NTCIR-3 CLIR Experiments at MSRA

Hongzhao He[1]
Department of Computer Science and Engineering of Tianjin University, China

Jianfeng Gao
Natural Language Computing Group, Microsoft Research
jfgao@microsoft.com

Abstract

This paper describes three statistical models for the purpose of resolving query translation ambiguity for cross-language information retrieval (CLIR). First, a decaying co-occurrence model is present. It is an extension of traditional co-occurrence models in that it contains a decaying factor which decreases the mutual information when the distance between the terms increases. Second, a phrase translation model is described aiming to detect and translate noun phrases that are not stored in the dictionary. Finally, a triple translation model is proposed which provides a way of exploiting linguistic dependency information. We show experimentally improvements of using these models on TREC and NTCIR corpus.

1 Introduction

Microsoft Research Asia (MSRA) participated in the English-Chinese CLIR track at NTCIR-3. We focus our researches on resolving query translation ambiguity using statistical models.

While the dictionary-based CLIR is very popular due to its simplicity and the increasing availability of online machine readable dictionary, we always face the problem of translation ambiguity, i.e. how to select the correct translations among that provided by the dictionary. In this paper, two statistical models are explored to address this problem. They are (1) decaying co-occurrence model, and (2) triple translation model. The detailed description of these two models can be found at [3].

The decaying co-occurrence model is an extension of the traditional co-occurrence model commonly used in previous research. It contains a decaying factor which decreases the mutual information when the distance between the terms increases. The triple translation model is presented to capture syntactic dependence relations between words, and translate triples as a unit.

In the remainder of this paper, we will discuss in turn each model together with our approaches and CLIR results. The results include official runs we submitted and additional runs that we designed to help us explore the issues. Finally, we give our conclusions and present our future work.

2 Previous Research

There are amount of previous researches on resolving query translation ambiguity for CLIR.

Several recent studies [2, 5] suggested the utilization of co-occurrence information. A term similarity is determined by the mutual information (or its variants) between terms. Then the most similar translation term among those in the dictionary is selected. While such a selection may lead to some improvements over a simple translation selection, the co-occurrence of two terms within a predefined window is treated in the same way, no matter how far they are from each other. This is obviously against our intuition that the strength of the underlying relation is stronger when the distance between the two terms is shorter. Therefore, we will extend the previous methods by incorporating a decaying factor that decreases the mutual information exponentially when the distance between the terms increases.

An alternative approach is to identify phrases from queries, and translate them as a unit [1]. The key issue is that none of phrase dictionaries is complete. While recent research reports promising result on using statistical models for new phrase identification and translation [4], the techniques are by no means mature.

[1] This work has been done while the author was visiting Microsoft Research Asia.

While linguistic structures such as syntactic dependence relations are proved to be useful to resolve word ambiguity [6] there is little research on using them for resolving translation ambiguity for CLIR. Triple translation model described below aims to make use of the syntactic dependence for query translation.

3 Our Approach

Our research has been focused on resolving query translation ambiguity using statistical models. Given a query in English, each term of the query is represented by a context vector where each element is extracted from the context of the term, i.e. the sentence/query that contains the term. Based on the different types of vector used, we have different statistical models for term translation. As an example, given an English sentence "*I read an interesting BOOK on the airline*", in order to select the correct translation of the word *BOOK*, a context vector of *BOOK* can be generated for instance, the following three ways.

- First, we simply treat each word (excluding the word *BOOK*) in the sentence as a feature element of the vector. This is the vector used by co-occurrence models.
- Second, if we identify noun phrases (NP) containing the word *BOOK*, such as "*an interesting BOOK*", and treat the NP as a feature element of the vector. We then get the vector which can be used for phrase translation model.
- Finally, if we parse the sentence and extract the syntactic dependence relations of the word *BOOK*, such as (*read, object-verb, BOOK*) and (*interesting, adj-noun, BOOK*), we then generate the context vector which is used for triple translation model which we will describe in Section 3.3.

In NTCIR-3 CLIR experiments, we focus on the use of the co-occurrence model and the triple translation model. For completeness, we will also briefly describe the phrase translation model in Section 3.2. Readers who have interests can find more detailed discussion of these models at [3, 4].

3.1 Decaying co-occurrence model

It is assumed that the correct translations of query terms tend to co-occur in target language documents and incorrect translations do not. Therefore, given a set of original English query terms, we select for each term the best translation such that it co-occurs most often with other translation words in Chinese documents. Finding such an optimal set is computationally very costly. Therefore, an approximate algorithm is used [4, 5]. It works as follows. Given a set of n original query terms $\{s_1,...,s_n\}$, we first determine a set T_i of translation words for each s_i through the lexicon. Then we try to select the word in each T_i that has the highest degree of cohesion with the other sets of translation words. The set of best words from each translation set forms our query translation.

The cohesion between a term x and a set T of other terms is estimated as follows:

$$Cohesion\ (x,T) = \log(\sum_{y \in T} SIM\ (x,y))\ . \qquad (1)$$

Here, *SIM(x,y)* represents the similarity between two terms x and y. The traditional co-occurrence approach uses mutual information (or its variants) as term similarity. Mutual information is defined as follows:

$$MI(x,y) = P(x,y) * \log(\frac{P(x,y)}{P(x) * P(y)}), \qquad (2)$$

where

$$P(x,y) = \frac{C(x,y)}{\sum_{x',y'} C(x',y')}, \text{ and } P(x) = \frac{C(x)}{\sum_{x'} C(x')}.$$

Here *C(x, y)* is the frequency of co-occurrences of terms x and y within predefined windows (e.g. sentences) in the collection, *C(x)* is the number of occurrences of term x in the collection.

We observe that any co-occurrence within the windows is treated in the same way, no matter how far they are from each other. In reality, we find that closer words usually have stronger relationships, thus should be more similar. Therefore, we add a distance factor $D(x, y)$ in the mutual information calculation. This factor decreases exponentially when the distance between two terms x and y, increases, i.e.

$$D(x,y) = e^{-\alpha * (Dis(x,y)-1)}, \qquad (3)$$

where α is the decay rate, which is determined empirically (which is set to 0.8 in our experiments), and $Dis(x,y)$ is the average distance between x and y in the corpus. Therefore, term similarity in the extended co-occurrence model consists of two components: (1) the mutual information $MI(x,y)$ as defined before, and (2) the decaying factor $D(x,y)$:

$$SIM(x,y) = MI(x,y) * D(x,y) . \quad (4)$$

3.2 Phrase translation model

The translation of multi-word phrases as unit is expected to be more precise than a word-by-word translation since phrases usually have fewer senses. The major problem is how to identify and translate new phrases that are not stored in the dictionary.

We find that there are some translation patterns between Chinese and English. For example, an English noun phrase *ENP* of the pattern (NOUN-1 NOUN-2) is usually translated into the Chinese noun phrase *CNP* of the pattern (NOUN-1 NOUN-2), and a (NOUN-1 of NOUN-2) phrase is usually translated into the (NOUN-2 NOUN-1) sequence in Chinese. In our experiments, we defined 40 high-frequent translation patterns. Therefore, the phrase translation consists of two steps: given an *ENP* we first guess the Chinese NP patterns according to translation patterns, we then guess the *CNP* using the Chinese language model. The phrase translation model is used to rank the translated Chinese NP candidates.

The procedure can be formally described as follows: Given an English NP, $ENP=\{e_1,...,e_n\}$, with its NP pattern, *EPT*; for each English term e_i in *ENP*, we retrieve all the possible Chinese translations from the bilingual dictionary. We also get all the possible translation patterns *CPT* for *EPT*. Then the best Chinese translated phrase, $CNP^*=\{c_1,...,c_m\}$, is the one that maximizes the Equation (5) below.

$$CNP^* = \arg\max_{CNP} P(CNP \mid ENP) \quad (5)$$
$$= \arg\max_{CNP} P(ENP \mid CNP) \times P(CNP)$$

where $P(CNP)$ is a priori probability of words of the translated Chinese NP estimated by maximum likelihood estimation (MLE) using Chinese corpus, and $P(ENP|CNP)$ is the phrase translation probability. To incorporate translation patterns, we decompose $P(ENP|CNP)$ as Equation (6): We consider an NP (*ENP* or *CNP*) as a set of words (*E* or *C*) assembled by an NP pattern (*EPT* or *CPT*). Assuming that the translation of words and NP patterns are independent, we have

$$\begin{aligned} P(ENP \mid CNP) &= P(E, EPT \mid C, CPT) \\ &= P(E \mid C, CPT) \times P(EPT \mid C, CPT) \\ &= P(E \mid C) \times P(EPT \mid CPT) \end{aligned} \quad (6)$$

where $P(E|C)$ is the translation probability from Chinese words *C* in *CNP* to English words *E* in *ENP*. $P(EPT|CPT)$ is the probability of the translation pattern *EPT* (i.e. the order of translation words), given the Chinese pattern *CNP*. Both probabilities are estimated by MLE using word-aligned bilingual corpus. The detailed description of the phrase translation model can be found at [4].

3.3 Triple translation model

A triple represents a dependence relationship between two words, such as verb-object, subject-verb, etc. We represent a triple as (w,r,w'), where w and w' are words and r is the dependence relation. It means that w' has an r relation with w. For example, a triple (*read, object-verb, book*) means that "book" is the object of the verb "read".

Among all the dependence relations, we only consider the following four that can be detected precisely using our parser: (1) sub-verb, (2) verb-object, (3) adjective-noun, and (4) adverb-verb[2].

It is our observation that there is a strong correspondence in dependence relations in the translation between English and Chinese, despite the great differences between the two languages. As reported in [3], more than 80% of the above four dependence relations have one-one mapping between English and Chinese. For example, an object-verb relation in English (e.g. (*read, object-verb, book*)) is usually translated into the same object-verb relation in Chinese (e.g. (读, *object-verb*, 书).

This means, for an English triple $ETP = (w_e,r_e,w_e')$, the most likely Chinese translation should also be a triple $CTP = (w_c,r_c,w_c')$, where w_c

[2] Although noun-noun relation is also very important for IR and for appropriate translation of phrases, it is handled by phrase translation model described in the last section.

and w_c' are the Chinese translations of the English terms w_e and w_e', respectively, and r_c is the Chinese counterpart of r_e. The triple translation model is used to rank the translated Chinese triple candidates.

The procedure can be formally described as follows: Given an English triple $ETP = (w_e, r_e, w_e')$, and the set of its candidate translating Chinese triples CTP, the best Chinese triple $CTP^* = (w_c, r_c, w_c')$ is the one that maximizes the Equation below:

$$CTP^* = \arg\max_{CTP} P(CTP \mid ETP) = \arg\max_{CTP} P(ETP \mid CTP) \times P(CTP), \quad (7)$$

where $P(CTP)$ is the *a priori* probability of words of the translated Chinese triple, which is estimated by MLE using Chinese corpus. Now, the remaining problem is how to estimate the triple translation probability $P(ETP|CTP)$. It can be of course, estimated from parallel bilingual corpus. But the parallel bilingual corpus which should be sufficient enough for reliable estimation is not always available since the number of triples is much larger than the number of words. We then decompose the triple translation probability into 3 components as follows:

$$P(ETP \mid CTP) \propto Score(ETP \mid CTP) \times P(r_e \mid r_c) = Sim(w_e, w_c) \times Sim(w_e', w_c') \times P(r_e \mid r_c) \quad (8)$$

As the correspondence between the same dependence relation across English and Chinese is strong, we simply assume $P(r_e|r_c) = 1$ for the corresponding r_c and r_e, and $P(r_e|r_c) = 0$ for the other cases. Here, we use the word similarity $Sim(w_e,w_c)$ instead of translation probability between w_c and w_e because we do not suppose to have enough parallel bilingual corpus, and the word similarity can be estimated using non-parallel bilingual corpus as described below.

Consider an English/Chinese word pair w_e/w_c which is represented by a context vector with each element a triple pair $(w_e, r_e, w_e')/(w_c, r_c, w_c')$. If w_e' and w_c' form a translation pair stored in a bilingual dictionary, and r_e and r_c are corresponding dependence relations, we say that w_e and w_c have a common triple pair $(w_e, r_e, w_e')/(w_c, r_c, w_c')$. Using an information-theoretic definition, $Sim(w_e,w_c)$ is measured by the ratio between the amount of information needed to describe the commonality of w_e and w_c (denoted by $I(common(w_e,w_c))$, which can be estimated approximately in our experiments, as the number of common triple pairs of the English/Chinese word pair w_e/w_c) and the information needed to fully describe w_e and w_c (denoted by $I(describe(w))$, which can be estimated approximately in our experiments, as the number of triples containing w.):

$$Sim(w_e, w_c) = \frac{I(common(w_e, w_c))}{I(describe(w_e)) + I(describe(w_c))} \quad (9)$$

The detailed description of the triple translation model can be found at [9, 3].

3.4 Summary of models

As a summary of the above three models, we compared them in terms of (1) natural language processing techniques we used, (2) linguistic knowledge we exploited, and (3) training corpus needed for parameter estimation. The comparison result is shown in Table 1.

4 Results

4.1 Tests on TREC-9

We first test the decaying co-occurrence model and the triple translation model on the TREC-9 Chinese corpus[3]. This corpus contains articles published in Hong Kong Commercial Daily, Hong Kong Daily News, and Takungpao. They amount to 260MB. It also contains 25 English queries (with translated Chinese queries) evaluated by the NIST (National Institute of Standards and Technology). We use long queries in our experiments. Chinese texts are segmented into words using a dictionary containing 220,000 words. The bilingual lexicon we used contains 401,477 English entries, including 109,841 words, and 291,636 phrases.

The Okapi system with BM2500 weighting (Robertson and Walker 2000) is used as the basic retrieval system. The main evaluation metric is interpolated 11-point average precision. The following methods are compared to investigate the effectiveness of our models for query translation:

[3] The results have been reported in [3]

Model	NLP techniques	Use of linguistic knowledge	Training data
Decaying Co-occurrence	Word morphological analysis	Little	Chinese corpus
Phrase Translation	NP chunking	Shallow (NP)	Parallel bilingual corpus
Triple Translation	Full parsing	Deep (syntactic dependency)	Non-parallel bilingual corpus

Table 1: Summary of three models

1. *Monolingual*: retrieval using the manually translated Chinese queries provided with the corpus.
2. *Simple translation*: retrieval using query translation obtained by taking the first translations from the bilingual dictionary.
3. *Best-sense translation*: retrieval using translation words selected manually from the dictionary, one translation per word. This method reflects the upper bound performance using the dictionary.
4. Our methods that incorporate the use of triple translation model and the decaying co-occurrence model.

Previous work [5, 8] showed that if multiple translations of a term were accepted in query translation, it is possible to obtain better performance of cross-language retrieval than that of monolingual retrieval, partly because of the query expansion effect. In order to separate the impact of query expansion from that of query translation, in our experiments, each English query term is translated by only one Chinese term[4].

The results of this series of experiments on query translation are shown in Table 2 and Figure 1. As shown in rows 4 and 5 of Table 5, both the co-occurrence model and the triple translation model bring substantial improvements over simple translation. The use of the decaying co-occurrence model results in a 48% improvement, which is statistically significant (p-value = 0.008). Row 6 corresponds to the preferred translation strategy of combining two models: triples are first identified and translated, remaining terms are then translated by the decaying co-occurrence model. It shows that using both models in our query translation process, we achieve the best performance. It is better than using the co-occurrence model alone by 5%.

	Methods	**Avg. P.**	**% Mono. IR**
	Monolingual	0.2862	
	Simple translation	0.1613	56%
	Best-sense translation	0.2730	95%
	2 + co-occurrence model	0.2392	84%
	2 + triple model	0.1908	67%
	5 + co-occurrence model	0.2517	88%

Table 2: Retrieval effectiveness on TREC-9 corpus

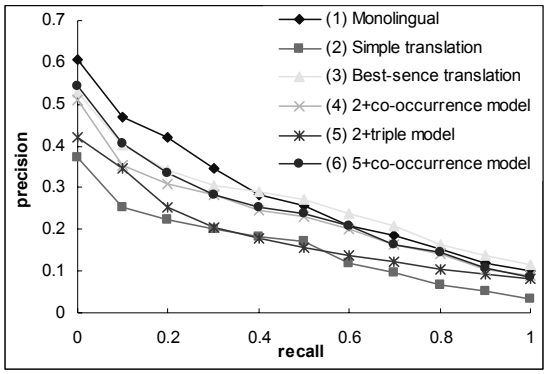

Figure 1: P-R curves (TREC-9 queries)

4.2 Experiments in NTCIR-3

NTCIR-3 collection contain 381,681 Chinese documents and 42 topics[5]. All Chinese documents are news articles. Some statistical data are shown in Table 3.

[4] This follows a suggestion by Douglas Oard.

[5] The IDs of topics in 1998-1999 Topic Set used in this collection are 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20, 21, 22, 23, 24, 25, 27, 32, 33, 34, 35, 36, 37, 38, 39, 40, 42, 43, 45, 46, 47, 48, 49, and 50.

Collection	# of Documents
CIRB011 (1998-1999): Chinese	132,173
CIRB020 (1998-1999): United Daily News (1998-1999): Chinese	249,508
EIRB010 (1998-1999): Taiwan News and Chinatimes English News (1998-1999): English	10,204

Table 3. Statistics of Document Set for CLIR Task of NTCIR Workshop 3

For each topic-document pair, 4-level relevance is defined, as show in Table 4. Therefore, for each submitted run, 2 scoring results are created. One is rigid and the other is relaxed. S and A will be regarded as relevant in the rigid mode; S, A, and B will be regarded as relevant in the relaxed mode.

Meaning	Symbolic Score	Numerical Score
Highly Relevant	S	3
Relevant	A	2
Partially Relevant	B	1
Irrelevant	C	0

Table 4. 4-level relevance used in NTCIR

We submitted 3 runs to English-Chinese CLIR track. Table 5 shows the average precision (Avg. P.) of each run together with comparison with the average performance (Avg. IR) of all participants.

In all three runs, we used preferred query translation strategy of combining decaying co-occurrence model and triple translation model as described in Section 4. In MSRA-01, we used the short query set, while in MSRA-02 and MSRA-02 we used the long query set. In MSRA-01 and MSRA-02, we performed the 2-stage pseudo relevance feedback, while in MSRA-03, no query expansion is used.

5 Conclusions and Future Work

This paper describes several approaches to resolving query translation ambiguity using statistical models. For the sake of selecting the correct translation of a query terms, a context vector is generated by extracting features from the query/sentence that contains the term. By using different levels of linguistic knowledge, extracted features vary from (1) words, (2) phrases, to (3) syntactic dependence relations. Depending on different types of feature we used, we propose three different statistical models:

1. The decaying co-occurrence model uses a bag of words as context vector. It is an extension of the traditional co-occurrence model in that it contains a decaying factor which decreases the mutual information when the distance between the terms increases.
2. The phrase translation model aims to detect and translate NPs that are not stored in the dictionary.
3. The triple translation model provides a way of exploiting linguistic dependency information.

We show experimentally that translating NPs and triples as a unit achieves much more precise translations. While difficult-to-obtain parallel bilingual corpus is required for phrase translation model training, non-parallel corpus can be used for triple translation model training. This indicates that the triple translation model may be more applicable in realistic systems.

Our future work will be focused on the methods of combining three proposed models. The combination can be achieved by combining different model scores or combining feature sets. Another important problem we currently have with the triple translation is the robustness of the parser. A certain portion of incorrect triples are extracted, especially those with low frequencies. Many other triples cannot be extracted because the parser fails to parse the sentence completely. In order to increase the impact of triple translation, the robustness of the parser has to be improved. In fact, as we only need to extract triples, a partial parser

Methods	Relax			Rigid		
	Avg. P.	Avg. IR	% Avg. IR	Avg. P.	Avg. IR	% Avg. IR
MSRA-01	0.1921	0.1144	168%	0.1629	0.0903	175%
MSRA-02	0.2781	0.1861	149%	0.2361	0.1509	156%
MSRA-03	0.2581	0.1861	139%	0.2179	0.1509	144%

Table 5. Average precision of the submitted runs

may be more suitable. This alternative will also be investigated in the future.

Acknowledgements

We would like to thank Changning Huang, and Ming Zhou for their comments, thank Guihong Cao and Min Zhang for their help in our experiments. We would also like to thank Stephen Robertson and Stephen Walker for providing us the Okapi system.

References

[1] Ballesteros, L., and Croft, W. B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In: *ACM SIGIR'97*. pp. 84-91.

[2] Ballesteros, L., and Croft, W. B. (1998). Resolving ambiguity for cross-language retrieval. In: *ACM SIGIR'98.* Melbourne, Australia., pp. 64-71

[3] Gao, J., Nie, J. Y., He, H., Chen, W., and Zhou, M. (2002). Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations. In: ACM SIGIR'02, Tempere, Finland, pp 183 – 190.

[4] Gao, J., Nie, J. Y., Zhang, J., Xun, E., Zhou, M., and Huang, C. (2001) Improving query translation for CLIR using statistical Models. In: *ACM SIGIR'01,* New Orleans, Louisiana, pp. 96-104.

[5] Gao, J., Nie, J. Y., Zhang, J., Xun, E., Su, Y., Zhou, M., and Huang, C. (2000). TREC-9 CLIR experiments at MSRCN. In *TREC-9*, pp. 343-353..

[6] Lin, D. (1997). Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of ACL/EACL-97*, Madrid, pp. 64-71.

[7] Robertson, S. E., and Walker, S. (2000). Microsoft Cambridge at TREC-9: Filtering track. In *TREC-9*.

[8] Xu, J., and Weischedel, R. (2000). TREC-9 cross-lingual retrieval at BBN. In *TREC-9*, pp. 106-116.

[9] Zhou, M., Ding, Y., and Huang, C. (2001). Improving translation selection with a new translation model trained by independent monolingual corpora. *Computational linguistics and Chinese Language Processing*. Vol. 6, No. 1, pp 1-26.