TREC-9 CLIR at CUHK
Disambiguation by Similarity Values Between Adjacent Words

Honglan Jin Kam-Fai Wong

Systems Engineering and Engineering Management Department
The Chinese University of Hong Kong
{hljin, kfwong@se.cuhk.edu.hk}

Abstract

We investigated the dictionary-based query translation method combining the translation disambiguation process using statistic co-occurrence information trained from the provided corpus.

We believe that neighboring words tend to be related in contextual meaning and have higher chance of co-occurrence particularly if adjacent words (two or more) compose a phrase. The correct translation equivalents of co-occurrence pattern in a source language are more likely to co-occur in a target language documents than in conjunction with any incorrect translation equivalents within a certain range of contextual window size.

In this work, we tested several methods to calculate the degree of co-occurrence and used them as the basis of disambiguation. Different from most disambiguation methods which usually select one best translation equivalent for a word, we select the best translation equivalent pairs for two adjacent words. The final translated queries are the concatenation of all overlapped adjacent word translation pairs after disambiguation.

System Description

The well-known vector space modeled SMART information retrieval system, Version 11, is used as our platform. We adopted the weighting strategy for documents and queries as *Lnu.Ltu* [1,2], which has been proved more successful than cosine normalization.

The queries were produced after query translation and ambiguity resolution. We fed them to the SMART system to get the retrieval result.

Query Translation

Bilingual Dictionaries

A bilingual English to Chinese machine readable dictionary (MRD) produced by Earth Village (http://www.samlight.com/ev/eng/) is used as our translation resource. This MRD has many entries exactly the same with those in the bilingual dictionary edited by LDC (http://morph.ldc.upenn.edu/Projects/Chinese). The reason we chose Earth Village was that Earth Village provided POS (parts of speech) information. We thought it was useful at the early stage of our experiments but ended up not using it. For phrase translation purpose, we combined three sources: a Chinese to English MRD (http://www.mindspring.com/~paul_denisowski/cedict.html), Earth Village, and the one from LDC. The Chinese to English MRD was converted to English->Chinese and we extracted all the phrase translations from the above resources and complied a single phrase level English Chinese MRD. From our previous experiments, we found the more the number of translations for a word, the higher the chance of introducing extraneous translations for this word. For this reason, we used only Earth Village MRD as our word level translation resource. There are 61729 entries, 2.3 translations in average for each word. However, for the 25 queries in TREC-9, each word in the source language (English) has 5 translations in Chinese after the translation by Earth Village MRD.

Phrase level translation was performed before word level translation. All English words were morphologically transformed to its original word root by using WordNet (http://www.wordnet.com). The root was used as the key to search for its corresponding translations in the dictionary. To perform phrase level translation, we created from bigrams to five-grams composed by adjacent words first in the queries. If a higher gram translation failed, a lower one would be tried until bigram was reached. If it still failed, word level translation was adopted. Otherwise, phrase level translation was performed and the same procedure starting from the next word position was repeated.

Chinese Segmentation

The corpus and the translated queries were segmented by using the perl coded software developed by Erik Peterson (http://www.mandarintools.com). But we replaced the original word list dictionary with our own, a word list of Hong Kong style words.

Post processing after translation

After the initial translation, we did some pruning based on our previous experience and some ad hoc rules. Earth Village is basically a mainland Chinese language style dictionary while the corpus used is in Hong Kong style Chinese. For the same concept, two styles may have totally different representations in the bilingual dictionary. For the translated segmented queries (mainland style), we did the following pruning:

1. Delete the translations having longer than five Chinese characters unless there's only one translation: If a translation is too long (exceed five characters for example), this translation is highly likely the description of the word meaning instead of direct translation of the word.
2. Delete the translated entries being segmented unless there's only one translation. If a translated word is segmented, very probably it is because (1) there's no entry in the dictionary for the word segmentation, (2) it has different translations in China and Hong Kong.
3. Keep only the first three translations with the highest term frequency (*TF*) in the corpus. From our previous experiments, translations with high term frequency in the target language tend to have higher chance of being the correct translations than rare appearing ones.

After the above processing, each word has no more than three translation candidates.

Disambiguation

There are several scenarios of resolving translation ambiguity by using co-occurrence (CO) information.

First, a NLP parser can be used to recognize all the grammatical sub-components such as a phrase. Then the CO information is used to calculate the coherent values in the target language among the composite words within a phrase. The translation for this phrase is the one that has the highest coherent values among all the translation combinations for the phrase. However, a parser is not always reliable. Further, individual words which are not associated to any phrases are isolated in meaning; we can do nothing to resolve their translation ambiguity.

Second, ambiguity is resolved in sentence level rather than phrase level like method one. We create all the translation combinations in the target language for a sentence and choose the one that has the highest coherent values as the final translation. Obviously, as a sentence is usually much longer than a phrase, the number of translation combinations in this method is much larger and thus the computation cost can be too high. Another problem with this method is that when the corpus is not large enough, the coherent values trained from it may be misleading. The longer a sentence is, the more costly is the computation and the larger the corpus is required. The rate of increase of both computation cost and size of the corpus required is exponential.

Third, the disambiguation is done between two adjacent words. Among all the translation combinations between two words, we choose the pair with the highest coherent values as the final translation. The cost is low and the corpus size requirement is much less restricted. We adopted this method for its easy computation and the corpus condition.

Co-occurrence information such as mutual information (MI) [3] was used to calculate the degree of cohesion between two words. MI measure however strongly favors rarely appearing words. We apply the methods to calculate the similarity values between all adjacent word pairs in queries to reduce the translation errors.

If two words always co-occur within a particular contextual range such as adjacent positions, a sentence or even a whole document, they should have similar distribution pattern within that contextual range throughout the document collection. Higher similar distribution means higher degree of co-occurrence pattern or coherent values. The correct translation equivalents of co-occurrence pattern in source language is more likely to co-occur in the target language documents than in conjunction with any incorrect translation equivalents within a certain range of contextual window size.

We calculate this degree of similarity as the inner product of two vectors each representing a word distribution in the collection. For disambiguation purpose, a fine-grained context for a co-occurrence scope is essential. We chose the window size to be a sentence in target language. The dimension of the vectors are the number of windows (or the number of sentences in the collection). The value of each dimension is 1 if a word appears in that sentence and 0 otherwise. We made two assumptions here: a word always appears no more than once in a sentence and the variation of sentence length can be ignored. By considering only the distribution throughout the corpus as the normalization factor, we assigns *idf* value to each dimension of a vector of a word as the weight, i.e.,

$$idf = \log(N / N_c)$$

where N is the total number of documents in the corpus and Nc is the number of documents where the word appears. The similarity of two words by their inner product is the sum of

$$tf(ab) * idf(a) * idf(b)$$

in each dimension where $tf(ab)$ is the co-occurrence indicator (1 or 0) in sentence scope and $idf(a)$, $idf(b)$ are the *idf* values for words a and b respectively.

We calculate similarity values for all possible pairs of translation between two adjacent non-stopword words in queries and select the translated pairs with the highest similarity value in the target language as the final translations.

The final translated queries are the concatenation of all overlapped adjacent word translation pairs after disambiguation.

Our method is different from others in that we did not select the best translation candidate for a word. We select the best translation pairs instead. By considering all overlapping pairs, each word in fact has two translations (except the first and the last words in a sentence). But if a translation has strong similarity value with the translation of the word adjacently before and after it, two translations should be the same.

There are several features for this arrangement: First, no grammatical boundary such as phrase boundary recognition is needed during disambiguation. Second, even if two adjacent words are not a phrase, many of them are related in contextual meaning and have a higher chance of co-occurrence. Overlapped concatenation makes each word's translation be selected twice. If two translations are the same, such a word appearing in queries in the target language would have higher weighting than a word having two different translations when the *TF* value is considered in query weighting. We believe the former case would produce more correct translations. If this is the case, more correct translations would enforce higher weighting values, which would help the retrieval performance.

Experiments

We submitted two official runs. One was monolingual and the other cross-lingual. However, we will describe more runs here to support our analysis.

We used all three parts of a query: title, description and narratives. All our queries are long queries. We used SMART *Lnu.Ltu* weighting and SMART Rocchio query expansion (mono run) before and after query translation (xlingual run). Three parameters of expansion were set to alpha=8, beta=16, gamma=8. For monolingual query expansion, we added 35 terms extracted from the top 10 documents. From our previous experiment, we trained the optimal number of terms to be 10 terms. But as there was a copyright statement at the end of each document, we increased the number to 35 terms. For the same reason, the number to be added for cross-lingual run is increased from 20 terms to 50 terms from the top 20 documents. We also did query expansion before query translation using the corpus from TREC data vol. 5, the Foreign Broadcast Information Service (FBIS) files. FBIS is more than 400 MB in size and contains many international related documents. The number of expanded terms were 10 terms from the top 10 documents. The translation for the added terms in the source language were done by selecting the first two translations in the dictionary. All the parameters mentioned above were trained from our previous experiments if not otherwise stated.

Table 1:Official run results

Run	11-point	Relevant	R precision
CHUHK00CH1	0.2419	552	0.2524
CHUHK00XEC1	0.2583	514	0.2618

Table 1 is our official run results. **CHUHK00CH1** is the **monolingual** run and **CHUHK00XEC1** is the **cross-lingual run**. There are 663 relevant documents altogether in TREC-9. Table 2 is the component result for monolingual CHUHK00CH1 run and Table 3 is the component result for cross-lingual CHUHK00XEC1 run.

Table 2: Monolingual component results

Run	11-point
Lnu.Ltu	0.2288
above+expansion (top 10 docs, 35 terms)	0.2419 (+6%)

Table 3: Cross-lingual component results

Run	11-point
Lnu.Ltu	0.1862
above+expansion before query translation (top 10 docs, 10 terms)	0.2642
above+expansion after query translation (top 20 docs, 50 terms)	0.2583

There are some interesting phenomena from the results. Our final cross-lingual run exceeds its corresponding monolingual run, the performance ratio is 0.2583/0.2419=106.8%. However, if we compare the performance before any query expansions, that ratio is 0.1862/0.2288=81%.

For the cross-lingual run, the improvement of query expansion (from 0.1862 to 0.2642) before query translation is as high as 42%. We contribute the drastic improvement to the following reasons: First, the corpus "Foreign Broadcast Information Service" seems to contain many relevant documents to the queries in the source language and thus it is ideal for the source of blind relevance feedback. Second, selecting the first two translations for the expanded terms seems to be very successful in this context. Due to the time limitation, we could not investigate carefully on how to select the best translation candidates for isolated terms.

By looking at the results produced from the final query expansion, the improvement for monolingual is 6% (from 0.2288 to 0.2419), which is reasonable. However, the query expansion after translation led to performance degradation, from 0.2642 to 0.2583 even though the retrieved relevant documents increased from 495 to 514.

Table 4 concludes our performance comparing with other groups.

Table 4: Result comparison

Run	best	median	worst
CHUHK00CH1(mono)	3	12	10
CHUHK00XEC1(xlingual)	2	16	7

Analysis

In this section, we present the results from more runs to support our analysis. We aim to compare our proposed method with other related ones such as MI (mutual information) and highest term frequency methods. To do this, we did the following experiments.

1. The disambiguation is done by selecting the translation pairs with the highest MI value (denoted as **sim_mi**). MI is calculated as

$$I(a,b) = \log_2 \frac{P(x,y)}{p(x,y)}$$

2. The disambiguation is done by selecting the translation candidate with the highest term frequency appeared in the target corpus (denoted as **htf**). The similarity measure used in our official runs was used here except *idf* normalization, i.e., the disambiguation is done by selecting the translation pairs with the highest value of co- occurrence numbers (denoted as **sim_tf**).

Table 5 shows the retrieval results for the above runs in average precision (11-point). These runs were all done by query expansion before and after query translation with the same parameters used in our official cross-lingual runs.

Table 5: Comparative results

Run	11-point (b)	11-point (a)
mi	0.2552	0.2473
htf	0.2613	0.2544
sim_tf	0.2638	0.2564

MI method was worse than the others while htf, sim_tf and our sim_idf performed better. It is surprising that htf, the simplest method produced such a good result considering the efforts it takes. The result of sim_tf reveals similar message: high term frequency translations in the target are good indication of good translations. MI has the disadvantage of strongly favoring rarely appearing words.

We performed a final experiment trying to support our hypothesis that our overlapped concatenation of best selected translation pairs would enforce more correct translations to have higher weighting if the term frequency factor in the query is properly considered. If this is the case, it would be helpful for retrieval performance. To test this, we did Lnu.ntu weighting retrieval where term frequency factor is "augmented" comparing with Lnu.Ltu weighting.

The average 11-point recall precision is **0.2649** before the query expansion and **0.2596** after the query expansion. Although the increase is not obvious (0.2642 and 0.2583 in our official cross-lingual run), this result gives the highest figure comparing with all Lnu.Ltu runs.

We also observed consistent retrieval degradation after the final query expansion in all cross-lingual runs.

Conclusion:

We presented our disambiguation method by using similarity values between all adjacent words in the target language. It is based on the co-occurrence numbers within a sentence scope in the whole collection. On top of that, *idf* values of a word pair are used to normalize the co-occurrence numbers. We have shown that both co-occurrence number with or without normalization worked better than MI method. In particular, *idf* normalization is 4.5% (0.2583/0.2473) better than MI method in our experiments. More tests will be performed to further verify the improvement reported here.

This is our first participation in TREC. We reckon that this is a good start for our future research.

Acknowledgements

This project is partially supported by DARPA, USA, under the TIDES program (grant No.: N6601-00-1-8912) and Research Grant Committee, Hong Kong, under the direct grant initiative (project No.: 2050253).

References

[1] Amit Singhal, Chris Buckley, Mandar Mitra. Pivoted Document Length Normalization. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 21-29, 1996.

[2] Chris Buckley, Amit Singhal, Mandar Mitra, Gerard Salton. New Retrieval Approaches Using SMART: TREC 4. In D.K. Harman, editor, *Proceedings of the Fourth Text REtrieval Conference (TREC-4).* NIST Special Publication 500-236, 1996.

[3] Church, K.W. and Hanks, P. 1990. 'Word collocation norms, mutual information, and lexicography' *Computational linguistics* 16: 22-29