combined feedback improved average precision, while pre-translation feedback did not have any effect on performance. This query contained several E7 errors. During the automatic translation of the source query and term vector, multiword English terms (*waldheim affair*, *austrian president kurt waldheim*, and *nazi war crimes*) were broken down into separate words by prepositions: *affaire de waldheim*, (*la participation*) *autrichienne de waldheim de kurt de président*, *crimes de guerre de nazi*. With pre-translation feedback, the target query vector contained only separate words (*affaire*, *waldheim*, *kurt*, *président*, *autrichienne*, *crimes*, *guerre*, *nazi*) and the wrong phrase *participation autrichienne*. In contrast to this, post-translation and combined feedback returned the appropriate phrases (*affaire waldheim*, *président autrichien kurt waldheim*, *président autrichien*, *kurt waldheim*, *président waldheim*, *crimes de guerre nazis*), which were included in the target term vectors, in addition to separate words.

For eight queries, post-translation feedback improved average precision, while pre-translation and combined feedback lowered average precision. Most of these queries had E3 errors, in addition to other errors. In these queries, translation errors affected important terms, but other key terms in the same queries were translated correctly and provided sufficient context for useful post-translation feedback. Pre-translation and combined feedback did not improve the performance for this query set, because English thesaurus terms were an additional source of translation errors and noise. On one hand, some relevant thesaurus terms may be translated incorrectly or left without translation. For example, pre-translation feedback for Topic 1016 ("Tuberculosis") returned several ambiguous terms (e.g., *cases*, *test*) and acronyms (e.g., *AIDS*, *CDC*, *HIV*). Automatic translation of the ambiguous terms resulted in the wrong disambiguation (*case* translated as *case* "box, compartment" or *caisse* "case, container", instead of *cas* "occurrence (of a disease)"; *test* translated as *essai* "trial, analysis", instead of *test* "medical test to detect virus, bacteria, etc."), while the acronyms were included in the final query vector without translation. On the other hand, feedback in the source language may contain extraneous terms that increase noise even though they are translated correctly. For example, pre-translation feedback for Topic 1010 ("Solar-Powered Cars") introduced many extraneous terms related to automobile air pollution and its effects on the environment and climate, while the query asked for information on "alternative energy sources to replace the continued exploitation of the world's finite fossil fuels". In such a case, combining pre-and post-translation feedback only exacerbates the problem, because it introduces even more extraneous terms during the second feedback stage. Post-translation feedback alone contains fewer translation errors and less noise.

Only one query (Topic 1020) demonstrated decreased average precision with all three feedback methods. The no-feedback experiment with this query achieved a very high average precision (0.6527), which was 2% higher than the monolingual performance for the corresponding French query (Topic 2020). There was only one translation error (E6), which did not affect the term vector. Although all feedback methods found many relevant thesaurus terms, the average precision decreased slightly. Pseudo relevance feedback also lowered average precision for the ideal French query. This suggests that when the quality of the original query is good and the quality of translation is good, pseudo relevance feedback does not contribute much to improve performance.

Our analysis has shown that the effectiveness of different feedback methods varies depending on the types of translation errors and the relative importance of the terms affected by these errors. Pre-translation and combined feedback can neutralize the effect of translation errors caused by incorrect disambiguation and literal translation of idiomatic phrases. The first feedback (prior to translation) creates a stronger base query in the source language, while the second feedback (after the translation) further improves the query by finding appropriate terms in the target language. On the other hand, pre-translation feedback may create noise by introducing extraneous terms, which are further amplified by the second feedback in the combined feedback method. Even if additional English terms are relevant for the topic, they may still be the source of translation errors because of their ambiguity. Pre-translation feedback often returns English proper names and acronyms that are either translated incorrectly (because of the removed capitalization) or included in the target vector without translation. All these factors reduce the effectiveness of the pre-translation and combined feedback. Post-translation feedback is effective when there is sufficient context in the translated query even if some terms are translated incorrectly. It often restores the key phrases that were broken into separate words during the query translation, and finds additional useful multiword terms. Post-translation feedback alone may fail to improve the query performance, especially when one or more important key terms are translated incorrectly and other query terms do not provide sufficient information about the query topic.

6. Summary and Future Work

In this work, we adopted pseudo relevance feedback for query expansion in cross-language retrieval with MT-based query translation. We experimented with three feedback methods: feedback prior to translation, after translation, and at both points. All feedback methods demonstrated significant performance improvement in cross-language experiments compared to not using feedback. Such results are consistent with reports of the effectiveness of pseudo relevance feedback in cross-language information retrieval using parallel corpora and bilingual dictionaries (Carbonell et al., 1997; Ballesteros and Croft, 1998). Post-translation feedback outperformed pre-translation and combined feedback for the majority of topics and demonstrated higher average levels of improvement for the set of twenty-two topics. The use of feedback in cross-language retrieval allowed us to achieve and even exceed monolingual no-feedback performance.

The effectiveness of different feedback methods depends on the types of translation errors and the relative importance of the terms affected by these errors. Our analysis of errors in the translated queries shows that wrong sense disambiguation and inappropriate translation of multi-word terms are the most frequent translation errors produced by a machine translation system. In most cases, the use of pseudo relevance feedback helps to reduce the negative effect of translation errors.

In our experiments, we used long topic statements expressed in sentences. It would be beneficial to determine the effect of query length and the effect of query formulation using our current approach. For example, how effective is this approach for short queries that do not contain much context, and how effective is this approach for word lists that are not syntactically well-formed? In our future work, we intend to investigate these issues and develop technologies to deal with wrong sense disambiguation and inappropriate translation of multi-word terms.

Acknowledgements

We thank Chengxiang Zhai and two anonymous reviewers for their helpful comments.

References

Ballesteros, L., & Croft, W.B. (1998). Statistical methods for cross-language information retrieval. In G. Grefenstette (Ed.), *Cross-Language Information Retrieval* (Chapter 3). Boston, MA: Kluwer Academic Publishers.

Carbonell, J., Yang, Y., Frederking, R., Brown, R.D., Geng. Y., & Lee, D. (1997). Translingual information retrieval: A comparative evaluation. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence* (pp. 708—714).

Davis, M., & Ogden, W. (1997). QUILT: Implementing a large-scale cross-language text retrieval system. In *Proceedings of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 92—98).

Diekema, A., Oroumchian, F., Sheridan, P., & Liddy, E.D. (1999). TREC-7 evaluation of Conceptual Interlingua Document Retrieval (CINDOR) in English and French. In E.M. Voorhees & D.K. Harman (Eds.), *Information Technology: The Seventh Text REtrieval Conference (TREC-7)* (NIST Special Publication 500-242, pp. 169-180). Washington, DC: U.S. Government Printing Office.

Evans, D.A., Bennett, J., Tong, X., Huettner, A., Zhai, C., & Stoica, E. (2000). CLARIT TREC-8 Manual Ad-Hoc Experiments. (To appear) In E.M. Voorhees & D.K. Harman (Eds.), *Proceedings of The Eighth Text REtrieval Conference (TREC-8).* Washington, DC: U.S. Government Printing Office.

Evans, D.A., Huettner A., Tong, X., Jansen, P., & Bennett, J. (1999). Effectiveness of Clustering in Ad-Hoc Retrieval. In E.M. Voorhees & D.K. Harman (Eds.), *Information Technology: The Seventh Text REtrieval Conference (TREC-7).* (NIST Special Publication 500-242, pp. 143-148). Washington, DC: U.S. Government Printing Office.

Evans, D.A., & Lefferts, R.G. (1994). Design and evaluation of the CLARIT-TREC-2 system. In D.K. Harman (Ed.), *Information Technology: The Second Text REtrieval Conference (TREC-2).* (NIST Special Publication 500-215, pp. 137-150). Washington, DC: U.S. Government Printing Office.

Gachot, D.A., Lange, E., & Yang, J. (1998). The SYSTRAN NLP browser: An application of machine translation technology in cross-language information retrieval. In G. Grefenstette (Ed.), *Cross-Language Information Retrieval* (Chapter 9). Boston, MA: Kluwer Academic Publishers.

Gey, F.C., Jiang, H., Chen, A., & Larson, R.R. (1999). Manual queries and machine translation in cross-language retrieval and interactive retrieval with Cheshire II at TREC-7. In E.M. Voorhees and D.K. Harman (Eds.), *Information Technology: The Seventh Text REtrieval Conference (TREC-7),* (NIST Special Publication 500-242, pp. 527-540). Washington, DC: U.S. Government Printing Office.

Hull, D.A., & Grefenstette, G. (1996). Querying across languages: a dictionary-based approach to multilingual information retrieval. In *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 49-57).

Landauer, T.K., & Littman, M.L. (1990). Fully automatic cross-language document retrieval using latent semantic indexing. In *Proceedings of the Sixth Conference of University of Waterloo Center for the New Oxford English Dictionary and Text Research* (pp. 31-38).

Milic-Frayling, N., Zhai, C., Tong, X., Jansen, P., & Evans, D.A. (1998). Experiments in query optimization, the CLARIT system TREC-6 report. In E.M. Voorhees and D.K. Harman (Eds.), *Information Technology: The Sixth Text REtrieval Conference (TREC-6)*. (NIST Special Publication 500-240, pp. 415-454). Washington, DC: U.S. Government Printing Office.

Oard, D.W. (1999). TREC-7 experiments at the University of Maryland. In E.M. Voorhees and D.K. Harman (Eds.), *Information Technology: The Seventh Text REtrieval Conference (TREC-7)*. (NIST Special Publication 500-242, pp. 541-546). Washington, DC: U.S. Government Printing Office.

Oard, D.W., & Hackett, P. (1998). Document translation for cross-language text retrieval at the University of Maryland. In E.M. Voorhees and D.K. Harman (Eds.), *Information Technology: The Sixth Text REtrieval Conference (TREC-6).* (NIST Special Publication 500-240, pp. 687-696). Washington, DC: U.S. Government Printing Office.

Qu, Y., Jin, H., Eilerman, A.N., Stoica, E., & Evans, D.A. (2000). CLARIT TREC-8 CLIR Experiments. (To appear) In E.M. Voorhees & D.K. Harman (Eds.), *Proceedings of The Eighth Text REtrieval Conference (TREC-8)*. Washington, DC: U.S. Government Printing Office.

Salton, G., & Buckley. C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41, 288-297.

Sheridan, P. & Ballerini, J.P. (1996). Experiments in multilingual information retrieval using the spider system. In *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 58-65).

Voorhees, E.M., & Harman, D.K. (Eds.). (1998). *Information Technology: The Sixth Text REtrieval Conference (TREC-6).* NIST Special Publication 500-240. Washington, DC: U.S. Government Printing Office.

Appendix

A typical topic from TREC consists of a title, description, and narrative. The title field usually consists of up to three words that best describe the topic. The description field gives a one-sentence description of the topic area. The narrative field states the criteria used to judge a document as relevant. In our experiments, the three fields are combined together for constructing a query.

```
<num> Number: CL1
<E-title> Waldheim Affair

<E-desc> Description:
Reasons for controversy surrounding Waldheim’s
World War II actions.

<E-narr> Narrative:
Revelations about Austrian President Kurt
Waldheim’s participation in Nazi crimes during
World War II are argued on both sides.  Relevant
documents are those that express doubts about the
truth of these revelations.  Documents that just
discuss the affair are not relevant.
```

Figure 2: An example of a source English topic.

```
<num> Number: CL1
<F-title> Affaire Waldheim

<F-desc> Description:
Raisons de la controverse à l'égard des
agissements de Waldheim pendant la deuxième
guerre mondiale.

<F-narr> Narrative:
Les révélations sur la participation du président
autrichien Kurt Waldheim aux crimes nazis pendant
la deuxième guerre mondiale font l'objet de
controverses.  Les documents pertinents font état
de doutes sur la culpabilité de Waldheim.  Les
articles qui ne font que mentionner l'affaire ne sont
pas valables.
```

Figure 3: An example of an ideal French topic.