Resolving Query Translation Ambiguity using a Decaying Co-occurrence Model and Syntactic Dependence Relations

Jianfeng Gao[*], Jian-Yun Nie[**], Hongzhao He[#], Weijun Chen[##], Ming Zhou[*]
[*] Microsoft Research, Asia, Email: {jfgao, mingzhou, cnhuang}@microsoft.com
[**] Université de Montréal, Email: nie@iro.umontreal.ca
[#]Tianjin University, China.
[##]Tsinghua University, China.

ABSTRACT
Bilingual dictionaries have been commonly used for query translation in cross-language information retrieval (CLIR). However, we are faced with the problem of translation selection. Several recent studies suggested the utilization of term co-occurrences in this selection. This paper presents two extensions to improve them. First, we extend the basic co-occurrence model by adding a decaying factor that decreases the mutual information when the distance between the terms increases. Second, we incorporate a triple translation model, in which syntactic dependence relations (represented as triples) are integrated. Our evaluation on translation accuracy shows that translating triples as units is more precise than a word-by-word translation. Our CLIR experiments show that the addition of the decaying factor leads to substantial improvements of the basic co-occurrence model; and the triple translation model brings further improvements.

Categories and Subject Descriptors
H.3 [**Information Systems**]: Information Storage and Retrieval. F.1.2 [**Computation by Abstract Devices**]: Modes of Computation – *Probabilistic computation.* H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *Natural language*

General Terms
Algorithms, Measurement, Documentation, Performance, Design, Reliability, Experimentation, Languages, Theory.

Keywords
Query translation, CLIR, Statistical model, Parse, Co-occurrence

1. INTRODUCTION
Bilingual dictionaries have been commonly used for query translation in CLIR, but we are always faced with the problem of translation ambiguity, i.e. several translation words are provided for a source word by the dictionary, but only part of them are appropriate.

Recently, several studies [2, 4, 7, 11, 14, 19] have been proposed the use of co-occurrence information to deal with this problem. In those studies, usually a mutual information value between terms is estimated according to their co-occurrences within a predefined window. The translation word that has the highest mutual information score with the other translation words is selected. While these approaches have successfully improved the quality of query translation, there are still two main drawbacks. First, the standard calculation of mutual information does not take into account the distance between terms; whereas our intuition is that closer terms have stronger relationships. Second, the syntactic dependency between terms is not considered.

This paper presents two methods to improve the basic co-occurrence approaches. First, we incorporate a *decaying factor* that decreases the mutual information when the distance between the terms increases. Second, we present a statistical *triple* translation model, in which syntactic dependence relations (represented as triples) are integrated.

Our evaluation results show that the triple translation is more precise than the word-by-word translation with the co-occurrence model. The CLIR experiments on TREC collections show that the decaying co-occurrence method performs better than the basic co-occurrence method, and the triple translation model brings additional improvements.

The remainder of this paper is organized as follows: Section 2 provides a brief description on the related work. Sections 3 and 4 describe our co-occurrence approach and the triple translation model. Section 5 evaluates the proposed approaches. Section 6 presents our conclusion.

2. DEALING WITH TRANSLATION AMBIGUITY
To deal with the selection of the correct translation terms from a bilingual dictionary, several recent studies [2, 4, 7, 11, 14, 19] 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, we notice that in those studies, the co-occurrence of two terms within

SIGIR'02, August 11-15, 2002, Tampere, Finland,

[# ##] This work was done while these authors were visiting Microsoft Research, Asia.

a predefined scope is treated in the same way, no matter how far they are from each other. This conception is 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 when the distance between the terms increases. A similar idea has been applied successfully to statistical language modeling [5], showing improved performance of the cache language model. We expect similar improvements on CLIR, and this will be confirmed by our experiments.

While there is little research on using syntactic approaches for resolving translation ambiguity for CLIR, linguistic structures have been successfully exploited in other applications. For example, in [12], syntactic dependency was exploited for resolving word sense ambiguity. A term similarity measure based on triples was first proposed in [13]. [18] extended this measure to a cross-language term similarity term, and presented a statistical translation model for collocation translation. This model is adopted in this study for triple translations.

Another problem of simple dictionary approaches is the poor coverage of the dictionary. We demonstrate that the use of the triple translation model provides an interesting alternative solution to this problem.

In this paper, we focus the use of co-occurrence model and triple model for query translation. In our CLIR experiments, the query translation process may be summarized as follows:

(1) Triples are first identified in English (source language) queries using a dependency parser.

(2) The translation of the identified triples is determined using a triple translation model, which is trained on unrelated English and Chinese corpora.

(3) The remaining words in the query are translated as terms by the decaying co-occurrence model, which determine the best translation term among all those stored in the dictionary.

3. USING CO-OCCURRENCE INFORMATION FOR TRANSLATION SELECTION

3.1 Principle of the basic co-occurrence approach

A correct translation word is the one that fits well the context of the whole sentence. The original sentence (or query) reflects well the context of the sentence; but it would be difficult to find a way to directly compare a translation word with the original sentence, unless there is a similarity measure between words across languages. Another possible way is to consider that all the translation words selected for the other words of the source sentence form an alternative specification of the context. Then a good translation word is the one that has a high *cohesion* with the other translation words. The advantage of this alternative approach is that there is no need to have a cross-language word comparison. Only relationships between words of the same language are used. They can be obtained through their co-occurrences in a monolingual text corpus. This is the principle of co-occurrence approach to translation selection. It can also be expressed as follows: correct translations of query words tend to co-occur in the target language and incorrect translations do not [2]. A similar principle is also used in [14].

3.2 Decaying Co-occurrence Model

The basic co-occurrence approach uses mutual information (or its variants) as term similarity. Mutual information is defined as follows:

$$MI(x,y) = P(x,y) * \log(\frac{P(x,y)}{P(x) * P(y)}), \quad (1)$$

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.

$$D(x,y) = e^{-\alpha * (Dis(x,y)-1)}, \quad (2)$$

where α is the decay rate, which is determined empirically (see Section 3.3), and *Dis*(*x,y*) is the average distance between *x* and *y* in the corpus.

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*):

$$SIM(x,y) = MI(x,y) * D(x,y). \quad (3)$$

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

$$Cohesion\ (x,T) = \log(\sum_{y \in T} SIM(x,y)). \quad (4)$$

In query translation, each source word is translated by one target word. The internal cohesion of a set of target words is the sum of the cohesions of each target word with all the other target words. The set with the highest internal cohesion is selected as the query translation.

Finding an optimal set of translation words may be computationally very costly. Therefore, an approximate greedy algorithm has been proposed in [8] to select the best set. We use the same algorithm here.

3.3 Model Estimation and Evaluation in Monolingual IR

The decaying co-occurrence model is estimated on a collection of Chinese newspaper articles consisting of approximately 80

million characters. To investigate the effectiveness of the decaying model, we performed a preliminary test with query expansion in monolingual IR. For each query term, we expand it by an additional term that has the highest cohesion value with the other words of the original query. This expansion task is very similar to the translation selection in CLIR. Therefore, it gives a good indication on the possible impact on query translation.

A Chinese synonym dictionary is generated for query expansion from the LDC bilingual dictionaries[1] as follows: For each Chinese term *c*, and each of its English translation *e*, we consider all the Chinese translations *c'* of *e* as synonyms of *c*.

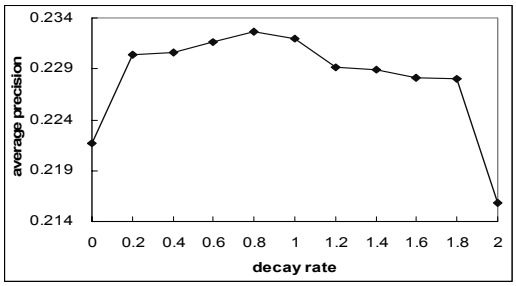

Figure 1: The impact of decaying factor on query expansion

Our experiments are carried out on the TREC-9 Chinese collection [16]. Figure 1 shows the retrieval results with query expansion while the decay rate varies. It can be seen that the decaying co-occurrence model performs generally better than the basic co-occurrence model (when decay rate = 0). With a decay rate of 0.8, we obtain the best performance of the average precision 23.3%, which is 5% better than the basic model. These experiments show that the decaying factor allows us to better distinguish strong and weak term relationships. As the problem of translation selection in CLIR is similar to this expansion task, we can expect a similar effect with the decaying factor. In our later experiments, the decay rate will be set to 0.8.

4. THE USE OF TRIPLES

The incorporation of triples is an attempt to take into account some syntactic dependences in translation selection. Our hypothesis is that strong syntactic dependences usually remain in the translations, and an ideal query translation should contain the same syntactic dependences as in the original query. Syntactic dependences also provide an additional criterion to the earlier cohesion measure: A good translation word should not only co-occur with other translation words, but also have the required syntactic dependence relations with them.

4.1 Principle

A triple represents a dependence relationship between two words, such as verb-object, subject-verb, and so on. 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, (have, sub-verb, I) means that "I" is the subject of the verb "have". Figure 2 shows an English sentence and the triples that one can extract from it.

[1] http://morph.ldc.upenn.edu/Projects/Chinese/.

Sentence	I have a brown dog.
Triples	(have, sub-verb, I) (have, verb-obj, dog) (dog, adj-noun, brown) (dog, det-noun, a)

Figure 2: Examples of Dependence relations or triples

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. For example, a sub-verb relation in English (e.g. (have, sub-verb, I)) is usually translated into the same sub-verb relation in Chinese (e.g. (有, sub-verb, 我). 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 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.

4.2 Correspondence of Dependence Relations between English and Chinese

To test the assumption of strong correspondence of dependence relations between Chinese and English, we used a word-aligned bilingual corpus, which consists of 60,000 pairs of Chinese and English sentences. The corpus was first parsed using an English and Chinese parser– NLPWIN[3]. The four types of dependency relationship were extracted. We analyzed the correspondence on dependence relations between Chinese and English. The results are shown in Table 1. As we can see, more than 80% of dependence relations of sub-verb, adj-noun, and adv-verb have one-one mappings between English and Chinese, while the mapping rate of verb-object is approximately 65%.

Further analyses showed that the mapping errors of verb-object occur in the following situations: (1) one single English verb maps a Chinese triple (e.g. read → 读[read] v | 书[book] o), or (2) an English verb-prep-object sequence maps a Chinese verb-object sequence (e.g. change-to-currency → 用[use] v | 货币[currency] o). As the first case is not a triple translation, and will not affect the triple model, it is ignored. The second problem is quite common. In fact, the combination of verb-preposition in English is very often translated into a single verb in Chinese. If we consider such a combination in English as "verb", the mapping rate of verb-object is increased to more than 80% (see Table 1 – (*) case). This is the way we will use to map verb-object patterns between Chinese and English.

[2] Although noun-noun relation is also very important for IR and for appropriate translation of phrases, it is beyond the scope of this paper. Please see [8] for a detailed description of noun phrase detection and translation for CLIR.

[3] NLPWIN parser is developed at Microsoft Research. It construct a parse tree fro a sentence, then a logical form. Triple is one of the logical forms.

So, globally, the mapping rate is generally higher than 80%. We can conclude that there is indeed a very strong correspondence between Chinese and English in the four dependence relations considered.

Dependency	sub-verb	adj-noun	adv-verb	verb-obj	verb-obj (*)
Mapping Rate	81.2%	81.0%	80.9%	64.8%	80.7%

Table 1: Triple correspondence between Chinese and English

4.3 Triple Translation Model

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:

$$CTP^* = \arg\max_{CTP} P(CTP \mid ETP) = \arg\max_{CTP} P(ETP \mid CTP) \times P(CTP), \quad (5)$$

where $P(ETP|CTP)$ is the translation probability, $P(CTP)$ is the *a priori* probability of words of the translated Chinese triple. $P(CTP)$ can be determined using maximum likelihood estimation (MLE) by

$$P(CTP) = \frac{|w_c, r_c, w_c'|}{|*,*,*|}, \quad (6)$$

where $|w,r,w'|$ is the frequency count of the triple (w,r,w'). The wildcard symbol * means that it can be any word/relation.

$P(ETP|CTP)$ cannot be estimated directly because there is no large triple-aligned corpus available (the one we used in Section 4.2. is too small). Therefore, we assume that the translations of the English terms w_e and w_e', and the English dependency relation r_e are independent, so $P(ETP|CTP)$ can be decomposed as follows:

$$P(ETP \mid CTP) = P(w_e, r_e, w_e' \mid w_c, r_c, w_c') = P(w_e \mid w_c) \times P(w_e' \mid w_c') \times P(r_e \mid r_c). \quad (7)$$

Notice that $P(w_e|w_c)$ and $P(w_e'|w_c')$ are translation probabilities within triples. They are different from the unrestricted probabilities such as the ones in IBM models [3].

Between the same dependence relation r_c and r_e, $P(r_e|r_c)$ could be estimated from Table 1. However, we do not have the statistics for different r_c and r_e. 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.

The remaining elements to be estimated are the within-triple word translation probabilities $P(w_e|w_c)$ and $P(w_e'|w_c')$. We further assume that they can be estimated by a similarity score (its estimation will be described later) between the words, i.e.

$$P(w_e|w_c) \propto Sim(w_e, w_c),$$

and $$P(w_e'|w_c') \propto Sim(w_e', w_c').$$

Then Equation (7) can be rewritten as:

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

Now, the problem is the definition of $Sim(w_e, w_c)$[4]. This is done by extending the approach of [13] which estimates the similarity $Sim(w_1, w_2)$ for w_1 and w_2 in the same language. Let us first describe the approach of [13]. The basic idea is to consider all the dependence relations with a word w as forming its *dependence context*, denoted by $T(w)$. For example, in Figure 2, the dependence context of "dog" consists of three triples, i.e. $T(\text{"dog"})$={(have, verb-obj, *), (*, adj-noun, brown), (*, det-noun, a)}. It is assumed that two words are likely to have similar meanings if their dependence contexts are identical. For the sake of simplicity, in what follows, we use (r,w') to denote either $(*,r,w')$ or $(w',r,*)$.

Using an information-theoretic definition, $Sim(w_1, w_2)$ is measured by the ratio between the amount of information needed to describe the commonality of w_1 and w_2 (denoted by $I(common(w_1, w_2))$) and the information needed to fully describe what w_1 and w_2 are (denoted by $I(describe(w_1, w_2))$):

$$Sim(w_1, w_2) = \frac{I(common(w_1, w_2))}{I(describe(w_1, w_2))} \quad (9)$$

Let $I(w,r,w')$ be the amount of information needed to describe a triple (w,r,w'), which is estimated by Equation (10)[5].

$$I(w, r, w') = \log \frac{P(w \mid r, w')}{P(w \mid r)}, \quad (10)$$

where

$$P(w \mid r, w') = \frac{|w, r, w'|}{|*, r, w'|}, \text{ and}$$

$$P(w \mid r) = \frac{|w, r, *|}{|*, r, *|}.$$

Then, assume that information can be additive, $I(common(w_1, w_2))$ is calculated as the sum of the information contained in common triples belonging to both dependence context sets $T(w_1)$ and $T(w_2)$:

$$\sum_{(r,w) \in T(w_1) \cap T(w_2)} (I(w_1, r, w) + I(w_2, r, w)) \quad (11)$$

Similarly, $I(describe(w_1, w_2))$ is calculated as the sum of the information contained in triples belonging to either dependency context sets $T(w_1)$ or $T(w_2)$:

$$\sum_{(r,w) \in T(w_1)} I(w_1, r, w) + \sum_{(r,w) \in T(w_2)} I(w_2, r, w) \quad (12)$$

Now, let us explain how Equation (9) is extended in [18] to the cross-language word similarity $Sim(w_e, w_c)$.

The basic idea is the same: two words are similar if they occur in similar dependence contexts. Simply, for an English dependence context (r_e, w'_e), a Chinese context (r_c, w'_c) is similar if r_c corresponds to r_e, and w'_c is a translation of w'_e stored in a bilingual dictionary. In this case, we call the Chinese dependence context a possible translation of the English dependence context.

The probability of such a translation may be estimated as $P(w'_c|w'_e) * P(r_c|r_e)$, where $P(w'_c|w'_e)$ is the translation probability from w'_e to w'_c. We assign an equal probability to all the w'_c

[4] The definition is a simplified version of that in [18].

[5] This formula is slightly changed from that used in [13].

stored in the bilingual dictionary for w'_e, i.e. $P(w'_c|w'_e) = 1/(\#$ of translation words for $w'_e)$ if w'_c is a dictionary translation of w'_e; otherwise, $P(w'_c|w'_e) = 0$. $P(r_c|r_e)$ is estimated as before.

So, the cross-language commonality between w_e and w_c $I(common(w_c,w_e))$ is modified from Equation (11) as follows:

$$\sum_{(r_c,w_c')\in T(w_c)} I(w_c,r_c,w_c')\times P(w_c'|w_e')\times P(r_c|r_e) \quad (13)$$
$$+\sum_{(r_e,w_e')\in T(w_e)} I(w_e,r_e,w_e')\times P(w_e'|w_c')\times P(r_e|r_c)$$

The descriptions of w_c and w_e are defined as before as follows:

$$I(describe\ (w_c)) = \sum_{(r,w)\in T(w_c)} I(w_c,r,w) \text{ and} \quad (14)$$

$$I(describe\ (w_e)) = \sum_{(r,w)\in T(w_e)} I(w_e,r,w)$$

4.4 Model Training

One advantage of the triple translation model is that it can be trained on unrelated English and Chinese corpora. The triple model only requires the estimation of triple probabilities separately in the two languages, then through a bilingual dictionary and the assumption of strong correspondence of dependence relations, the whole triples in the two languages become related. There is a risk that unrelated triples in Chinese and English can be connected with this method. However, as the conditions that are used to make the connection are quite strong (i.e. possible words translations in the same triple structure), we believe that this risk, although exists, is not overwhelming. Our later experiments will show that the translation relations created among triples in this way bring significant benefits.

In our training, we used an English text corpus that contains articles published in the Wall Street Journal from 1987 to 1992, which amount to 750MB. The Chinese text corpus we used contains articles published in the People's Daily from 1980 to 1998. They amount to 1200MB. NLPWIN is used to extract triples in both corpora. Table 2 shows the number of different triples extracted.

As described in Section 4.3, for triple translation, two parameters have to be estimated: $P(CTP)$ and $Sim(w_e,w_c)$. They are respectively estimated according to Equations (6) and Equations (13) and (14).

Language	sub-verb	adj-noun	adv-verb	verb-object
Chinese	26,773,214	14,707,246	10,191,300	22,259,701
English	6,475,461	2,026,177	741,719	6,558,566

Table 2: Statistics of the extracted triples

5. EXPERIMENTS AND EXTENSION

In this section, we first present our experiments on the accuracy of triple translations, then the evaluations on CLIR.

5.1 The Evaluation of Triple Translations

5.1.1 Verb-object triple translation results

Verb-object triples seem to be the ones that represent the most serious translation ambiguity problem among the four main triple types. Therefore, we focus on this type of triple in our first experiments.

From the set of Chinese and English verb-object triples described in Section 4.4, 80% of them served as the training data. From the remaining 20% triples, several test sets are extracted, corresponding to different characteristics:

(1) **Test-set 1 (frequent verbs)**: This set contains 1000 triples with frequent verbs.

(2) **Test-set 2 (infrequent verbs)**: This set contains 275 triples with infrequent verbs.

(3) **Test-set 3 (frequent triples)**: This set contains frequent 1000 triples.

(4) **Test-set 4 (infrequent triples)**: This set contains 700 infrequent triples.

The translation performance is measured by accuracy, defined as (# of correct translated triples / # of triples in the test set).

The following three translation methods are compared:

(1) **Method A**: Each English word in a triple is translated by the most frequent translation word in the dictionary.

(2) **Method B**: The translation words are determined by the decaying co-occurrence model.

(3) **Method C**: The triple model is used.

The results are shown in Table 3. We can see that Method C achieves the highest performance in all four test sets. This shows that the most precise translation of a triple is obtained by using the triple model. These results also confirm that the risk of relating unrelated triples by using non-parallel corpora for training is small. We also notice that the translation accuracy varies very consistently among different test sets. This shows that the triple model performs well not only for frequent triples and verbs, but also for infrequent ones.

	Test-set 1	Test-set 2	Test-set 3	Test-set 4
Method A	42.4%	44.0%	54.1%	54.1%
Method B	55.0%	67.3%	62.3%	67.6%
Method C	74.2%	80.9%	73.3%	81.0%

Table 3: Verb-object triple translation results

We also observe that Method B is generally better than Method A. This determines the following preferred query translation strategy: When a triple is identified, it is translated by the triple model; the remaining words are translated by the co-occurrence model if covered by it; otherwise, the simple translation method is used.

5.1.2 Beyond the limit of the bilingual dictionary

Due to the limited coverage of the dictionary, a correct translation may not be stored in the dictionary. This naturally limit the coverage of triple translations. In order to generate more triple translations, the following two expansion methods are used to create new candidate translation triples:

(1) *Expansion by synonyms*: An English synonym dictionary is first generated from the bilingual dictionaries using the method described in Section 3.3. Translations of all the synonyms of the term w_e are added as the possible translations of w_e. This expansion corresponds to the method of pre- and post-translation expansion in [2].

(2) *Expansion by triples*: Given an English triple (w_e, r_e, w'_e), for each Chinese translation w_c of w_e stored in the dictionary, we select all Chinese words w'_c such that $I(w_c, r_c, w'_c)>0$ as the extended translations of w'_e; The $I(w_c, r_c, w'_c)>0$ condition means that (w_c, r_c, w'_c) is a significant triple (not a noise) in the corpus. The expansion on Chinese words is done similarly.

We now define a new word translation relation $s(w_c|w_e)$ between words in the expansion context. This relation is no longer a probability, but a generalized score. We use it to replace $P(w_c|w_e)$ in Equation (13). It is defined as follows. Let $|Tran1(w_e)|$, $|Trans2(w_e)|$ and $|Trans3(w_e)|$ be respectively the size of the translation set obtained by the non-expanded method and by the two expansion methods.

- For *unexpanded* method, each translation in $Tran1(w_e)$ will be assigned an equal score $1/|Tran1(w_e)|$ (this is the same as before).
- For *synonym expansion*, if a translation belongs to both $Tran1(w_e)$ and $Tran2(w_e)$, the score is $0.6/|Tran1(w_e)| + 0.4/|Tran2(w_e)|$; if it only belongs to $Tran1(w_e)$, then it is $0.6/|Tran1(w_e)|$; otherwise, it is 0.
- Similarly, for *triple model expansion*, if a translation belongs to both $Tran1(w_e)$ and $Tran3(w)_e$, the score is $0.6/|Tran1(w_e)| + 0.4 / |Tran3(w_e)|$; if it only belongs to $Tran1(w_e)$, then it is $0.6/|Tran1(w_e)|$; otherwise, it is 0.

The values of 0.6 and 0.4 are chosen empirically through experiments.

For example, in the triple expansion case, initially the dictionary does not contain "怀" (be pregnant [of a child]) as a translation of "bear". The triple model expansion will generate an additional translation "怀" (be pregnant [of a child]) for "bear" because "孩子" (child) and its translation "child" are frequent objects of these two verbs. That is s(怀 |bear) is relatively high. In addition, the triples (怀, verb-obj, 孩子) and (bear, verb-obj, child) are quite frequent in the corpora, so both I(怀, verb-obj, 孩子) and I(bear, verb-obj, child) are strong. So globally, the commonality between 怀" and "bear is high.

We performed comparison experiments on test-set 1, using Method C for triple translation. The results are shown in Table 4.

Method	**Accuracy**
No expansion	74.2%
Expansion by synonyms	69.8%
Expansion by triples	80.3%

Table 4: Results of translation expansion

	English triple	No expansion	Triple expansion
Positive examples	bear | child	忍受 | 孩子 (suffer from child)	怀 | 孩子 (bear child)
	break | silence	打碎 | 沉默 (smash silence)	打破 | 沉默 (break silence)
Negative examples	build | road	修建 | 公路 (build road)	制定 | 办法 (set up a method)
	make | sound	发出 | 声音 (make sound)	发表 | 讲话 (give a speech)

Figure 3: Example translations using triple model expansion

We found that synonym expansion does not bring any benefit because it introduces much noise along with some correct expansions. On the other hand, the triple model expansion achieves a substantial improvement. Its translations have a better coverage than the unexpanded method, and it also carries out a better expansion selection than the expansion by synonyms. Therefore, we integrate triple expansion in our triple translation method.

Unfortunately, it can also introduce wrong triples. Figure 3 shows some positive and negative examples. We found that the wrong expanded triple translations usually occur when the combinations of wrongly expanded translations are frequent ones in the corpus, thus cannot be filtered out using the triple translation model. The Chinese triples in the two negative examples of Figure 3 are very frequent ones. In these cases, even if the component words separately are strongly related to the original English words, their combination corresponds to a meaning different from the original one. Our current triple expansion is unable to deal with this problem.

5.2 CLIR Results

The two proposed query translation methods are tested on the TREC-9 Chinese corpus. 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 lexical resources we used include three human compiled bilingual lexicons (including the LDC English-Chinese dictionary) and a bilingual lexicon generated from a parallel bilingual corpus automatically. The resulting combined dictionary contains 401,477 English entries, including 109,841 words, and 291,636 phrases. The use of the combined dictionary is motivated by previous studies [9, 17], which showed that larger lexicon resource improves CLIR performance significantly.

The Okapi system with BM2500 weighting [15] is used as the basic retrieval system. The main evaluation metric is interpolated 11-point average precision. Statistical t-test and query-by-query analysis are also employed. To decide whether the improvement by method X over method Y is significant, the t-test calculates a p-value based on the performance data of X and Y. The smaller the p-value, the more significant is the improvement. Usually, if the p-value is small enough (p-value<0.05), we can conclude that the improvement is statistically significant.

The following methods are compared to investigate the effectiveness of our models for query translation:

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 [8, 15] 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[6].

The results of this series of experiments on query translation are shown in Table 5 and Figure 5. 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).

#	Methods	Avg. P.	% Mono. IR
1	Monolingual	0.2862	
2	Simple translation	0.1613	56%
3	Best-sense translation	0.2730	95%
4	2 + co-occurrence model	0.2392	84%
5	2 + triple model	0.1908	67%
6	5 + co-occurrence model	0.2517	88%

Table 5: Retrieval effectiveness on TREC-9 corpus

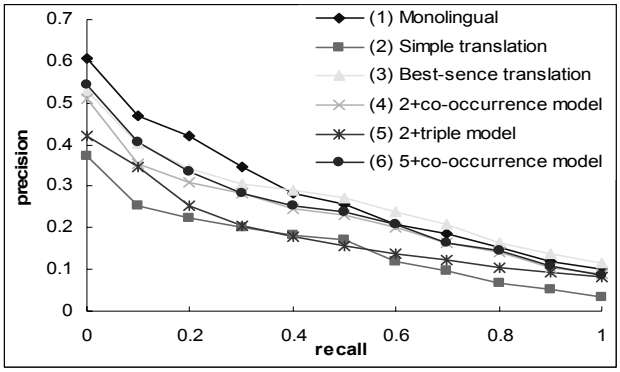

Figure 4: P-R curves (TREC-9 queries)

Row 6 corresponds to the preferred translation strategy. 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%. This performance is partly due to the process of triple expansion.

We also observe that the combination of the triple model with the simple translation method (raw 5) leads to an effectiveness well below that with the co-occurrence model (raw 4). This is expectable because triples have a much lower coverage than co-occurrences. Therefore, in the combination of raw 5, only a few triples from 11 queries out of 25 have been translated by the triple model, while all the other words are translated by the first translation word in the dictionary. So this "counter-performance" is not surprising. Figure 5 shows a closer view on the 11 queries.

From the 11 queries, NLPWIN extracted 52 triples which appear at least 5 times in the corpus. The minimal occurrences of 5 is set due to the fact that many low frequency triples are in fact noise. The 52 triples include 12 verb-object triples, 8 sub-verb triples, and 32 adj-noun triples, and no adv-verb triples. As shown in Figure 5, for these queries, the triple translation has positive impact on the co-occurrence method for almost all the 11 queries, except for #61. In the case of query #61, only one translation word differs: (coal) consumption→消耗 (by the co-occurrence model), and consumption→ 消费 (by the triple model). Both translations are correct, but the first translation word is often used for industrial consumption (which is the case for this query); whereas the second is often used for consumption of particular consumers. For all the 11 queries, globally, the triple method makes a statistically significant improvement of 56% over simple translation (p-value = 0.015), and 10% over the decaying co-occurrence model (p-value = 0.02).

A further analysis shows that the triple translation model is able to assemble translation words correctly in a triple because the triples can capture the syntactic dependency between not only sequential words (i.e. phrases) but also non-sequential words in a sentence as the (originate, sub, computer-virus) triple in query #64 shown in Figure 6. All the translations in Figure 6 are correct.

We can compare some of examples of Figure 6 with those by the decaying co-occurrence model (Method 5). For Queries #63 and #64, the co-occurrence model gives some different but correct translations for the following words: develop→发展 originate→ 发源. However, the word "hacker" in Query #65 is wrongly translated as 恶作剧者 (joker), and "charge" as 保护 (protect) by the co-occurrence method. For these cases, the triple translation method generates the correct translations. These examples show that the translations with triples can successfully correct some of the incorrect selections of translation words.

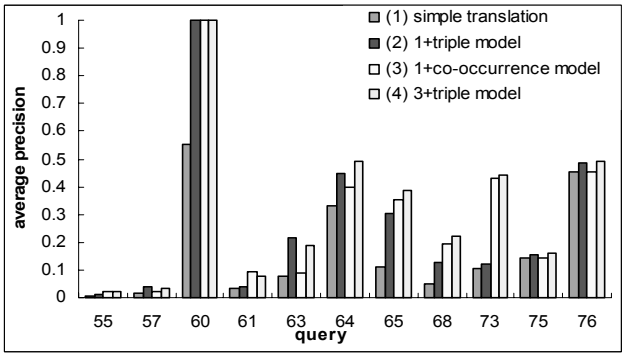

Figure 5: The queries containing triples

Q#	Sentence, the extracted triples and translations
63	What new or renewable energy sources are being developed in China?
	(energy, adj-noun, renewable)→ 可再生 | 能源 (energy, adj-noun, new)→ 新 | 能源 (develop, verb-obj, energy)→ 开发 | 能源
64	Have any computer viruses been discovered to have originated in Asia?

[6] This follows a suggestion by Douglas Oard.

	(discover, verb-obj, originate) →发现 | 起源 (originate, sub-verb, computer-virus)→计算机病毒 | 起源
65	Have any computer hackers been charged with crimes in Asia?
	(hacker, adj-noun, computer) →计算机 | 黑客 (charge, verb-obj, hacker) →控告 | 黑客

Figure 6: Some translations by the triple model

6. CONCLUSION

This paper proposed two statistical models for dealing with the problem of query translation ambiguity. We focused on the method that exploits linguistic dependency information represented as triples. The translations using triples showed three main benefits: a more precise translation; an extension of the coverage of the bilingual dictionary; and the possibility to train the model using unrelated bilingual corpora. Our experiments of CLIR showed that the triple translation has a positive impact on the query translation, and results in significant improvements of CLIR performance over the co-occurrence method. We also presented a revised version of the co-occurrence model. It differs from previous ones in that it includes a distance component that decays the mutual information between terms when the distance between them increases. Our experiments showed that the decaying co-occurrence model performs better than the standard co-occurrence model, and brings significant improvements over the simple dictionary approaches in CLIR.

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

ACKNOWLEDGEMENTS

The authors would like to thank Ashley Chang, and Chang-Ning Huang for their suggestions and comments on a preliminary draft of this paper. Thanks also to four anonymous reviews for valuable and insightful comments.

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] Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R.L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2): 263-311

[4] Bian, G.W. and Chen, H.H. (1998). Integrating query translation and document translation in a cross-language information retrieval system. *Machine Translation and Information Soup*, Lecture Notes in Computer Science, #1529, Spring-Verlag, pp. 250-265.

[5] Clarkson, P., and Robinson, A. (1997). Language model adaptation using mixtures and an exponentially decaying cache. In *Proceedings of ICASSP-97*, pp. 799--802.

[6] Davis, M. W., and Ogden, W. C. (1997). Free resources and advanced alignment for cross-language text retrieval. In: *TREC-6*, pp. 285-402.

[7] Fung, P., Liu, X., and Cheung, C. S. (1999). Mixed language query disambiguation. In *ACL-99*. The 37th Annual Meeting of the Association for Computational Linguistics, College Park, Maryland, USA, pp. 333-340.

[8] 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.

[9] 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..

[10] Hull, D. A., and Grefenstette, G. (1996). Querying across languages: a dictionary-based approach to multilingual information retrieval. In: *ACM SIGIR'96.* pp. 49-57.

[11] Jang, M.G., Myaeng, S. H., and Park S. Y. (1999). Using mutual information to resolve query translation ambiguities and query term weighting. In *ACL-99*. College Park, Maryland, pp. 223-229.

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

[13] Lin, D. (1998). Automatic retrieval and clustering of similar words. In *COLING-ACL98,* Montreal, Canada, August, pp. 768-774.

[14] Peters, C., and Picchi, E. (1996). Cross language information retrieval: A system for comparable corpus querying. In *SIGIR'96 Workshop on Cross-linguistic Information Retrieval*, pp. 24--33.

[15] Robertson, S. E., and Walker, S. (2000). Microsoft Cambridge at TREC-9: Filtering track. In *TREC-9*, pp. 361-368.

[16] Voorhees, E., Harman, D. (2001). Overview of the ninth text retrieval conference (TREC-9). In *TREC-9* pp. 1-14.

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

[18] 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.

[19] Mandala, R., Tokunaga, T., and Tanaka, H. (1999). Combining multiple evidence from different types of thesaurus for query expansion. In: *ACM SIGIR'99*. pp 191-197.