CLEF Experiments at Maryland: Statistical stemming and backoff translation

Douglas W. Oard[1], Gina-Anne Levow[2], and Clara I. Cabezas[3]

[1] College of Information Studies & Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742 USA,
oard@glue.umd.edu, WWW home page: http://www.glue.umd.edu/~oard/
[2] Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742 USA,
gina@umiacs.umd.edu, WWW home page: http://umiacs.umd.edu/~gina/
[3] Department of Linguistics & Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742 USA,
clarac@umiacs.umd.edu

Abstract. The University of Maryland participated in the CLEF 2000 multilingual task, submitting three official runs that explored the impact of applying language-independent stemming techniques to dictionary-based cross-language information retrieval. The paper begins by describing a cross-language information retrieval architecture based on balanced document translation. A four-stage backoff strategy for improving the coverage of dictionary-based translation techniques is then introduced, and an implementation based on automatically trained statistical stemming is presented. Results indicate that competitive performance can be achieved using four-stage backoff translation in conjunction with freely available bilingual dictionaries, but that the the usefulness of the statistical stemming algorithms that were tried varies considerably across the three languages to which they were applied.

1 Introduction

One important goal of our research is to develop cross-language information retrieval (CLIR) techniques that can be applied to new language pairs with minimal language-specific tuning. So-called "dictionary-based" techniques offer promise in this regard because bilingual dictionaries have proven to be a useful basis for CLIR [6] and because simple bilingual dictionaries are becoming widely available on the Internet. Although bilingual dictionaries sometimes include useful information such as part-of-speech, morphology and translation preference, it is far more common to find a simple list of translation equivalent term pairs—what we refer to as a "bilingual term list." The objective of our participation in the Cross-Language Evaluation Forum (CLEF) was to explore techniques for dictionary-based CLIR using bilingual term lists between English and other European languages. We applied techniques that we have used before (balanced document translation, described below), and chose to focus our contrastive runs on improving translation coverage using morphological analysis and an unsupervised morphological analysis approach that we refer to as "statistical stemming." In the next section we describe our balanced document translation architecture, explain how morphological analysis can be used to improve translation coverage without additional language-specific resources, and introduce two statistical stemming algorithms. The following section presents our CLEF results, which demonstrate that the additional coverage achieved by four-stage backoff translation can have a substantial beneficial effect on retrieval effectiveness as measured by mean average precision, but that our present statistical stemming algorithms perform well only in French. In the final section we draw some conclusions regarding the broader utility of our techniques and suggest some additional research directions.

2 Experiment Design

We chose to participate in the multilingual task of CLEF 2000 because the structure of the task (English queries, documents in other languages) was well matched to a CLIR architecture based on document translation that we have been developing. Document translation is an attractive approach in interactive applications if all queries are in a single language because the pre-translated documents that are retrieved can immediately be examined by the user. Although storage overhead is doubled (if the documents are also stored in their original language), that may be of little consequence in an era of rapidly falling disk prices. The principal challenge in a document translation architecture is to balance the translation speed and translation accuracy. In our initial experiments with document translation, we found that a commercial machine translation system required about 10 machine-months to translate approximately 250,000 documents – resource requirements that would clearly be impractical in many applications [5]. With simpler techniques, such as looking up each word in a bilingual term list, we can translate a similar number of documents in only three machine-hours—a period of time comparable to that required to build an inverted index. In our CLEF experiments we have thus chosen to focus on improving the retrieval effectiveness of dictionary-based CLIR without introducing a significant adverse effect on translation efficiency.

Figure 1 illustrates our overall CLIR system design. Each non-English collection was processed separately using the appropriate bilingual term list. We grouped the articles from *Der Spiegel* and *Frankfurter Rundschau* into a single German collection and formed a French collection from the *Le Monde* articles and an Italian collection from the *La Stampa* articles. The documents were normalized by mapping all characters to lower case 7-bit ASCII through removal of accents. Term-by-term translation was then performed, applying a four-stage backoff statistical stemming approach to enhance translation coverage. For translation, we tokenized source-language terms at white space or terminal punctuation (which had the effect of ignoring all source-language multiword expressions in our bilingual term lists). When no translation was known for a clitic contraction, automatic expansion was performed (e.g. *l'heure* → *la heure*) and the resulting words were translated separately.[1] Other words with no known translation were retained unchanged, which is often appropriate for proper names. We produced exactly two English terms for each source-language term. For terms with no known translation, the untranslated term was generated twice. For terms with one known translation, that translation was generated twice. Terms with two or more known translations resulted in generation of each of the "best" two translations once. In prior experiments we have found that this strategy, known as "balanced translation," outperforms the still fairly common (unbalanced) technique of including all known translations because it avoids overweighting terms that have many translations (which are often quite common, and hence less useful as search terms) [4].

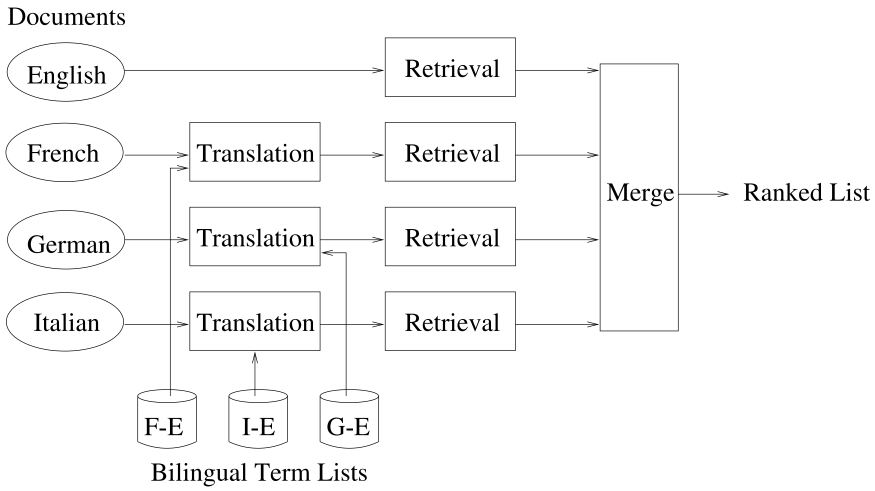

Fig. 1. Production of a merged multilingual ranked list using document translation.

Each of the four resulting English collections (the fourth consisting of *Los Angeles Times* articles, which did not require translation) was then indexed using Inquery (version 3.1p1), with Inquery's kstem stemmer and default English stopword list selected. Queries were produced by enclosing each word in the title, description, and narrative fields (except for stop-structure) in Inquery's #sum operator. In our official runs, two types of stop-structure were removed by hand: "find documents" was removed at the beginning of any description field in which it appeared, and "relevant documents report" was removed at the beginning of any narrative field in which it appeared. Because this stop structure was removed

[1] Clitic contractions are not common in German, so we did not run the splitting process in that case.

manually after examining the queries, our runs are officially classified as being in the "manual" category. We generated separate ranked lists for each collection and then used the weighted round-robin merging technique that we had developed for the TREC CLIR track to construct a single ranked list of the top 1000 documents retrieved for each query [7]. We expected our (monolingual) English system to outperform our French and German systems, and we expected our Italian system to be adversely affected by the small size of the bilingual term list for that language pair. We thus chose a 10:5:5:3 ratio as the relative weights for each language.

We used the same bilingual term lists for CLEF 2000 that we had employed in the TREC-8 CLIR track [7]. Table 1 shows the source and summary statistics for each dictionary. Source language terms in the bilingual term lists were normalized in a manner similar to that used for the documents, although clitic contractions were not split because they were not common in the bilingual term lists. Balanced document translation becomes unwieldy beyond two translations, so the number of translations for any term was limited to the two that most commonly occurred in written English. All single word translations were ordered by decreasing unigram frequency in the Brown corpus (which contains many genres of written English), followed by all multi-word translations (in no particular order), and finally by any single word entries that did not appear at all in the Brown corpus. Translations beyond the second for any English term were then deleted; this had the effect of minimizing the effect of infrequent words in non-standard usages or misspellings that might appear in the bilingual term list.

Pair	Source	English Terms	non-English Terms	Avg Translations
E-G	http://www.quickdic.de	99,357	131,273	1.7
E-F	http://www.freedict.com	20,100	35,008	1.3
E-I	http://www.freedict.com	13,400	17,313	1.3

Table 1. Sources and summary statistics for bilingual dictionaries.

2.1 Four-Stage Backoff Translation

The coverage problem in CLIR arises when the object being translated (in this case, a document), contains a term that is not known to the translation resource (in this case, the bilingual term list). Bilingual term lists found on the web often contain an eclectic mix of root forms and their morphological variants, and our experience with the TREC-8 CLIR track suggested that morphological analysis of terms contained in documents and bilingual term lists could discover plausible translations when no exact match is found. We thus developed a four-stage backoff strategy that was designed to maximize coverage while limiting the introduction of spurious translations:

1. Match the **surface form** of a document term to **surface forms** of source language terms in the bilingual term list.
2. Match the **morphological root** of a document term to **surface forms** of source language terms in the bilingual term list.
3. Match the **surface form** of a document term to **morphological roots** of source language terms in the bilingual term list.
4. Match the **morphological root** of a document term to **morphological roots** of source language terms in the bilingual term list.

The process terminates as soon as a match is found at any stage, and the known translations for that match are generated. Although this process may result in generation of an inappropriate morphological variant for a correct English translation, the use of English stemming in Inquery should minimize the effect of that factor on retrieval effectiveness.

2.2 Statistical Stemming

The four-stage backoff strategy described above poses two key challenges. First, it would require that an efficient morphological analysis system be available for every document language that must be processed. And second, the morphological analysis systems would need to produce accurate results on words presented out of context, as they are in the bilingual term list. This is a tall order, so we elected to explore a simplification of this idea in which morphological analysis was replaced by stemming. Stemmers are freely available for French and German,[2] and stemming has proven to be about as effective as more sophisticated morphology in information retrieval applications where (as is the case in our application) matching is the principal objective [3]. In TREC-3, Buckley, et al. demonstrated that a simple stemmer could be easily constructed for Spanish without knowledge of the language by examining lexicographically similar words to discover common suffixes [1]. We decided to try to push that idea further, automating the process so that it could be applied to new languages without additional effort. We call this approach "statistical stemming," since the stemmer is learned from the statistics of a text collection, in our case the collection that was ultimately to be searched.

Statistical stemming is a special case of unsupervised acquisition of morphology, a specialized topic in computational linguistics. Of this work, the closest in spirit to our objectives that we know of is a program known as Linguistica [2]. Linguistica examines each token in a collection, observing the frequency of stems and suffixes that would result from every possible breakpoint. An optimal breakpoint for each token is then selected by applying as a constraint that every instance of a token must have the same breakpoint and then choosing

[2] French and German stemmers are available as part of the PRISE information retrieval system, which is freely available from the U.S. National Institute of Standards and Technology. Stemmers for a broader collection of languages, including Italian, are also available from the Muscat project at http://open.muscat.com/developer/index.html

breakpoints for each unique token that minimize the number of bits needed to encode the collection. This "minimum description length" criterion captures the intuition that breakpoints should be chosen in such a way that each token is partitioned into a relatively common stem and a relatively common suffix. Linguistica is freely available,[3] but the implementation we used could process only about 200,000 words on a 128 MB Windows NT machine. This is certainly large enough to ensure that breakpoints will be discovered for most common words, but breakpoints might not be discovered for less common terms—quite possibly the terms that would prove most useful in a search. We therefore augmented Linguistica with a simple rule induction technique to handle words that were outside Linguistica's training set.

We implemented rule induction as follows. We first counted the frequency of every one, two, three and four-character suffix that would result in a stem of three or more characters for the first 500,000 words of the collection. Each instance of every word was used to compute the suffix frequencies. These statistics alone would overstate the frequency of partial suffixes—for example, "-ng" is a common ending in English, but in almost every case it is part of "-ing". We thus subtracted the frequency of the most common subsuming suffix of the next longer length from each suffix.[4] The adjusted frequencies were then used to sort all two, three and four-character suffixes in decreasing order of frequency. We observed that the count vs. rank plot for an English training case was convex, so we selected the rank at which the second derivative of the count vs. rank plot was maximized as the limit for how many suffixes to generate for each length. In tuning experiments with English, this approach did not work well for single-character suffixes because the distribution of character frequency (regardless of location) is highly skewed. We thus sorted single characters by the ratio between their word-final likelihood and their unconditioned likelihood, and again used the maximum of the second derivative as a stopping point.[5] For each word, the first matching suffix (if any, from the top of the list) was then removed to produce the stemmed form.

The heuristics we chose were motivated by our intuition of what constituted a likely suffix, but the details were settled only after a good deal of tweaking with a training collection. Of note, the training collection contained only English documents and the tweaking was done by the first author, who has no useful knowledge of French, German or Italian. Table 2.2 shows the suffix removal rules for those languages that were automatically produced with no further tuning. Many of the postulated suffixes in that table accord well with our intuition, as in the case the French adverbial suffix *-ment* or third-person plural inflectional suffix *-ent*. However, some others suggest insufficient generalization. Consider the suggested German suffixes: *-ngen*,*-nden*,*-sen*,*-nen*,*-gen*,*-den*, and *-ten*. The more appropriate suffix would be *-en*; however, the preference for longer subsuming

[3] Linguistica is available at http://humanities.uchicago.edu/faculty/goldsmith/index.html

[4] We did not adjust the frequencies of four-character suffixes since we did not count the five-character suffixes.

[5] If a more precise specification of the process is desired, the source code for the rule induction software is available from the first author.

strings selects the less general suffixes. A large number of single character suffixes are suggested for Italian, including letters such as *-k* and *-w* which do not typically appear in word-final position in this language. This somewhat counterintuitive set suggests that further optimization of threshold setting may be needed.

French	German	Italian
ment	chen	ione
tion	ngen	ente
ique	nden	ioni
ions	sche	ento
ent	rung	enti
res	lich	ato
tes	sten	are
es	ten	to
re	ung	ta
x	den	re
s	gen	ti
	nen	no
	ter	la
	sen	y
	en	o
	er	e
	te	a
	y	k
	t	i
		x
		w

Table 2. Candidate stems, in order of removal.

Three official runs were submitted. In our baseline run ("unstemmed"), we used no pre-translation stemming (i.e., step one alone). In our Linguistica run ("backoff4Ling"), we implemented the complete four-stage backoff strategy using Linguistica for terms with known breakpoints, and added a fifth stage that replicated stage four using the rule induction stemmer in place of Linguistica that would be invoked if none of the first four stages found a translation. The rule induction process was considerably faster than Linguistica (less than 5 minutes, compared with 30-40 minutes for Linguistica) so we also submitted a third run in which which we implemented four-stage backoff with rule induction alone. Table 2.2 summarizes these conditions.

3 Results

Our backoff4 run was judged, and all three runs were scored officially. The top line in Table 4 summarizes the results. Overall, a four-stage backoff document

	unstemmed		backoff4Ling		backoff4	
Stage	Document	Term List	Document	Term List	Document	Term List
1	None	None	None	None	None	None
2			Linguistica	None	Rule Induction	None
3			None	Linguistica	None	Rule Induction
4			Linguistica	Linguistica	Rule Induction	Rule Induction
5			Rule Induction	Rule Induction		

Table 3. Backoff translation steps for the three official runs.

translation strategy using statistical stemming achieved an improvement in retrieval effectiveness over the unstemmed approach that was found to be statistically significant by a paired two-tailed t-test ($p < 0.05$ in both cases). Figure 2 illustrates the advantage of backoff translation on a topic-by-topic basis.

	Unstemmed	Backoff4	Backoff4Ling
Multilingual	0.1798	0.1952	0.1938
English	0.4348	0.4348	0.4348
French	0.1877	0.2823	0.2649
German	0.2421	0.2421	0.2425
Italian	0.2127	0.2045	0.2022

Table 4. Multilingual and language-specific mean uninterpolated average precision, averaged over 40 topics.

Surprisingly, our simple (and quite *ad hoc*) rule induction technique produced results that were statistically indistinguishable from those obtained using the more sophisticated Linguistica system. As Figure 3 shows, Linguistica does better on some topics, but worse on others.

As Figure 4 shows, on balance backoff translation with statistical stemming performed somewhat better than the the median of the submitted CLEF multilingual runs in the automatic category. Since the effect of our limited manual stop-structure removal was likely quite small, we interpret these results as indicating that we have achieved a credible degree of retrieval effectiveness using only freely available linguistic resources.

Although we can conclude from these results that four-stage backoff resulted in improved retrieval effectiveness and that statistical stemming appears to be a viable substitute for more sophisticated morphological analysis in this application, the multilingual task design can easily mask single-language effects. We therefore performed a *post hoc* language-specific analysis by segregating the selected documents and the relevance judgments by language and then scoring each ranked list against the appropriate relevance judgments. The ranked lists that we scored thus contained 130 (for Italian) or 217 (for French and German)

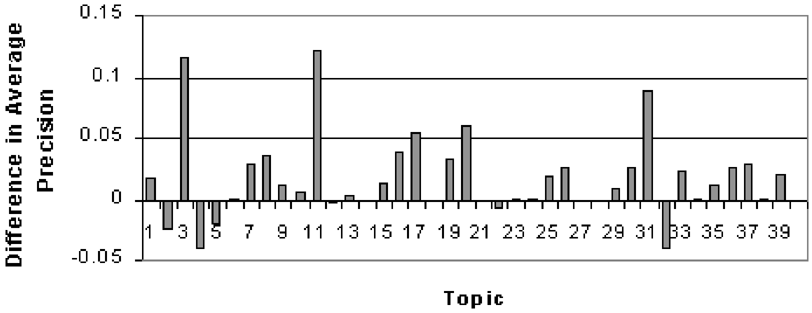

Fig. 2. Improvement (above axis) of Backoff4 over Unstemmed.

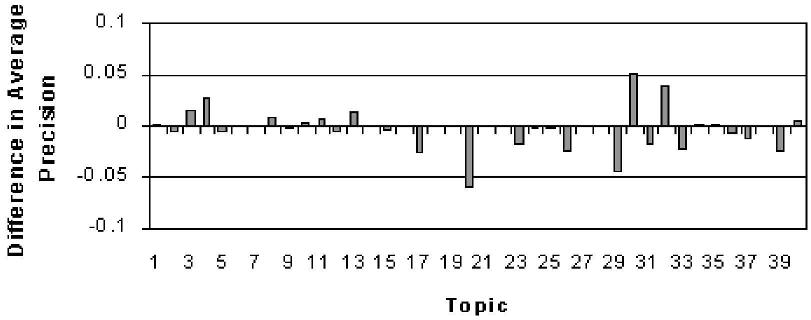

Fig. 3. Improvement (above axis) of Backoff4 over Backoff4Ling.

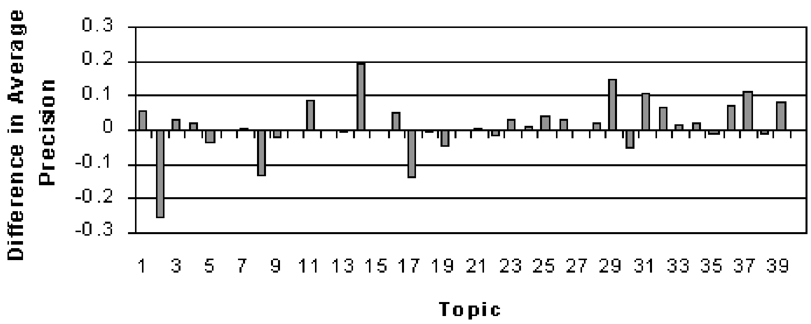

Fig. 4. Comparison of Backoff4 (better above axis) with median CLEF results.

documents from each language. The resulting mean uninterpolated average precision values are shown in the lower portion of Table 4. For French, we found that that both implementations of backoff translation achieved a 55% relative improvement over the unstemmed case, and we found that result to be statistically significant by a paired two-tailed t-test at $p < 0.05$. As Figure 5 illustrates, the magnitude of the improvement varies somewhat across topics, although some of the observed variation may be due to differences in the number of relevant topics for each document.

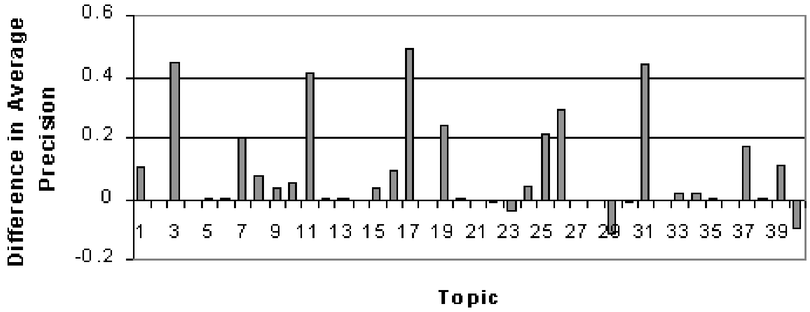

Fig. 5. Improvement (above axis) of backoff translation over unstemmed translation for French documents.

As Table 4 shows, no similar beneficial effect was observed from backoff translation in German or Italian. Many CLEF participants observed that it was important to split German compounds, something that we did not do. Further analysis of the German results may thus not be productive until we have given some thought to how backoff translation and statistical stemming might be integrated with automatic compound splitting. Our disappointing results in Italian might be explained by two possible causes. One possibility is that our statistical stemming techniques are not well suited to some characteristic of Italian. The alternative hypothesis is that our Italian-English bilingual term list (the smallest of the three that we used) may simply be too small.

To explore this issue further, we conducted an additional set of *post hoc* experiments in which we substituted a freely available manually constructed rule-based Italian stemmer from the Muscat project[6] for the Italian rule induction statistical stemmer. We call the new runs "Backoff4Muscat." As Table 5 shows, Backoff4Muscat outperforms Backoff4, although the improvement over Backoff4 is not statistically significant in either the multilingual of the Italian-specific case. The improvement of Backoff4Muscat over Unstemmed was found to be significant at $p < 0.05$ by a two-tailed paired t-test in the multilingual

[6] http://open.muscat.com/developer/index.html

case, but not for Italian alone. We thus conclude that four-stage backoff translation is helpful even with relatively small bilingual term lists, and that the poor performance of Backoff4 for Italian results from a deficiency in the statistical stemmers that we used.

	Unstemmed	Backoff4	Backoff4Muscat
Multilingual	0.1798	0.1952	0.1994
Italian	0.2127	0.2045	0.2338

Table 5. Comparison of backoff translation using statistical stemming and the Muscat Italian stemmer.

4 Conclusion

We have introduced two new techniques, four-stage backoff translation and statistical stemming, and shown how they can be used together to improve retrieval effectiveness in a document translation architecture. Four-stage backoff translation appears to help when using impoverished lexicons that contain a few tens of thousands of terms. Our initial experiments with statistical stemming produced promising results in French, but it is clear from our German and Italian results that more work is required before similar techniques can be reliably applied to a broader range of languages.

Our experiments suggest a number of promising directions for future work. A new version of Linguistica is now available, and trying that is an obvious first step. A detailed analysis of the threshold selection in Italian for our rule induction statistical stemmer is also clearly called for—a more appropriate threshold selection technique might produce far better results with little effort. Our original threshold selection strategy was chosen after inspection of English training data, but an alternative would be to learn a set of language-specific thresholds using test collections from the Text Retrieval Conference's Cross-Language Information Retrieval track. Extending that line of reasoning, the correct answer might not be a single set of thresholds but rather an iterative technique that takes advantage of our ranking of possible stems. We might, for example, first stem using a conservative set of thresholds, and then restem more aggressively if no match is found. Finally, it might be productive to explore the middle ground between our simple rule induction stemmer and the full Linguistica system in order to learn which techniques are particularly helpful in this application. Linguistica provides for fine-grained control over its operation, and exploring a range of possible parameter settings would be a first step in this direction.

When coupled with other language-independent techniques such as blind relevance feedback for query expansion and for post-translation document expansion [4], the techniques that we have explored in this work can potentially provide developers with a robust toolkit with which to design effective dictionary-based CLIR systems using only a bilingual term list and some modest query-language resources (specifically, a comparable collection from which to obtain term statistics). This first CLEF evaluation has proven to be a suitable venue for exploring these questions, and we look forward to continued participation in future years.

Acknowledgments

The authors wish to thank Patrick Schone, Philip Resnik and David Yarowsky for helpful discussions of the unsupervised morphology acquisition, John Goldsmith for making Linguistica available, and Jianqiang Wang for his help with Inquery and the bilingual term lists. This work was work was supported in part by DARPA contract N6600197C8540 and DARPA cooperative agreement N660010028910.

References

1. Buckley, C., Salton, G., Allan, J., Singhal,A.: Automatic query expansion using SMART: TREC 3. In D. K. Harman, editor, *Overview of the Third Text REtrieval Conference (TREC-3)* (1994) 69–80
2. Goldsmith, J.: Unsupervised learning of the morphology of a natural language. http://humanities.uchicago.edu/faculty/goldsmith/ (2000)
3. Hull, D. A.: Stemming algorithms - A case study for detailed evaluation. Journal of the American Society for Information Science, **47**(1) (1996) 70–84
4. Levow, G., Oard, D. W.: Translingual topic tracking with PRISE. In Working Notes of the Third Topic Detection and Tracking Workshop (2000)
5. Oard, D. W.: A comparative study of query and document translation for cross-lan guage information retrieval. In Proceedings of the Third Conference of the Association for Machine Translation in the Americas (1998)
6. Oard, D. W., Diekema, A. R.: Cross-language information retrieval. Annual Review of Information Science and Technology **33** (1998)
7. Oard, D. W., Wang, J., Lin, D., Soboroff, I.: TREC-8 experiments at Maryland: CLIR, QA, and routing. The Eighth Text Retrieval Conference (TREC-8) (1999) http://trec.nist.gov.