Resolving Translation Ambiguity using Monolingual Corpora
A Report on Clairvoyance CLEF-2002 Experiments

Yan Qu, Greg Grefenstette, David A. Evans
Clairvoyance Corporation
5001 Baum Boulevard, Suite 700
Pittsburgh, PA 15213-1854
{y.qu, g.grefenstette, dae}@clairvoyancecorp.com

Abstract

Choosing the correct target words is a difficult problem for machine translation. In cross-language information retrieval, this problem of choice is mitigated since more than one translation alternative can be retained in the target query. Between choosing just one word as a translation and keeping all the possible translations for each source word, one can apply a range of filtering techniques for eliminating some words and keeping others. In the bilingual track of CLEF 2002, focusing on word translation ambiguity, we experimented with several techniques for choosing the best target translation for each source query word by using co-occurrence statistics in a reference corpus consisting of documents in the target language. One of two distinct corpora was used, the target-language test corpus or the World Wide Web. Our techniques give one best translation per source query word. We also experimented with combining these word choice results (providing up to three translations for each word) in the final translated query. The source query languages were Spanish and Chinese; the target language documents were in English. We submitted four automatic runs for each language pair. When the methods were combined, mixing results obtained with different reference corpora, the recall and average precision of Spanish-to-English retrieval reached 95% and 97%, respectively, of the recall and average precision of an English monolingual retrieval run. For Chinese-to-English text retrieval, the recall and average precision reached 89% and 60%, respectively, of the English run.

1. Introduction

Choosing among lexical or translation variants to find correct target words is a difficult problem for machine translation. In cross-language information retrieval (CLIR), the problem of choosing a single best translation is less critical in theory since more than one translation alternative can frequently be retained in the target query with only minimal harm to retrieval performance. However, there is still a range of possible performance tradeoffs between choosing just one word as a translation and keeping all the possible translations for each source word (Grefenstette 1998). To exploit the possible advantages of limiting the alternative translations to "a few (or one) best", one can apply a range of filtering techniques for eliminating some words and keeping others. In the bilingual track of CLEF 2002, we focused specifically on this problem of word translation ambiguity and experimented with several techniques for choosing the best target translation for each source query word, using co-occurrence statistics over target-language reference corpora. In particular, we derived statistics from two different corpora, the target-language (English) test corpus and the World Wide Web. Our techniques give one "best" translation per source query word per reference corpus. We also experimented with combining these word choice results (providing up to three translations for each word) in the final translated query. The source query languages were Spanish and Chinese; the target language documents were in English. We submitted four automatic runs for each language pair. In this report, we describe our translation disambiguation methods and present their performance results.

2. CLARIT Cross-Language Information Retrieval

For cross-language information retrieval, both the documents and the queries need to be represented in the same language at some point in the process (Oard & Dorr 1996). In our experiments, we adopted the query translation approach. First, the query terms in the source languages were translated into all possible terms in the target language using translation lexicons. Some of these translations were retained according to the methods described below. The retained terms in the target language (English, here) were used for retrieving documents from the target collection. For all document processing—including query and document indexing and retrieval—we used the Clarit system (Evans & Lefferts 1995), in particular, the functions for NLP (morphological analysis and phrase recognition), IR (term weighting and phrasal decomposition), and "thesaurus extraction" (for effecting pseudo-relevance feedback).

2.1 Spanish Topic Processing and Translation

Spanish queries were processed as follows. The text of the Spanish query was tokenized and morphologically analyzed using a Spanish version of Clarit. Only nouns, verbs, adjectives, and adverbs were retained for further treatment. Some additional words were removed via a stop list containing a total of 400 words. This list includes all prepositions, pronouns, and articles (which were already removed using the morphological analyzer); common stop words such as "es", "cada"; and query meta-language from previous CLEF queries such as "describir" and "discutir".

As an example of our processing consider the Spanish query on the Leaning Tower of Pisa (Topic 136): *Torre inclinada de Pisa. ¿En qué estado se encuentra la torre inclinada de Pisa?* After morphological analysis and stop word removal, this query becomes *torre, inclinar, pisa, estado, torre, inclinar, pisa.* Each of these words is then looked up in a Spanish–English word-to-word dictionary, which contains the following translations:

estado	inclinar	pisa	torre
state states statis	apt bow drooping incline inclined inclining sloping stooping titling verging	pisa	high tower towers

We created our Spanish-English gloss lexicon by combining various lexicons available on the Web. The final collation was not manually edited; stop words were automatically removed from the English translations (unless the only translation was a stop word); and the Spanish side of the dictionary was lemmatized (e.g., an original gloss such as "inclinado—apt to" was reduced to "inclinar—apt" in our experimental version). If a source word was not found in the dictionary then the original source word was retained in lieu of a translation. The resulting translations formed the basis of the English queries that were generated by the methods described below.

We can note here two limitations of our technique: (1) The dictionaries are neither clean nor complete. Notice that "leaning" is missing from the above translations of "inclinar". (2) We have restricted ourselves to word-to-word translations for engineering reasons, even though we know that phrasal translation results in superior performance in CLIR (Hull & Grefenstette 1996). If we were to go beyond a research version of our system, investments in reducing these limitations would need to be made.

2.2 Chinese Topic Processing and Translation

Since spaces are not used in Chinese text for word segmentation, we first broke Chinese text into individual words. We used the longest-match method, which greedily recognizes an initial string of characters as a word if the string matches a word in the segmentation dictionary. We obtained a Chinese-to-English wordlist from the Linguistics Data Consortium[1] (LDC). This bilingual wordlist contains a list of Chinese words together with their possible English translations—a total of 188,474 entries. We did not edit or further "clean" the lexicon. Our run-time segmentation dictionary consisted of the Chinese words from this wordlist augmented with all possible single Chinese characters and symbols. By using the words from the bilingual wordlist, we ensured that the words identified during segmentation would have glosses during translation.

Once we obtained the segmented Chinese words, we first removed stop words automatically via a stop word list. The list contains a total of 3,894 entries, including closed-class words (e.g., symbols, prepositions, pronouns, particles), numerals, and query specific terms such as 报导 and 文章, collected from CLEF 2001 topics. Then we translated the remaining words into English using the LDC bilingual wordlist. Lastly, we used the translation disambiguation methods (described below) to select the best translation for a query word.

[1] http://www.ldc.upenn.edu/Projects/Chinese/LDC_ch.htm#e2cdict

As an example of our processing, consider the Chinese version of the Leaning Tower of Pisa query (Topic 136): 比萨斜塔; 比萨斜塔的健康情况如何？ First the query is segmented as:

比萨 ; 斜 ; 塔 ; 比萨 ; 斜 ; 塔 ; 的 ; 健康 ; 情况 ; 如何 ; ？;

After the stop-word removal, the query contains unique words: 比萨 ; 斜 ; 塔 ; 健康 ; 情况. Each of these words is then looked up in the Chinese–English bilingual wordlist, which yields the following translations:

比萨	斜	塔	健康	情况
pisa	askant slanting	ter pagoda tower	hygeia exuberance health healthiness	circumstance circumstantiality situation state affairs

We should note that there are several problems with the longest-match method that can cause segmentation errors. First, the greedy algorithm may break words in the wrong places when word boundaries are ambiguous. Generally, a wrong segmentation will result in more subsequent segmentation errors. Second, the coverage of the dictionary will affect segmentation quality. Missing dictionary words will result in single characters being generated during segmentation. For bilingual retrieval, this not only reduces term quality, but also increases ambiguity, since single characters are generally more ambiguous than multiple character words. This is especially a problem with proper names, where the meanings of the single characters in the names bear no relation to the meaning of the name. In the pre-processing of topics, we eliminated any sequence of more than three consecutive single characters. This helped reduce translation noise, but without proper handling of the names, we still lost the specific information associated with deleted characters, which tended to render the topics too general.

3 Translation Disambiguation Methods

Given that a source language term can give more than one possible target translation, we want to find the best or the best few translations for each non-stop word in the source topic. Our methods for disambiguating alternative possible translations are based on two observations:

- The correct translation for the query term, given its potential translations in a target language, is generally not ambiguous when context (i.e., other terms in the query) is considered.

- The Web and reference corpora can be used as practical resources for estimating the coherence of the translated terms. Each provides a language model of how words co-occur. In particular, we expect that words that are found to co-occur are lexically-semantically cohesive.

For example, suppose the query terms $s_1,\ldots,s_5$ in the source language have the translations in the target language as follows:

s_1	s_2	s_3	s_4	s_5
t_{11} t_{12} t_{13}	t_{21} t_{22}	t_{31} t_{32} t_{33} t_{34}	t_{41}	t_{51} t_{52}

Term s_1 has three possible translations: t_{11}, t_{12}, and t_{13}. A context for t_{11} can be constructed as one of the possible sequences including the other translations in the target language, such as,

$<t_{11}, t_{21}, t_{31}, t_{41}, t_{51}>$
$<t_{11}, t_{21}, t_{32}, t_{41}, t_{51}>$
$<t_{11}, t_{21}, t_{33}, t_{41}, t_{51}>$
$<t_{11}, t_{21}, t_{34}, t_{41}, t_{51}>$
…
$<t_{13}, t_{22}, t_{34}, t_{41}, t_{52}>$

In this example, there are a total of $3 \times 2 \times 4 \times 1 \times 2$ (i.e., 48) possible sequences or paths through the translation space. Each path establishes a context for the translated terms with respect to their neighbors. We assume that the best path of all combinations will demonstrate the best coherence among the translated terms.

We have developed several practical methods to measure the quality (or coherence) of translation paths based on evidence of actual word co-occurrence in reference corpora. One of these methods takes advantage of the World Wide Web (WWW) and two use the actual target test collection, as described in the following sections.

3.1 Web Method

The Web method is an elaboration of the ideas explored by Grefenstette (1999), namely, using the WWW as the language model for choosing translations. The idea of using the Web to acquire general language models is becoming more popular. (See also Zhu & Rosenfeld 2001 for an application for speech.) To exploit the Web, we first create sequences of possible translations. Each sequence is sent to a popular Web portal (here, AltaVista) to discover how often the combination of translations appears. The number of occurrences of a translation is used as the score for the sequence. The complete algorithm is as follows:

```
1  Get translations for each term in the source language query.
2  Construct a hypothesis space of translated sequences (overlapping n-grams, n=3 in our experiments)
   by obtaining all possible combinations of the translations in a source sequence of n query words.
3  For each translated sequence in the hypothesis space,
     3.1  Send it to a Web portal (e.g., AltaVista);
     3.2  Call the number of pages on which that translated sequence occurs the coherence score
          for the query;
     3.3  Select the translation for each source word that has the best coherence score.
4  Collate the selected translations into a new target language query.
```

Since word order is not preserved from one language to another, the query sent to the Web uses an operator that enforces presence of the translated sequence but not sequence order. AltaVista's advanced search supports the operator NEAR, which ensures that the words so linked appear within ten words of each other, in any order. We use this operator to calculate the score of the sequence.

Since the Web searches do not stem search terms, we expanded each translation by linking all surface forms of the search term by OR. For example, if the translated sequence being scored was "big black dogs" then the following, advanced AltaVista query was generated:

> (big OR bigger OR biggest) NEAR (black OR blacks OR blacked OR blacking OR blackest OR blacking) NEAR (dog OR dogs)

The Web method has the advantage of a massive reference corpus: most candidate translations (paths) will result in some "hits" and the number of hits for meaningful combinations of words will typically be much greater than for meaningless ones. It has the potential disadvantage that the texts on the Web may bear little or no direct relation to the texts (or domain) of the target search. In theory, a narrowly focused target search (e.g., in a technical domain reflected by many documents concentrated in the database to be searched) might be under-represented in the Web corpus compared to alternative, more common documents. To mitigate this possibility, we also explored two methods that use only the target texts for co-occurrence evidence.

3.2 Target Corpus Methods

An alternative reference corpus for language modeling, particularly modeling the coherence of combinations of translation alternatives, is the target corpus itself. We use the target corpus as the basis for choosing a "best" translation of a query, exploiting approaches developed by Evans (2000; 2001). We implemented two target-corpus methods ("Corpus1" and "Corpus2").

3.2.1 Corpus1

The Corpus1 method has the following steps:

```
1  Get translations for each term in the query.
2  Construct a hypothesis space of translated queries by obtaining all possible combinations
```

```
   of the translations.
3  For each translated sequence in the hypothesis space,
      3.1  Send it to the target database;
      3.2  Compute the sum of the similarities scores of the top $N$ retrieval documents as the coherence
           score of the sequence.
4  Select the sequence (or sequences) with the best coherence score.
```

The Corpus1 method computes the coherence score for every path in the hypothesis space. This can be computationally expensive when the query terms have many possible translations. In our experiments, we reduced the hypothesis space by using a maximum of three translations for each query term. In cases where there were more than three alternative translations, we chose the three terms with the smallest distribution scores in the target corpus. For summing similarity scores, we set N to 100.

3.2.2 Corpus2

The Corpus2 method makes use of the mutual information of two terms based on corpus statistics. The method works as follows:

```
1  Get translations for each term in the source language query.
2  Construct a hypothesis space of translated sequences (overlapping $n$-grams, $n=3$ in our experiments)
   by obtaining all possible combinations of the translations in a source sequence of $n$ query words.
3  For each translated sequence in the hypothesis space,
      3.1  Compute mutual information (MI) scores for all term pairs in the sequence;
      3.2  Sum up the scores from step 3.1 to give a coherence score for the sequence;
      3.3  Select as a translation of the first source word in the sequence the alternative that gives the best
           coherence score.
4  Collate the selected translations into a new target language query.
```

The mutual information between two term t_1 and t_2 is defined as:

$$MI(t_1, t_2) = \log \frac{p(t_1, t_2)}{p(t_1)\,p(t_2)}$$

3.3 An Illustrative Example

For Topic 136 (as presented in Section 2), we see that there are $3\times10\times1\times3$ or 90 possible ways the Spanish-to-English translations can be combined. For Chinese, the number is $1\times2\times3\times4\times5$ or 120. Here, we give the target translations for the source query terms as determined by the above three translation disambiguation methods. (This reflects a “combined” method in which all the “best” terms of each method are retained for the final translated query.) In cases where the methods produced different “best” translations, we separate each candidate translation by a comma (“,”).

Topic 136

English: *Leaning Tower of Pisa. What is the state of health of the Leaning Tower of Pisa?*
English terms: lean tower; pisa; lean; tower; health; state.

Spanish: *Torre inclinada de Pisa . ¿En qué estado se encuentra la torre inclinada de Pisa?*
English translations of the Spanish query terms: pisa; high, tower; state; droop, tilt, bow.

Chinese: 比萨斜塔; 比萨斜塔的健康情况如何？
English translations of the Chinese query terms: pisa; ter, tower; slant; health; affair, situation; health situation

For our actual submissions reflecting a particular method (e.g., Web), we naturally used only those “best” terms that the method itself nominated. In addition to our official runs, our experiments included a combined method that uses the full set of terms (as given above) for the final target query. We describe the performance results of the combined method along with our official and baseline results in the next section.

4 Experiments

Our CLIR experiment labels have the following convention: Cl<source-language>2<target-language><method>. Thus, "Cles2enw" denotes our Spanish-to-English Web-method run. The actual experiments involved separate runs for each of the three translation disambiguation methods described in Section 3 (the "w", "t1", and "t2" runs), and also runs with combined methods ("c1", based on a combination of the Web and Corpus1 methods, and "c2", based on all three methods). In all cases, combinations involved a simple concatenation of the selected translations nominated by each participating method. To establish a baseline for evaluating the quality of translation disambiguation, we used all possible translations in a default run ("all"). We ran English monolingual experiments to obtain the baseline with ideal translations.

All the experiments were run with post-translation pseudo-relevance feedback, as we have observed that post-translation pseudo-relevance feedback produces the best overall performance boost (Qu et al. 2000). The feedback-related parameters were based on calibration runs using CLEF 2001 topics. The settings for Spanish-to-English retrieval were: extracting T=50 terms from the top N=25 retrieved documents, with an additional term cutoff percentage set to P=0.1. For Chinese-to-English retrieval: T=50, N=25, P=0.01. For English monolingual retrieval: T=75, N=50, P=0.8. We used a variation of Rocchio weighting to identify terms for selection.

All runs were automatic. All the queries used the title and description fields (Ttitle+Description) of the topics provided by CLEF 2002. The results presented below are based on relevance judgments of 42 topics. The topics not evaluated include 93, 96, 101, 110, 117, 118, 127, and 132, as these were not listed among the official results (presumably because they have no relevant documents in the target corpus).

Table 1 and Table 2 give the results for our submitted runs, together with our other experimental runs for comparative analysis. For Spanish-to-English cross-language retrieval, compared with the baseline of keeping all possible translations, both the Web-based and corpus-based methods improve the average precision by 2.5% to 14.4%. The Web method and the Corpus2 method improve the exact precision by 17.2% and 5%, respectively. Overall recall decreases for the corpus-based methods, while it improves a little (0.6%) for the Web run. By combining the methods, the overall recall, average precision, and exaction precision all improve over the baseline. The combination of all three disambiguation methods produces the best average precision and exact precision, achieving 97.3% and 97.9% of the average precision and exact precision of the English monolingual retrieval results. The best recall is achieved by combining the Web method and the Corpus1 method, reaching 95.7% of that of English monolingual retrieval.

Run ID	Method	Recall (over baseline)	AP (over baseline)	EP (over baseline)
Cles2enw	**Web**	**720/821 (+0.6%)**	**0.3502 (+14.4%)**	**0.3399 (+17.2%)**
Cles2ent1	**Corpus1**	**664/821 (-7.3%)**	**0.3137 (+2.5%)**	**0.2871 (-1.0%)**
Cles2ent2	**Corpus2**	**706/821 (-1.4%)**	**0.3310 (+8.2%)**	**0.3046 (+5.0%)**
Cles2enc1	**Web, Corpus1**	**750/821 (+4.8%)**	**0.3478 (+13.7%)**	**0.3276 (+12.9%)**
Cles2enc2	Web, Corpus1, Corpus2	744/821 (+3.9%)	0.3583 (+17.1%)	0.3441 (+18.6%)
Cles2enall (baseline)	All possible translations	716/821	0.3060	0.2901
English	*original English topics*	*784/821*	*0.3682*	*0.3514*

Table 1: Spanish-to-English retrieval performance with post-translation pseudo-relevance feedback. The runs in boldface are our submitted runs.

For Chinese-to-English cross-language retrieval, compared with the bilingual baseline of keeping all possible translations, both the Web-based method and the Corpus2 method improve average precision and exact precision, while only the Corpus2 method improves recall. The Corpus1 method did not perform well compared with the baseline. Again, when the methods are combined, we observe improvements over the overall recall, average precision, and exact precision. The best run, with all three translation disambiguation methods combined, reached 89.0%, 59.7%, and 62.9% of the recall, average precision, and exact precision, respectively, of the English monolingual run.

Run ID	Method	Recall (over baseline)	AP (over baseline)	EP (over baseline)
Clch2enw	**Web**[2]	**591/821 (-9.2%)**	**0.1795 (+2.8%)**	**0.1752 (+3.2%)**
Clch2ent1	**Corpus1**	**558/821 (-14.3%)**	**0.1322 (-24.3%)**	**0.1262 (-25.6%)**
Clch2ent2	**Corpus2**	**655/821 (+0.6%)**	**0.1936 (+10.9%)**	**0.1853 (+9.2%)**
Clch2enc1	**Web, Corpus1**[3]	**653/821 (+0.3%)**	**0.1858 (+6.4%)**	**0.1841 (+8.5%)**
Clch2enc2	Web, Corpus1, Corpus2	698/821 (+7.2%)	0.2199 (+25.9%)	0.2209 (+30.2%)
Clch2enall (baseline)	All possible translations	651/821	0.1746	0.1697
English	*original English topics*	*784/821*	*0.3682*	*0.3514*

Table 2: Chinese-to-English retrieval performance with post-translation pseudo-relevance feedback. The runs in boldface are our submitted runs.

Since pseudo-relevance feedback can affect performance differently depending on the original query terms, in our follow-up experiments, we re-computed the results without using pseudo-relevance feedback to best estimate the quality of the selected translations against the baseline. Table 3 and Table 4 give the results without feedback for both language pairs.

Run ID	Method	Recall (over baseline)	AP (over baseline)	EP (over baseline)
Cles2enw-nf	Web	684/821 (+1.3%)	0.2940 (+27.5%)	0.2776 (+19.2%)
Cles2ent1-nf	Corpus1	620/821 (-8.2%)	0.2608 (+13.1%)	0.2417 (+3.8%)
Cles2ent2-nf	Corpus2	679/821 (+0.6%)	0.3035 (+31.6%)	0.3087 (+32.6%)
Cles2enc1-nf	Web, Corpus1	702/821 (+4.0%)	0.3110 (+34.9%)	0.2948 (+26.6%)
Cles2enc2-nf	Web, Corpus1, Corpus2	695/821 (+3.0%)	0.3079 (+33.5%)	0.2955 (+26.9%)
Cles2enall-nf (baseline)	All possible translations	675/821	0.2306	0.2328
English-nf	*original English topics*	*770/821*	*0.3331*	*0.3156*

Table 3: Spanish-to-English retrieval performance without pseudo-relevance feedback.

Run ID	Method	Recall (over baseline)	AP (over baseline)	EP (over baseline)
Clch2enw-nf	Web	521/821 (-11.5%)	0.1547 (+20.8%)	0.1630 (+27.5%)
Clch2ent1-nf	Corpus1	534/821 (-9.3%)	0.1094 (-14.6%)	0.1026 (-19.7%)
Clch2ent2-nf	Corpus2	575/821 (-2.4%)	0.1510 (+17.9%)	0.1460 (+14.2%)
Clch2enc1-nf	Web, Corpus1	588/821 (-0.2%)	0.1556 (+21.5%)	0.1608 (+25.8%)
Clch2enc2-nf	Web, Corpus1, Corpus2	613/821 (+4.1%)	0.1761 (+37.5%)	0.1858 (+45.4%)
Clch2enall-nf (baseline)	All possible translations	589/821	0.1281	0.1278
English-nf	*original English topics*	*770/821*	*0.3331*	*0.3156*

Table 4: Chinese-to-English retrieval performance without pseudo-relevance feedback.

Regardless of whether pseudo-relevance feedback is used or not, the Chinese–English retrieval is overall poor. Beside translation ambiguity, the bilingual translation lexicon is very noisy, with many wrong word choices, occasional misspellings, and interfering descriptive text. In addition, wrong word segmentation resulted from incomplete segmentation dictionary coverage and the greedy longest match segmentation algorithm.

For Spanish–English cross-language retrieval, all three translation disambiguation methods outperform the baseline in terms of recall, average precision, and exact precision (except recall with the Corpus1 method). The

[2] The statistics reported here are higher than the official evaluation statistics. In the official submission, we did not filter out the consecutive single Chinese characters (>3) for this run.

[3] The statistics reported here are higher than the official evaluation statistics since we fixed a formatting bug in the official submission.

combinations of the methods outperform any individual method participating in the combinations. For Chinese–English cross-language retrieval, both the Web method and the Corpus2 method improve average precision and exact precision, while the Corpus1 method performs less well with these measures compared with the baseline. All three methods resulted in a decrease in recall performance. The combinations generally outperform any individual method participating in the combinations, except for recall of the Clch2enc1-nf run.

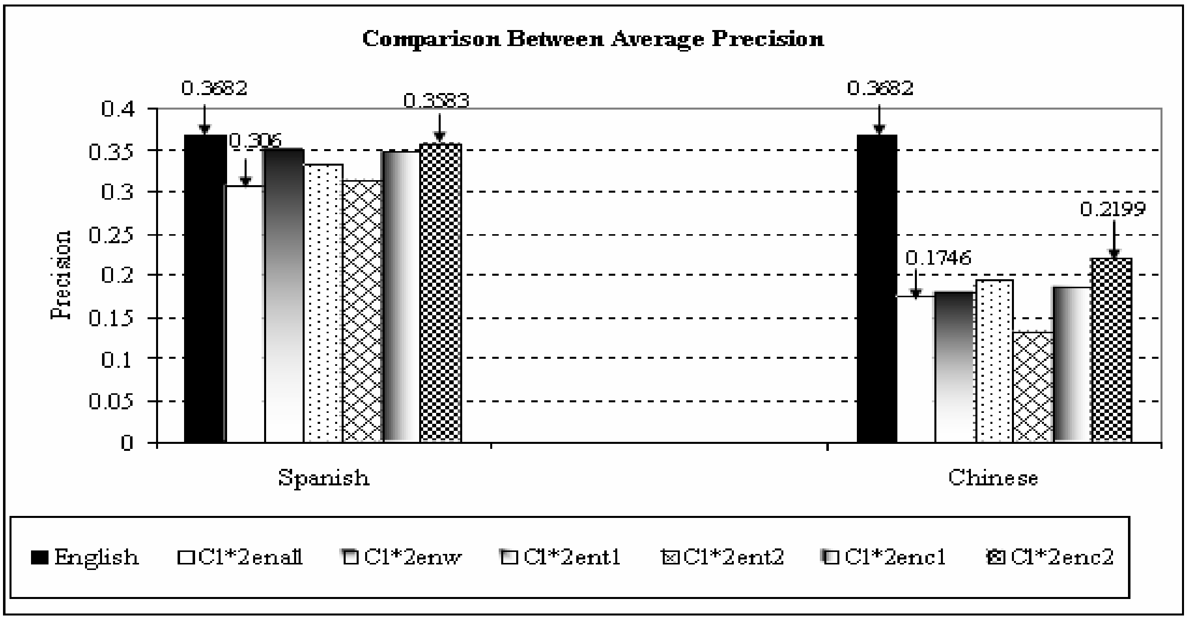

Figure 1: Comparative analysis of average precision, with the English run as the monolingual retrieval baseline and the Cl*2enall runs as the cross-language retrieval baseline. The best results are achieved by the combination of all three translation disambiguation methods, Cl*2enc2.

A summary of our results focusing on average precision is given schematically in Figure 1. In general, our best results for Spanish-to-English CLIR are virtually indistinguishable from English monolingual performance. Our best results for Chinese-to-English CLIR, suffering from the effects of poor resources, are at about 60% of the monolingual baseline.

5 Error Analysis

Comparing the English baseline run to the Spanish-to-English Web run (Cles2enw), we find that 16 of the translated Spanish-to-English queries actually give better results in terms of average precision than the English queries, 26 are worse (of which 12 are much worse, with less than half the average precision of the English queries). (See Figure 2.)

Some of the reasons for improved results are:

- Different word choice, e.g.:
    - *Population* in the English version of Topic 95 (Conflict in Palestine) becomes *town* in the Spanish-to-English translation of *poblacion*.
    - *Ski races* in the English version of Topic 102 becomes *ski competition* in the translations.
    - The English version of Topic 106 (European car industry) contains *countermeasures*, whereas the Spanish contains *medidas de recuperacion*, which is translated by the methods described above as *recovery measures*, which are more common words than *countermeasures*, which is usually found in political rather than business contexts in the CLEF documents.
    - In Topic 140 (Mobile phones), the English version contains *perspectives* and the Spanish version contains *pesrpectiva*, which can be translated as *outlook*, *perspective*, *prospect*, and *vista* via our dictionary, and our method chooses *outlook*, a more common word, which might well account for why this topic scores better after translation.

- Different word ordering: Clarit recognizes noun phrases in the English text; in the Spanish-to-English text, a query is repeated backwards and forwards so that different phrases may be recognized in the reconstituted topic, e.g.:
    - *Weapon destruction* appears in the translation of Topic 119, whereas *weapon* and *destruction* are not found in the same simple noun phrase in the English topic.
    - *Grunge rock* appears in the translation of Topic 130 while the English contains *grunge group.*

- Different formulations of the same topic in English and Spanish, e.g.:
    - Topic 121 (Successes of Ayrton Senna) contains *success* and *sporting achievements* while the Spanish version contains just the word *palmares*, which is not in the Spanish-English dictionary and not translated in the resulting English version of the Spanish topic. It could be that the English words *successes* and *achievements* are only distractors from the relevant documents for Ayrton Senna.
    - Topic 138 (Foreign words in French) contains *lengua*, which translates to *language* not present in the English version.

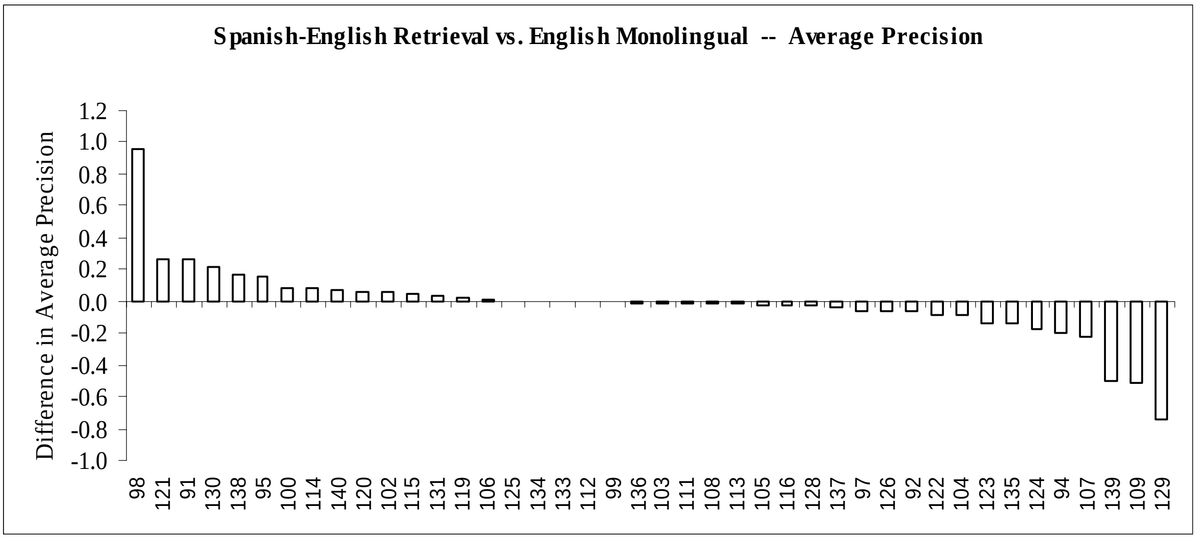

Figure 2: Query-by-query comparison of precision between the best Spanish–English retrieval run (Cles2enc2) and English monolingual retrieval.

Some of the reasons for worse results are:

- Proper names written differently and not in the dictionary, e.g.:
    - *Solzhenitsyn* (Topic 94) is written *Solzhenitsin* in the Spanish topic and is retained with the same spelling in the translated queries since it is not found in the dictionary.
    - *European Cup* (Topic 113) is written *Eurocopa* in Spanish and is not in the dictionary, so it passes through as-is, but is not found as a string in the English documents.

- Dictionary divergences, e.g.:
    - In Topic 107 (Genetic Engineering), *food chain* appears as *cadena alimentaria* in Spanish, but *alimentaria* does not have *food* among its translations.

Comparing the English baseline run to the Chinese-to-English runs, we observe that most of the translated queries are worse in average precision compared with that of the English run. As in Spanish-to-English retrieval, the combination of the three methods (Clch2enc2) produces the best overall performance: 12 of the translated queries give better average precision than the English queries and 30 are worse (Figure 3).

The better performance is due to:

- Reduced ambiguity in translation, e.g.:

  - In Topic 113 (European Cup), the English version uses the word *football*, while the target translations include the word *soccer*. Since *football* can mean *soccer* or *American football*, the translation makes the query more relevant to the topic.

- Difference in word choice, e.g.:

  - In Topic 133 (German Armed Forces Out-of-Area), the English version uses the word *area*, while the Chinese version uses *border*. As our system throws out *out* and *of* as stop words, the word *area* becomes too general, while *border* implies country boundaries.

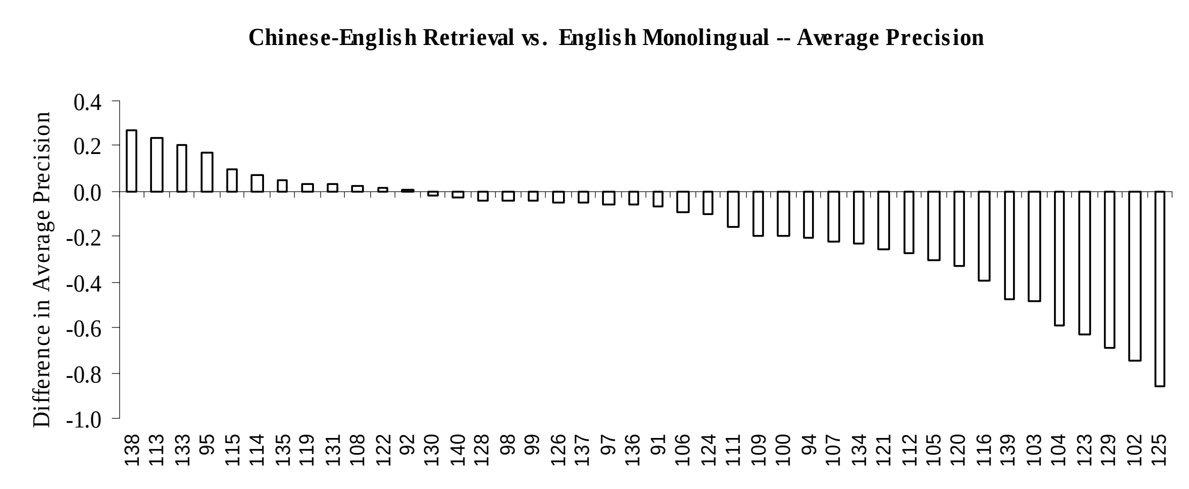

Figure 3: Query-by-query comparison of precision between the best Chinese–English retrieval run (Clch2enc2) and English monolingual retrieval.

Some of the reasons for the poor performance are:

- Improper segmentation of words, e.g.:

  - In Topic 111 (电脑动画), the word 动画 was segmented as two characters 动 (with translations of *act*, *arouse*, *get moving, move*, *stir*, *act*, *change*, *use*, *touch*, etc.) and 画 (with translations of *draw*, *painting*, *picture*). As a result, our system produced translations such as *use picture* instead of the correct translation *animation*.
  - Similar segmentation mistakes include: 银河 (galaxy) in Topic 129 as 银 (*silver*) and 河 (*river*); 选美 (beauty contest) in Topic 137 as 选 (*to choose, to elect, to pick, to select*) and 美 (*America, beautiful, pretty*).
  - 欧联渔获 (EU fishing) in Topic 139 was rendered as four single characters 欧 , 联 , 渔 , 获 , which were consequently filtered out by the system.

- Improper segmentation of transliterated names, e.g.:

  - Topic 102 (阿尔伯特·托姆巴的胜利) is segmented as:

    阿 ; 尔 ; 伯 ; 特 ; · ; 托 ; 姆 ; 巴 ; 的 ; 胜利 ;
    寻找 ; 有关 ; 阿 ; 尔 ; 伯 ; 特 ; 托 ; 姆 ; 巴 ; 获得 ; 滑雪 ; 竞赛的 ; 报导 ; 。 ;

    The name 阿尔伯特·托姆巴 has been incorrectly segmented into seven single characters that have no direct semantic relation to the transliterated name. During pre-processing of the topics, we

filtered out the consecutive single characters. As a result, the query contained only general terms such as 胜利 (victory), 滑雪 (ski), and 竞赛的 (competition). The best precision was achieved by the Web run at 0.1078. In contrast, in the English query (Victories of Alberto Tomba), the name, “Alberto Tomba”, as a term makes the query very specific, resulting in high precision (0.7491).
  - Other topics that suffered similarly include: Topic 94 (Return of Solzhenitsy), 98 (Films by the Kaurismäkis), 103 (Conflict of interests in Italy), 104 (Super G Gold medal), 120 (Edouard Balladur), 121 (Successes of Ayrton Senna), and 123 (Marriage Jackson-Presley).

- Different word choice, e.g.:

  - The translations of the word 遗传 in Topic 107 (遗传工程) are *transmit*, *transmittal*, and *hereditary*, instead of the desired translation *genetic*.
  - In Topic 99, Holocaust is represented in Chinese as 屠杀 (butchery, massacre).

6 Conclusions

We have explored three methods for selecting “best” target translations that take advantage of co-occurrence statistics among alternative translations of the source query words. We have demonstrated that, given a wide variety of possible translations that might be generated from a bilingual dictionary, the use of the Web or a local large corpus as a language model can provide a good basis for lexical choice, provided the gloss dictionary covers the source vocabulary. Combining the target words obtained from the different translation disambiguation methods can produce better cross-language retrieval performance, as compared to keeping all possible translations. For Spanish-to-English retrieval, our combined method achieved 95% and 97%, respectively, of the recall and average precision of our English monolingual run. For Chinese-to-English text retrieval, the recall and average precision reached 89% and 60%, respectively, of the English monolingual results.

Our experiments have shown that the quality of the translation resources has a significant impact on the performance of cross-language retrieval. The poor translation quality and poor coverage of the Chinese-to-English bilingual lexicon we used resulted in relatively poor performance of the Chinese-to-English retrieval. Names (transliterated names in CLEF topics, in particular) that are not covered in a translation lexicon need to be recognized and translated correctly for better retrieval performance.

With respect to the translation disambiguation methods, we believe our methods can be further improved by addressing the following issues:

- Identifying the optimal context span for the Web method and the Corpus1 method. Currently, we use a 3-word context, which may limit the access of contextual information.

- Exploring ways to prune paths in the translation hypothesis space when the space is large. This includes the problem of how to rank translations of a query term when many translations are possible (currently, for the Corpus1 method, we use the idf scores of the translations from the target corpus) and how to rank the combinations of the translations of a sequence of source terms.

- Identifying the optimal retrieval cutoff point for the Corpus1 method, instead of the arbitrary number of responses (i.e., 100) that we used in the experiments.

- Incorporating phrasal translations. Phrasal translations can give important performance gain in cross-language retrieval. We have begun to explore techniques for automatically translating phrasal terms.

In general, we feel our results demonstrate that it is possible to achieve remarkably high CLIR performance by exploiting relatively simple and available resources. Such approaches hold promise for cross-language retrieval in cases where machine translation, parallel corpora, or other knowledge resources may be difficult or impossible to obtain. We believe that incremental straightforward refinements of our approach will give better and more consistent results. We plan on testing this hypothesis in future work on other language pairs.

References

[Evans 2000] Evans, D.A. Method and Apparatus for Cross Linguistic Document Retrieval. U.S. Patent # 6,055,528, April 25, 2000.

[Evans 2001] Evans, D.A. Method and Apparatus for Cross Linguistic Database Retrieval. U.S. Patent # 6,263,329 (a division of U.S. Patent # 6,055,528), July 17, 2001.

[Evans & Lefferts 1995] Evans, D.A., and R.G. Lefferts. CLARIT–TREC Experiments. *Information Processing and Management*, Vol.31, No.3, 1995, 385–395.

[Grefenstette 1998]. Grefenstette, G. The Problem of Cross-Language Information Retrieval. In G. Grefenstette, editor, *Cross-Language Information Retrieval*, chapter 1. Kluwer Academic Publishers, Boston, pp.1-9, 1998.

[Grefenstette 1999] Grefenstette, G. The WWW as a resource for example-based MT tasks. In *Proc., ASLIB Translating and the Computer 21 Conference*, London, 1999.

[Hull & Grefenstette 1996] Hull, D.A. and G. Grefenstette. Experiments in Multilingual Information Retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Switzerland, 1996.

[Oard & Dorr 1996] Oard, D.W. and B.J. Dorr. *A survey of multilingual text retrieval*. Technical Report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies, 1996.

[Qu et al. 2000] Qu, Y., A.N. Eilerman, H. Jin, and D.A. Evans. The Effect of Pseudo-relevance Feedback on MT-Based CLIR. In *Proceedings of the Recherche d'Informations Assistée par Ordinateur (RIAO 2000),* 2000.

[Zhu & Rosenfeld 2001] Zhu, X. and R. Rosenfeld. Improving Trigram Language Modeling with the World Wide Web. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing* (ICASSP), 2001.