Evaluating Lexicon Coverage for Cross-Language Information Retrieval

Gina-Anne Levow	Douglas W. Oard
Inst. for Advanced Computer Studies	Coll. of Library & Information Services
University of Maryland	University of Maryland
College Park, MD USA 20742	College Park, MD USA 20742
gina@umiacs.umd.edu	oard@glue.umd.edu

Abstract

Cross-language retrieval systems use queries expressed in one natural language to retrieve documents that may be written in a different language. Term-by-term translation techniques based on bilingual lexicons are now widely used for cross-language retrieval, but little is known about the way in which retrieval effectiveness varies with dictionary coverage. This paper compares three types of coverage measures using two relatively large Chinese-English dictionaries. An average precision measure calculated using twenty Chinese queries to search the English portion of the Topic Detection and Tracking collection provides the baseline against which the coverage measures are compared. The results indicate that lexicon size is not a suitable coverage measure for this task. In contrast, we find that two other measures, joint "by token" coverage and joint "IDF-weighted, by token coverage" both successfully predict retrieval performance for different (non-merged) lexical resources.

1 Introduction

We live in an increasingly global community, and this fact has motivated rapidly accelerating investment in the development of information systems to help users with a broad range of linguistic skills find information that might be expressed in any of a large number of languages. Cross-Language Information Retrieval (CLIR), in which queries in one language are used to retrieve documents written in another, is a key challenge in this regard. One common approach to CLIR, known as Dictionary-based Query Translation (DQT), is to look up each query term in a simple translation lexicon (generally a bilingual term list), and then replace that term with appropriate terms in the language(s) that documents might be written in (Oard and Diekema, 1998).

Lexicon coverage is an important issue for CLIR systems based on DQT, but little is presently known about how lexicon coverage should be measured. The most obvious statistic, the number of query-language terms in the lexicon, provides no insight into how well matched those terms are with the intended application. Grefenstette suggested that collection-sensitive coverage measures could be more insightful, and illustrated that claim using by-token coverage of query and document collections from the Text Retrieval Conference (Grefenstette, 1998). In this paper we extend that work, introducing a new collection-sensitive coverage measure that reflects the design of modern ranked retrieval systems. We then compare the predictions made using three coverage measures with actual retrieval effectiveness measures obtained using a moderately large bilingual test collection. Our ultimate goal is to discover insightful measures of lexicon coverage for the CLIR task that could be applied to languages for which large (and expensive) information retrieval test collections are not available. Two of new coverage measures - joint "by token" coverage and joint "IDF-weighted, by token" coverage - provide important preliminary strides toward that goal.

2 A Task-Specific Coverage Measure

The basic operation in DQT is the replacement of a query language term with one or more document-language terms. Since this is only possible when the query-language term appears in the lexicon, query-language coverage is clearly essential. A representative query collection could be hard to obtain, so it might be difficult to compute query coverage directly. Large collections of query-language documents are often available in CLIR applications, and it makes sense to use such a collection as a surrogate for a collection of queries if term usage in the documents reasonably reflects the typical use of terms in the domain of discourse that the queries are expected to cover. We have thus chosen to base our new coverage measure in part on term usage statistics from a representative collection of texts in the query language.

A crucial step in any free-text retrieval system is the comparison of query terms with terms in the document. In CLIR systems based on DQT, this comparison cannot occur unless the document-language term appears in the translation lexicon.[1] It is, of

[1]We ignore here the cases such as proper names and loan

course, also necessary that it appear at the right place in the translation lexicon and that there be a sufficient basis for selecting that term when appropriate. Those issues are not easily represented at an aggregate level, however, so we have chosen not to consider them when constructing our coverage measure.

We estimate the probability that *some* translation will be found for a query term by computing the fraction of the query-language term occurrences that match a query-language term in the translation lexicon (i.e., coverage "by token" rather than "by type"). We estimate the probability that an *appropriate* translation will be found by computing the fraction of the document-language term occurrences that match a document-language term in *some entry* in the translation lexicon. These events are clearly not independent (since if the word is missing we would not be surprised to discover that a translation of that term is also missing), but we have chosen to make an independence assumption here because we have not yet developed a useful model of this interaction. We thus calculate the joint probability that a query-language term is present in the lexicon and that the appropriate translation is present in that lexicon entry as the product of the query-language coverage and the document-language coverage.

Modern ranked retrieval systems base the rank order assigned to a document on two factors that are computed for each term: the relative frequency of the term within the document (a measure of the degree to which the term reflects the "aboutness" of the document) and the relative rarity of the term within the collection (a measure of the degree to which the term adds "specificity" to the query). It is the within-document relative frequency factor that motivates our choice of "by token" coverage over "by type" coverage. But the collection-wide relative rarity factor, typically referred to as Inverse Document Frequency" (IDF), introduces an additional consideration. Although the interpretation of IDF is naturally associated with query terms, in monolingual retrieval systems the factor is more commonly applied to document terms for reasons of computational efficiency. The same approach works well in most CLIR systems that use DQT, although we recently found Chinese to be an important exception to that rule of thumb—apparently because segmentation errors propagate through the translation stage in ways that distort the IDF statistics (Oard and Wang, 1999). When IDF is an important factor in the retrieval system design, we can reflect that fact by computing the fraction of the IDF mass that is covered by the translation lexicon rather than the "by token" coverage. The adjustment should be applied to either the query language or the document language, depending on the design of the CLIR system, but not to both. Making this adjustment obviates the need to model the effect of stopword removal, since stopwords are typically so common that they would produce a very low IDF value.

words in which the query and document language terms might be written identically if the character set is the same.

3 Experimental Design

Evaluation of retrieval effectiveness depends on the availability of a suitable test collection that contains representative queries and documents as well as relevance judgments for appropriate (query, document) pairs. The Topic Detection and Tracking evaluation has recently developed a bilingual (Chinese/English) test collection that is well suited to our purpose, so we chose to work with Chinese queries and English documents. We have obtained two Chinese/English lexicons for use with TDT, but some initial analysis by the Linguistic Data Consortium suggested that one of the lexicons might not be as well suited to the task as its size might suggest (Huang, 1999). In this section we describe the design of an experiment to compare four lexical coverage measures for each of the lexical resources described in detail below.

3.1 Coverage Measures

We compute four coverage measures:

1. The number of Chinese terms in the lexicon (Ch Headwords)
2. The number of English terms in the lexicon (En Words)
3. The product of the "by token" coverage measures for each language (Joint BT)
4. The product of the "by token" coverage measure for Chinese and the "IDF mass" coverage measure for English. (Joint IDF-BT)

3.2 Lexical Resources

We sought to compare two Chinese-English bilingual lexicons: a term list provided by the Linguistic Data Consortium (LDC) and a lexicon that we derived automatically from the CETA (Optilex) Chinese-English dictionary. We compared each individually, and also evaluated a merged lexicon that we automatically created from the two resources. The LDC bilingual term list consists of Chinese terms paired with alternate translations for each term into English.[2] The list was compiled from a variety of sources, both internal to the LDC and available from the Internet. The LDC term list also includes some Chinese phrases, since part of the list was produced by automatically inverting an English-Chinese term list.

[2]The LDC term list is available at http://morph.ldc.upenn.edu/Projects/Chinese/

In contrast, the ChiChinese-English Translation Assistance (CETA) file is a manually compiled human-readable lexicon. It contains over 230,000 entries, compiled from 250 dictionaries, some general purpose, others domain-specific or multilingual (Russian-Chinese-French, etc.) as well as primary reference sources such as newspapers and periodicals. We utilize a subset of the entries drawn from contemporary general purpose and economic sources.[3]

For each original lexical resource and the merged lexicon, we convert the entries to a list of Chinese-English translation pairs. We remove any duplicates and delete target language forms that are descriptions of function, such as "question particle" or "exclamation indicating surprise or disgust" rather than actual translations, where automatically identifiable as such. In each case, alternate translations are ranked as follows: first all single word entries are ordered by decreasing target language unigram frequency calculated according to the Brown corpus, followed by all multi-word translations, and finally single word entries with zero unigram Brown corpus frequency. This approach attempts to minimize the damage due to infrequent words in translations (which typically are non-standard usages or misspellings), by ignoring them except when there are no more common alternatives available. We select the highest ranking translation alternative.

3.3 The Topic Detection and Tracking Task

The information retrieval evaluations are conducted in the context of the Topic Detection and Tracking (TDT) evaluation. We consider the part of the CLIR task in this evaluation that involves identification of relevant English documents based on four example documents in Mandarin Chinese. The document collection for retrieval includes two English newswire sources and automatic speech recognition (ASR) transcriptions of six English television and radio news sources. The exemplar query documents are drawn from two Chinese newswire sources and ASR transcriptions of one Mandarin radio broadcast news source.

The TDT task[4] focuses on event-based document retrieval. A seminal event is defined by up to four representative documents, and all related subsequent documents must be retrieved. Exhaustive relevance assessments have been performed by coders at the Linguistic Data Consortium (LDC), classifying each document as relevant/non-relevant for each query topic. The collection includes 19,215 Chinese documents totaling 6,054,556 words and 59,495 English documents totaling 20,245,556 words.

Because queries in TDT are derived from exemplar documents, our assumptions that term distribution in the queries can be well approximated by term distribution in a query language document collection are quite appropriate for this task. We construct a vector of the 180 terms that best distinguish the query exemplars from contemporaneous non-relevant documents by using a χ-squared test in a manner similar to that used by Schütze et al (Schütze et al., 1995). Because the TDT task design requires that all statistics be computed from prior documents, we use a "frozen" set of IDF values that are developed from a similar (but earlier) collection of training documents. The query exemplars for each topic occur at different points in the collection (Jan 1998-June 1998), so the document collections that are searched are right-nested subsets of the evaluation collection. To minimize the effects of different collection size on the comparability of precision and recall measures, we perform paired T-tests on average precision for each query to evaluate performance with different lexicons.

We perform a suite of experiments comparing average precision for:

1. the three bilingual lexicon alternatives, and
2. inclusion and exclusion of single Chinese character tokens.

We report the relationships between the different measures of lexicon coverage we have proposed and the effectiveness of different lexicons and treatments of single characters on CLIR.

4 Results

The results of the information retrieval experiments on the TDT corpus for the 6 lexicon configurations described above are presented in the table below. The results are reported as per-query average precision and overall query-averaged average precision.(Figure 1)

4.1 Overall Findings

For all lexical resources, exclusion of single-character Chinese words showed a significant decrease in retrieval performance relative to queries in which translations of single-character Chinese words were included ($p < 0.05$). Furthermore, overall the CETA/Optilex lexicon systems outperformed those that used the merged lexicon which in turn performed better than using the LDC term list alone. However, none of these source-based differences reached significance. Furthermore, the weaker performance of the merged lexicon relative to the original CETA/Optilex lexicon indicates that the effects of lexicon merging are complex.

[3] The commercial machine-readable version of the CETA dictionary, which we refer to as "Optilex" is available from the MRM corporation, Kensington, MD.

[4] Information about the TDT-3 evaluation may be found at http://www.ldc.upenn.edu/Projects/TDT3/

Topic	Comb Nosc	Comb	Opti Nosc	Opti	LDC Nosc	LDC	Opti-LDC-diff
1	0.1829	0.1873	0.17697	0.1835	0.1805	0.1697	0.0138
2	0.1561	0.1670	0.1498	0.1576	0.1654	0.1887	-0.0310
5	0.1836	0.2482	0.1489	0.2049	0.1886	0.2216	-0.0167
7	0.1172	0.1936	0.1321	0.1957	0.1304	0.1885	0.0073
13	0.1209	0.1294	0.1236	0.1348	0.1142	0.1328	0.0020
15	0.1441	0.1437	0.1412	0.1434	0.1475	0.1490	-0.0056
20	0.2427	0.2351	0.3008	0.3271	0.1631	0.1710	0.1562
23	0.1051	0.1111	0.1105	0.1196	0.1066	0.1186	0.0009
39	0.2293	0.2506	0.2320	0.2360	0.2430	0.2460	-0.0100
44	0.2466	0.2450	0.2655	0.2763	0.2033	0.2201	0.0561
48	0.1376	0.1227	0.1033	0.1032	0.1139	0.1259	-0.0227
57	0.1332	0.1526	0.1578	0.1708	0.1264	0.1821	-0.0113
70	0.1469	0.1547	0.1527	0.1707	0.1535	0.1568	0.0139
71	0.1416	0.1548	0.1358	0.1480	0.1293	0.1425	0.0054
76	0.1812	0.1815	0.1841	0.1802	0.1731	0.1762	0.0040
85	0.1644	0.2449	0.2677	0.2765	0.1999	0.2194	0.0571
88	0.2329	0.2304	0.2148	0.1945	0.2653	0.2697	-0.0751
89	0.1009	0.0979	0.1007	0.1013	0.0975	0.0930	0.0083
91	0.2381	0.2336	0.2539	0.2438	0.3302	0.2975	-0.0537
96	0.1168	0.1247	0.1157	0.1145	0.1220	0.1181	-0.0036
Overall	0.1661	0.1804	0.1734	0.1841	0.1677	0.1794	0.0048

Figure 1: Per-query Average Precision for three lexicon (Comb, Opti, LDC) with and without (Nosc) single characters. Shows improved average precision when retaining single characters.

While we find that only the differences in lexicon related to exclusion of single characters produced statistically significant differences in information retrieval performance on our small query set, we learn several lessons about necessary components for a useful measure of lexicon coverage: inadequacy of entry or headword count as predictor of performance and the importance of source and target language document token coverage. We also are able to begin to quantify our intuitions about the effect of general purpose lexicons on topic-specific information retrieval.

4.2 Lexicon Size vs. Collection-Based Lexicon Coverage

A natural initial measure of dictionary quality would be the size of the dictionary; one might hope that bigger would be better, in terms either of the number of headwords or the total number of entries in the lexicon. The table below lists the number of Chinese headwords, total numbers of Chinese-English pairs, number of headwords excluding single characters, and the number of English words in target translations for each lexicon ("Comb": merged LDC and CETA lexicons, "Opti": CETA/Optilex alone, and "LDC": LDC termlist alone). According to this table (Figure 2), the prediction would be: 1st: Comb, 2nd: LDC, 3rd: Opti. However, our retrieval experiments show that the Opti systems consistently outperform systems using materials translated by LDC. Furthermore the exclusion of single-character Chinese words reduces lexicon size far less than the difference between any of the primary dictionaries, a decrease of 5,000 words in contrast to differences of between 30,000 and 100,000. However, this change results in highly significant decreases in performance.

In contrast, for lexicon coverage, using either "by token" or "idf-weighted" metrics, the prediction is that Opti consistently exceeds LDC for all words, stopped or unstopped English and with and without single characters in the Chinese. The table below illustrates this contrast. Both of these types of coverage metric align with the observed retrieval performance for LDC and Opti.

There is a discrepancy in the relative improvement of LDC over Opti for stemmed cases, which we attribute to more inflected uses in the LDC lexicon, though further analysis is needed. Also, while the metric is effective for the individual lexicons, merged dictionaries clearly interact requiring a further refinement of the metric.

5 Exclusion of Single-Character Chinese Words

5.1 The Tokenization Problem

In deriving our measure of coverage, we note the importance of the first phase of query translation: tokenization, the identification of the individual terms

Lexicon	CH Headwords	CH Entries	CH HW Nosc	ENG words
Comb	195078	341187	188652	97603
Opti	91602	169067	85915	30322
LDC	127924	187130	121756	89003

Figure 2: Size of Chinese-English Lexicon in Chinese Headwords, Total Chinese Entries, Chinese Headwords excluding single characters, and English translation terms for each lexicon.

Lexicon	Ch Words	En Words	Joint BT Words	En stems	Joint BT Stems
"By Token"					
Comb	0.98	0.89	0.87	0.94	0.92
Opti	0.92	0.86	0.79	0.91	0.84
LDC	0.94	0.83	0.78	0.92	0.86
Lexicon	Ch Words	En Words	Joint IDF-BT Words	En stems	Joint IDF-BT Stems
"IDF-Weighted"					
Comb		0.86	0.84	0.92	0.90
Opti		0.81	0.74	0.87	0.80
LDC		0.79	0.74	0.89	0.83

Figure 3: "By Token" and "IDF-Weighted" Coverage Measures, for Chinese Words, English Words, Joint coverage (En * Ch), English Stemmed, and Joint Coverage Stemmed

in the query to be translated. In many Indo-European languages, such as English, this task is fairly simple and can be effectively reduced to little more than treating white space as term boundaries, although the status of multi-word phrases remains an issue. For some Asian languages such as Japanese or Korean, reliable cues are present in the written form, for instance, through morphology. However, written Chinese poses a particular problem at the tokenization level, since it has neither white space nor morphological cues to term segmentation, and correct segmentation is a matter of disagreement even among experts. Mis-segmentation not only fails to identify the correct term, but most often results in oversegmentation, particularly in dictionary-based approaches, since most single characters are valid words. However, these words may be especially problematic since they are highly polysemous, and many of these senses are uncommon. This poor segmentation leads to a cascade of errors that results in particularly poor performance for CLIR applications on Chinese documents.

5.2 Evaluation of Chinese Single-Character Deletion

One proposal for minimizing the impact of these incorrect segmentations is to exclude all single characters as product of missegmentation. We evaluate this alternative both through its impact on our coverage measures and on information retrieval performance. The exclusion of single characters from consideration affects both source and target components of coverage, decreasing the number of terms to be covered and removing those translation alternatives that arise only from single character words.

We find that removal of single characters from translation has a significant and negative impact on both our measure of coverage and on information retrieval performance. This finding holds across all lexicons which we evaluate: LDC, Optilex/CETA, and merged, as demonstrated by paired t-test, two-tailed. (Merged: $t=-2.35$, $p < 0.05$; Optilex/CETA: $t=-2.45$, $p < 0.025$; LDC: $t=-2.51$, $p < 0.025$)

5.3 Discussion

These results demonstrate that although many single-character Chinese word occurrences arise from over-segmentation, there is still significant useful information in these characters. An analogy to this situation may be found in the use of verbs in information retrieval. People focus on nouns in interactive query generation and often omit verbs. However, removing verbs from indexing and retrieval has a significant negative impact on retrieval performance.

6 Conclusion

We have presented three types of measures for evaluating bilingual lexicons in the context of CLIR: lexicon size, joint "by token" coverage

and joint "IDF-weighted, by token" coverage. We have computed these evaluation measures for each of three bilingual lexicons: LDC, CETA/Optilex, and a merged lexicon created from the first two. We then performed a CLIR task using each of these lexicons to perform dictionary-based query translation and computed average precision for each of these runs. We find that the joint "by token" coverage metric and joint "IDF-weighted, by token" coverage metric accurately predict relative performance of the two base lexicons (LDC and CETA/Optilex). In contrast, the lexicon size metric provides inappropriate predictions of performance on this retrieval task.

We use the alternative of single-character word exclusion in Chinese to further demonstrate the limitations of the lexicon size metric, as well as to explore the issues surrounding tokenization. The universal decrease in coverage values from the joint "by token" coverage value to the joint "IDF-weighted, by token" coverage value further provides analytic evidence for the intuition of CLIR researchers that general purpose lexicons fail to cover the highest IDF, and therefore most selective, terms in the collection, such as proper names and locations.

Currently, one can only evaluate a bilingual lexicon for CLIR by directly performing an information retrieval experiment. The evaluation of this experiment relies on the presence of large sets of relevance judgments, which are time-consuming and costly to create. A successful measure of lexicon coverage will act as an accurate proxy for this rare resource. The metrics described in this paper provide an improvement over previously considered lexicon coverage metrics and raise additional questions about evaluation of merged dictionaries and the impact of stemming.

Acknowledgments

The authors would like to thank Ruth Sperer for building the scripts to run the TDT evaluation and NIST for use of the PRISE information retrieval system and Philip Resnik for helping revise this paper. This work has been supported in part by DARPA contract N6600197C8540.

References

Gregory Grefenstette. 1998. Evaluating the adequacy of a multilingual transfer dictionary for the cross language information retrieval. In *First International Conference on Language Resources and Evaluation*, pages 755–758, May.

Shudong Huang. 1999. Evaluation of ldc's bilingual dictionaries. Unpublished manuscript.

Douglas W. Oard and Anne Diekema. 1998. Cross-language information retrieval. In *Annual Review of Information Science and Technology*, volume 33. American Society for Information Science.

Douglas W. Oard and Jianqiang Wang. 1999. Effects of term segmentation on Chinese/English cross-language information retrieval. In *Proceedings of the Symposium on String Processing and Information Retrieval*, September. http://www.glue.umd.edu/~oard/research.html.

Hinrich Schütze, David A. Hull, and Jan O. Pedersen. 1995. A comparison of classifiers and document representations for the routing problem. In Edward A. Fox, Peter Ingwersen, and Raya Fidel, editors, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 229–237, July.