Document Translation for Cross-Language Text Retrieval at the University of Maryland *

Douglas W. Oard and Paul Hackett
Digital Library Research Group
College of Library and Information Services
University of Maryland, College Park, MD 20742
{oard,pghtwoz}@glue.umd.edu

Abstract

The University of Maryland participated in three TREC-6 tasks: ad hoc retrieval, cross-language retrieval, and spoken document retrieval. The principal focus of the work was evaluation of a cross-language text retrieval technique based on fully automatic machine translation. The results show that approaches based on document translation can be approximately as effective as approaches based on query translation, but that additional work will be needed to develop a solid basis for choosing between the two in specific applications. Ad hoc and spoken document retrieval results are also presented.

1 Introduction

The principal goal of the University of Maryland's participation in the Sixth Text REtrieval Conference (TREC-6) was to evaluate the performance of a document translation strategy for Cross- Language Information Retrieval (CLIR). The Logos machine translation system[1] was used in a fully automatic mode for both document and query translation, and Inquery release 3.1p1 from the University of Massachusetts[2] was used for all runs. We participated in the Ad Hoc task as well in order to establish a baseline for the performance of this version of Inquery, and we also used Inquery for Quasi-Spoken Document Retrieval (QSDR) track runs in preparation for future work on speech-based information retrieval. No manual processing was done, and all of our runs were submitted in the automatic category.

2 Cross-Language Information Retrieval

Query translation has emerged as the most popular technique for CLIR, typically achieving between 50% and 75% of the retrieval effectiveness that is reported for comparable monolingual techniques when coupled with simple linguistic processing such as part-of-speech tagging or phrase indexing [4]. Query translation strategies are relatively efficient when short queries are presented, but a lack of adequate linguistic context in queries containing only a few words may limit the ability of systems to select the most appropriate translations for the query terms. Machine translation systems seek to exploit contextual clues in full-length documents to produce the best possible translations, and it is an open question whether a retrieval system based on automatic machine translation of each document can outperform query translation. We have thus sought to determine whether the additional effort required to translate every document would produce better retrieval effectiveness than query translation for the TREC-6 CLIR track.

The Logos machine translation system that we used for our experiments is a commercial product that is designed to assist human translators by automatically preparing fairly good translations of individual

*This work has been supported in part by DARPA contract N6600197C8540 and the Logos Corporation.
[1]Logos Corporation, 111 Howard Boulevard, Suite 214, Mount Arlington, NJ 07856 USA
[2]Center for Intelligent Information Retrieval, University of Massachusetts, Amherst, MA 01003

Technique	Title	Short	Long
Unstopped Monolingual	0.2480	0.1219	0.2396
Document Translation	0.1761	0.1829	0.2171
Query Translation	0.1668	0.1492	0.1561

Table 1: Non-interpolated average precision for the SDA/NZZ collection, averaged over 21 topics.

documents. The system is typically used by translation bureaus and other organizations as the first stage of a machine-assisted translation process, and we have previously used it for cross-language routing experiments [3]. The Logos system includes extensive facilities for adding domain-specific technical terminology and new linguistic constructs, but for TREC-6 we used only the machine readable dictionaries and semantic rules that are delivered as standard components of the product. The entire SDA and NZZ collections were translated from German into English, and only format-related preprocessing and postprocessing was performed. A brief description of the translation process is contained in Appendix A. The translated documents are available to TREC participants through the NIST FTP site, and the README file with those documents contains sufficient detail to reproduce the translation runs.

We used four SPARC 20 workstations and a fifth workstation that was upgraded from a SPARC 5 to a SPARC Ultra 1 after about three quarters of the documents had been translated. All of the workstations were shared with other users. Translation of the 48 months of news stories contained in the SDA and NZZ collections using these machines required approximately 2 months. About half of the CPU time was required to perform the translations themselves, the remainder being shared with other users of the same machines or lost due to operator- or system-induced problems. Even with these problems, this works out to a single-machine translation rate that is at least 5 times faster than the rate at which the news articles were originally generated.

Once all of the documents had been translated into English, a single Inquery index was built for the union of the SDA and NZZ collections. Index construction required a two hours on a dedicated Sparc 20, and retrieval results for all 25 queries were typically computed in a few minutes (varying slightly with query length). Approximately 5% of the translations, almost entirely NZZ documents, were unavailable when the original index was constructed, but those translations have been subsequently completed and are included in the corrected runs presented here. Appendix C relates these corrected runs to the official results scored by NIST.

Table 1 summarizes the non-interpolated average precision results for three retrieval approaches, averaged over the 21 topics for which relevant documents are known in the SDA/NZZ collection, and Figure 1 shows recall-precision graphs for the same data.[3] Three query lengths were used: only words appearing in the title field ("title"), only words appearing in the desc field ("short"), and all words appearing in the topic description except SGML markup ("long"). As Table 1 shows, short queries were not as good as titles alone, and a query-by-query analysis revealed greater variation across topics for short queries as well. We used words from the title field in both our "title" and "long" queries, and it is possible that omitting those (usually very informative) words from our "short" queries offset any improvement that might otherwise have resulted from extending the length of the query. In Figure 1 and what follows we have chosen to focus on title and long queries since including short queries would likely contribute more to clutter than to clarity.

The monolingual retrieval results in Figure 1 provide a useful baseline for evaluating cross-language retrieval performance. In those runs we used the untranslated SDA/NZZ document collection and the German queries. We did not have a German stemmer available, but we did construct a small stopword list (see Appendix B). As Figure 2 shows, the use of that German stopword list adversely impacted long queries and had no impact on title queries, so we have presented only unstopped results when using German documents.

For the document translation runs we used the Logos translations of the SDA/NZZ documents into

[3]No relevant documents are known in the German SDA/NZZ collection for topic CL22, and relevance judgments are not available for topics CL03, CL15 and CL25.

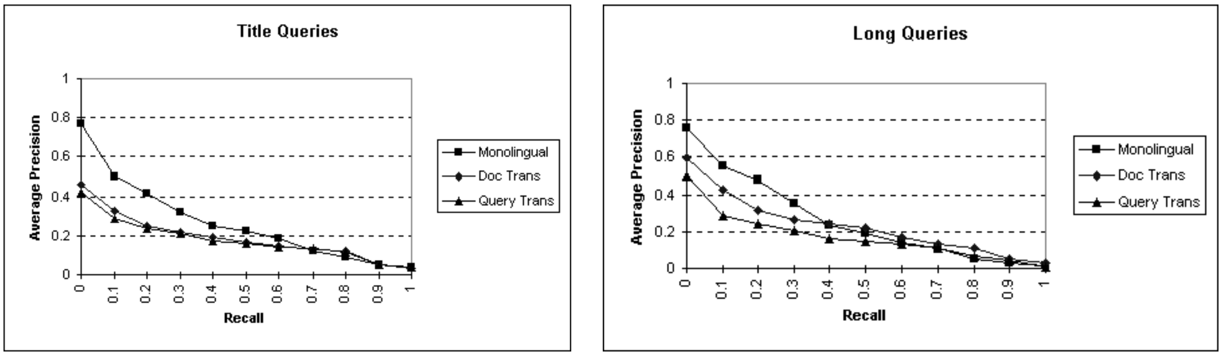
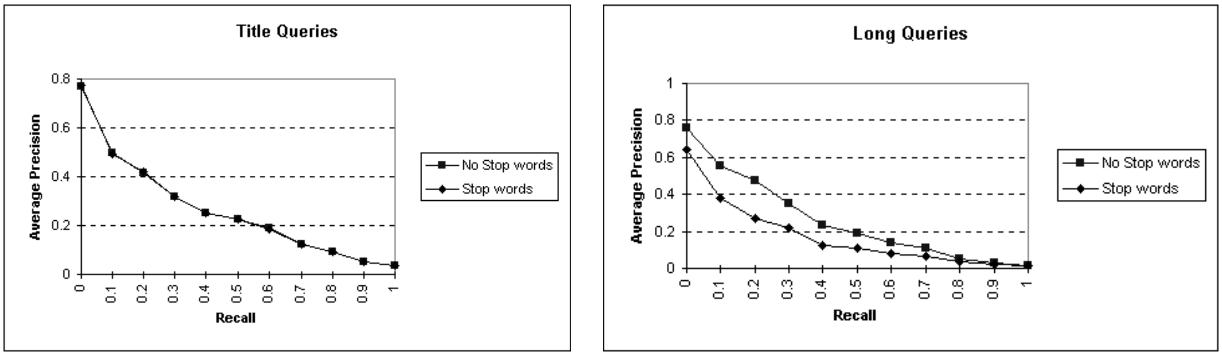


Figure 1: Comparison of retrieval approaches on the SDA/NZZ collection.



Title queries | Long queries

Figure 2: Recall-precision for monolingual retrieval on the SDA/NZZ collection with and without stopwords.

English and the English queries. Unlike the monolingual runs, both stemming and stopwords were used for the document translation runs. We used the Inquery "kstem" stemmer and Inquery's standard English stopword list. All other Inquery parameters were identical between the two sets of runs.

The SDA/NZZ query translation runs were made by using Logos to translate the English queries into German. The resulting queries were then used to retrieve untranslated SDA/NZZ documents. Again, Inquery was used without stemming or stopwords when processing German documents. Since Logos generates only a single "best guess" translation for any input, this approach differs in an important way from the more common approach based on cross-language query expansion. Cross-language query expansion techniques typically seek to replace each term in the query with every reasonable translation, including more than one possibility whenever unresolvable ambiguity is present [2]. By contrast, in the face of ambiguity Logos will simply choose whatever appears to be the best single translation.

Figure 1 shows that document translation and query translation perform about equally well on title queries, but that some advantage for document translation is apparent for long queries. Figure 3 depicts this result another way, showing the gain in uninterpolated average precision that results from using document translation rather than query translation on a query-by-query basis. Topic CL19 appears to account for much of the improvement in the long queries. It is difficult to draw strong conclusions from these results alone because the Logos "winner take all" approach to query translation has not been previously evaluated, but it does appear that document translation is performing at least as well as query translation and that both approaches are performing creditably, with results for title and long queries ranging between 67% and 90% of monolingual average precision on the SDA/NZZ collection.

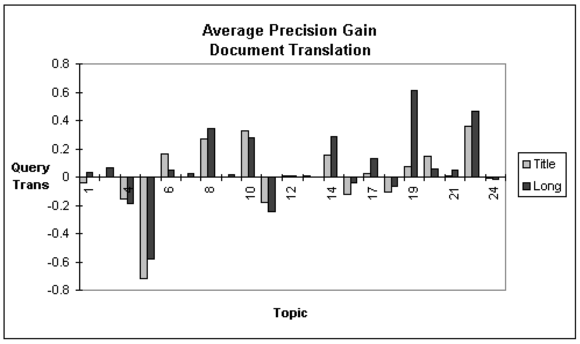


Figure 3: Relative advantage of document translation on the SDA/NZZ collection.

Technique	Title	Short	Long
Stopped Monolingual	0.3449	0.3121	0.3958
Query Translation	0.1928	0.1975	0.2455

Table 2: Non-interpolated average precision for the AP collection, averaged over 21 topics.

We did not try document translation on the CLIR track English AP collection, but we have obtained query translation and monolingual retrieval results for that collection using the untranslated AP documents, the "kstem" stemmer, and the standard Inquery stopword list. Table 2 and Figure 4 show those results. The monolingual results were obtained using English queries, while the query translation results were obtained with queries translated from German into English by Logos. Not surprisingly, a comparison of the results in Table 2 with those in Table 1 shows that retrieval effectiveness varies substantially across document collections, even when the same topics are used.

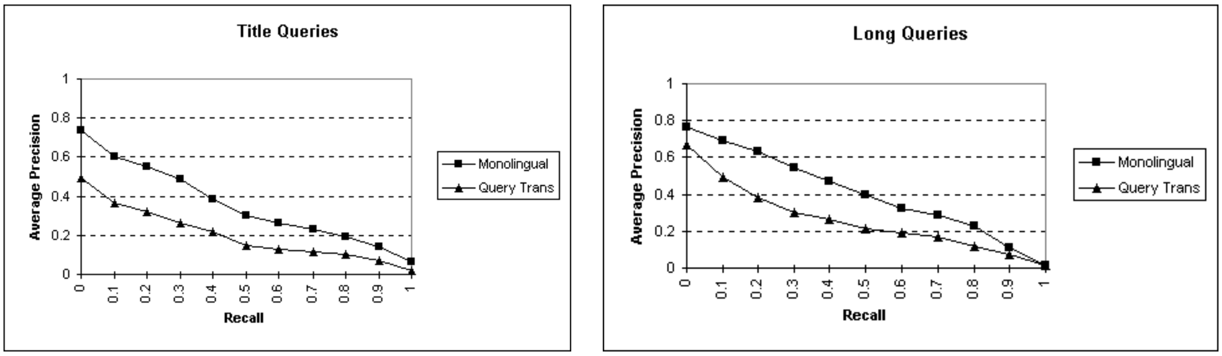


Figure 4: Query translation and monolingual retrieval results for the AP collection.

3 Ad Hoc Task

We used our participation in the ad hoc retrieval task to characterize the performance of our Inquery configuration in comparison with a broad range of participating systems. We submitted a single category A run with short queries based solely on the description field of each topic. Except for some content-neutral preprocessing to handle differing SGML markup, we used the same Inquery configuration for the ad hoc task that we used for our cross-language runs. The resulting non-interpolated average precision, averaged over 50 topics, was 0.1460. As Figure 5 shows, we achieved at or above median average precision for 33 of the 50 topics. It is difficult to draw strong inferences from this, however, given the general dissatisfaction with the performance of short queries on the ad hoc task this year. This was our first Category A submission, and we learned the usual lessons about the consequences of initially allocating far too little time and not quite enough disk space to the effort. We had no prior experience with Inquery and we estimate our overall effort to produce these results at 1 person-month. Based on installation effort and retrieval effectiveness, our assessment is that Inquery offers a practical alternative to the SMART version 11.0 system that we used in TREC-5 for modular cross-language retrieval experiments in which the translation and retrieval components are loosely coupled. We have not yet explored the Inquery API in sufficient detail to assess whether it will be practical to use Inquery to investigate more tightly coupled approaches in which unresolvable translation ambiguity must be preserved.

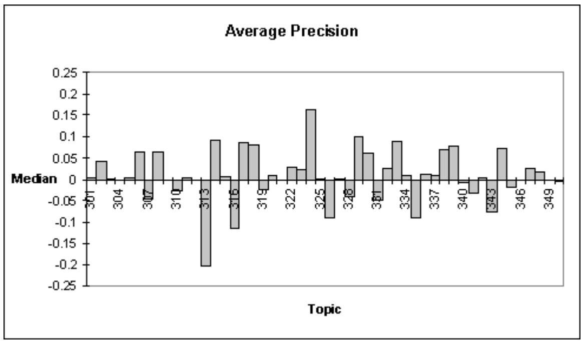


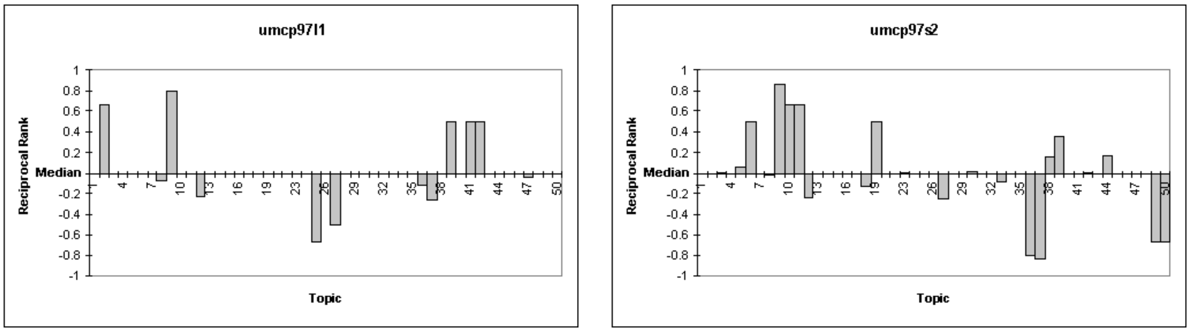
Figure 5: Monolingual retrieval for the ad hoc task.

4 Spoken Document Retrieval

We have recently initiated a project to investigate user interface design for information retrieval systems that provide access to large collections of recorded speech [5], and the Spoken Document Retrieval (SDR) track offered our first opportunity to gain experience with content-based retrieval using speech recognition output. We used Inquery to produce both a reference run from the transcripts and a QSDR run on the baseline recognizer output. Except for format-specific preprocessing, we made no other changes to our Inquery configuration for those runs. Figure 6 shows relative reciprocal ranks for our reference transcript and baseline recognizer runs, compared with the median reciprocal rank for each case. As Figure 7 illustrates, retrieval effectiveness declined substantially on about one quarter of the topics when the baseline recognizer output was substituted for the manually prepared transcripts.

5 Future Work

We are interested in exploring whether further improvements in cross-language retrieval effectiveness can be achieved by using the sort of linguistic analysis found in modern machine translation systems, but retaining any unresolvable ambiguity in a manner that can be effectively used by a text retrieval system. We are considering two approaches to this problem, one based on the extraction of intermediate representations from an existing machine translation system, and a second based on incorporation of more sophisticated linguistic representations into the retrieval system itself. This later approach has produced disappointing results in monolingual retrieval applications (c.f., [6]), but we believe that the presence of translation ambiguity in cross-language retrieval transforms the problem into one for which more sophisticated representations may



Reference transcript | Baseline recognizer

Figure 6: Speech Data Retrieval results — reciprocal rank vs. median reciprocal rank by query.

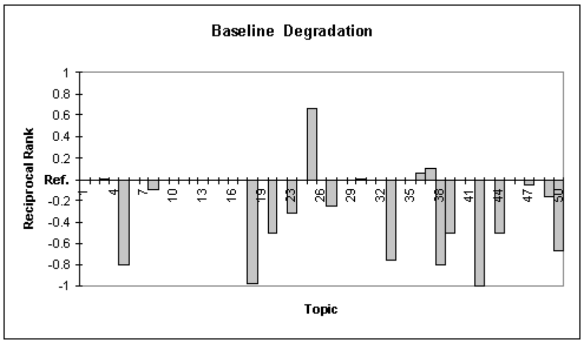


Figure 7: Degradation in reciprocal rank due to recognition errors.

be useful. Both of the approaches that we are considering should be able to exploit the linguistic context that is present in either documents or long queries, so both lead us in the direction of further experiments on cross-language retrieval based on document translation.

6 Conclusions

We have shown that document translation is a practical approach for cross-language text retrieval on moderately large collections, and we have observed some indications that document translation may ultimately be more effective than query translation for some applications. It appears that the CLIR test collection that has been developed at TREC-6 will be extremely useful for further investigation of these issues, and that is undoubtedly the most important legacy of this work. By providing a standard benchmark for evaluating the performance of competing approaches, the CLIR track has provided a sound basis for further advances in cross-language information retrieval.

Acknowledgments

The authors are grateful to Bonnie Dorr for providing facilities, resources and advice, to Scott Bennett and Harriet Leventhal for their assistance with the Logos translation system, to the University of Massachusetts for the use of Inquery, and to James Allan for help with Inquery configuration.

Appendices

A Document Translation Process

The translations were performed completely automatically using release 7.8.1 (or, for some NZZ documents, release 7.8.2) of the Logos machine translation system. System parameters were selected to use all available dictionaries and to maintain the imperative form where possible, but no new dictionaries were created for this purpose. The output was converted to the ISO 8859-1 (Latin-1) character set. Words that were not recognized by the Logos machine translation system were maintained in the original German, but characters with diacritical marks were mapped to the corresponding unmarked character.

In the SDA collection, only the LD, TI, TB, and TX fields were translated and indexed. In the LD field, the portion of the first line preceding the first ")" character was not translated. A total of 55 SDA documents failed to translate at all due to system errors. Those documents were removed from the translated collection but the corresponding untranslated documents were retained for the monolingual and query translation runs.

In the NZZ collection, the INDENT_TEXT, FOOTNOTE, TEXT, MAIN_TITLE, MAIN_TITLE_1, KURSIV_TITLE, KURSIV_TITLE_1, KURSIV_TITLE_2, LEAD, LINE_TITLE, LEGEND, MAGAZINE_TITLE, HEAD_TITLE, HEAD_TITLE_1, POETRY_TEXT, COLUMN_TITLE, SIDEHEAD_TEXT, FOOT_TITLE, FOOT_TITLE_1, FOOT_TITLE_2, INTRO_PARA, QUOTATION, SECTION_TITLE, and SECTION_TITLE_1 fields were translated and indexed. A total of 174 NZZ documents failed to translate due to system errors. Those documents were removed from the translated collection that was used for the document translation runs but the corresponding untranslated documents were retained for the monolingual and query translation runs.

B German Stopword List

The German stopword list that we tried for monolingual German runs was constructed by manually selecting stopwords from the German lexicon described in [1]. Terms were selected from prepositions, other functional elements, complementizers, pronouns, and a few contractions and other words, and selections were made by the developer of the lexicon, a non-native speaker of German. The following list contains every word in our stopword list:

ab aber alle allen aller am an andere anderem anderen anderer anderes ans auf auf aufwaerts aus bei beim das dein dem den denn der des dich die diese diese diesem diesen dieser dieser dieses dir drei dreie dreien dreier du du ein ein eine einem einen einer eines einige einigen einiger er es es euch euer für heraus herein herunter hinaus hinein hinter hinunter ich ihm ihn ihnen ihr im in ins jede jedem jeden jeder jedes jemand jemand jene jenem jenen jener jenes keine keinem keinen keiner keines man mein mein mich mir mit nach neben niemand niemand ob ohne sein selbst sich sich sie sie sie so über um und uns uns unser unser unter unter verschiedene verschiedenen verschiedener viele vielen vieler von vor wann warum was wegen weil weil welche welchem welchen welcher welches wem wen wer wes wessen wie wieviele wievielem wievielen wievieler wievieles wir wo zehn zu zu zum zur zwei zweie zweien zweier

C Official TREC Runs

Translations for approximately one sixth of the NZZ documents (scattered throughout the year) were not available in time for the official TREC submission, so those documents were not present in the translated collection that was used for the document translation runs. Formatting errors in the construction of two long English queries also resulted in submission of one official run without any selections for those topics. The results presented above reflect the corrected runs. Table 3 shows the correspondence of those runs to the identifiers of the official TREC runs.

D CLIR Track Questionnaire

```
1. OVERALL APPROACH:
```

Identifier	Collection	Queries	Approach	Remarks
umcpxgg1	SDA/NZZ	Title	Stopped monolingual	
umcpxgg2	SDA/NZZ	Short	Stopped monolingual	
umcpxgg3	SDA/NZZ	Long	Stopped monolingual	
umcpxgg4	SDA/NZZ	Title	Unstopped monolingual	
umcpxgg5	SDA/NZZ	Short	Unstopped monolingual	
umcpxgg6	SDA/NZZ	Long	Unstopped monolingual	
umcpxeg1	SDA/NZZ	Title	Document translation	
umcpxeg2	SDA/NZZ	Short	Document translation	
umcpxeg3	SDA/NZZ	Long	Document translation	Added CL12 and CL17
none	SDA/NZZ	Title	Query translation	New run
none	SDA/NZZ	Short	Query translation	New run
none	SDA/NZZ	Long	Query translation	New run
none	AP	Title	Stopped monolingual	New run
none	AP	Short	Stopped monolingual	New run
none	AP	Long	Stopped monolingual	New run
umcpxge1	AP	Title	Query translation	
umcpxge2	AP	Short	Query translation	
umcpxge3	AP	Long	Query translation	

Table 3: Official TREC identifiers corresponding to the corrected runs.

```
  1.1 What basic approach do you take to cross-language retrieval?
      [X] Document Translation

  1.2 Were manual translations of the original NIST topics used as a
      starting point for any of your cross-language runs?
      [X] No

  1.3 Were the automatically translated (Logos MT) documents used
      for any of your cross-language runs?
      [X] Yes, umcpxeg1, umcpxeg2, umcpxeg3

  1.4 Were the automatically translated (Logos MT) topics used
      for any of your cross-language runs?
      [X] Yes, umcpxge1, umcpxge2, umcpxge3

2. MANUAL QUERY FORMULATION:  N/A

3. USE OF MANUALLY GENERATED DATA RESOURCES:

  3.1 What kind of manually generated data resources were used?
      [X] Part-of-speech Lists (for stopword list development)

  3.2 Were they generated with information retrieval in mind or were
      they taken from related fields?
      [X] Machine Translation

  3.3 Were they specifically tuned for the data being searched (i.e.,
      with special terminology) or general-purpose?
```

[X] General purpose

3.4 What amount of work was involved in adapting them for use in your information retrieval system.
[X] 15 minutes

3.5 Size: See Appendix B

3.6 Availability: The source of the original part of speech list is cited in paper, the stopword list is provided in Appendix B.

3. USE OF MANUALLY GENERATED DATA RESOURCES:

3.1 What kind of manually generated data resources were used?
[X] Other, Logos MT

3.2 Were they generated with information retrieval in mind or were they taken from related fields?
[X] Machine Translation

3.3 Were they specifically tuned for the data being searched (i.e., with special terminology) or general-purpose?
[X] General purpose

3.4 What amount of work was involved in adapting them for use in your information retrieval system.
[X] 1 week

3.5 Size
[X] Est. 40,000 word dictionary

3.6 Availability? - Please also provide sources/references!
[X] Commercial, cited in paper.

4. USE OF AUTOMATICALLY GENERATED DATA RESOURCES: N/A

5. GENERAL

5.1 How dependent is the system on the data resources used? Could they easily be replaced if better sources were available?
[X] Easily replaceable

5.2 Would the approach used potentially benefit if there were better data resources (e.g. bigger dictionary or more/better aligned texts for training) available for tests?
[X] Yes, somewhat

5.3 Would the approach used potentially suffer a lot if similar data resources of lesser quality (noisier dictionary, wrong domain of terminology) were used as a replacement?
[X] Yes, somewhat

```
5.4 Are similar resources available for other languages than those
    used?
    [X] Yes, analysis in German and English, generation in
        German, English, Italian, French, Spanish
```

References

[1] Bonnie J. Dorr. *Machine Translation: A View from the Lexicon.* MIT Press, Cambridge, MA, 1993.

[2] David A. Hull and Gregory Grefenstette. Experiments in multilingual information retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996. http://www.xerox.fr/people/grenoble/hull/papers/sigir96.ps.

[3] Douglas W. Oard. Adaptive filtering of multilingual document streams. In *Fifth RIAO Conference on Computer Assisted Information Searching on the Internet*, June 1997. http://www.glue.umd.edu/~oard/research.html.

[4] Douglas W. Oard. Alternative approaches for cross-language text retrieval. In *AAAI Symposium on Cross-Language Text and Speech Retrieval.* American Association for Artificial Intelligence, March 1997. http://www.glue.umd.edu/~oard/research.html.

[5] Douglas W. Oard. Speech-based information retrieval for digital libraries. Technical Report CS-TR-3778, University of Maryland, College Park, March 1997. http://www.glue.umd.edu/~oard/research.html.

[6] Mark Sanderson. Word sense disambiguation and information retrieval. In W. Bruce Croft and C. J. van Rijsbergen, editors, *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 142–151. Springer-Verlag, July 1994. http://www.dcs.gla.ac.uk/ir/papers/Postscript/sanderson94b.ps.gz.