A Comparative Study for Query Translation using Linear Combination and Confidence Measure

Youssef Kadri
Laboratoire RALI, DIRO
Université de Montréal
CP 6128, Montréal, Canada, H3C3J7
kadriyou@iro.umontreal.ca

Jian-Yun Nie
Laboratoire RALI, DIRO
Université de Montréal
CP 6128, Montréal, Canada, H3C3J7
nie@iro.umontreal.ca

Abstract

In Cross Language Information Retrieval (CLIR), query terms can be translated to the document language using Bilingual Dictionaries (BDs) or Statistical Translation Models (STMs). Combining different translation resources can also be used to improve the performance. Unfortunately, the most studies on combining multiple resources use simple methods such as linear combination. In this paper, we drew up a comparative study between linear combination and confidence measures to combine multiple translation resources for the purpose of CLIR. We show that the linear combination method is unable to combine correctly different types of resources such as BDs and STMs. While the confidence measure method is able to re-weight the translation candidate more radically than in linear combination. It reconsiders each translation candidate proposed by different resources with respect to additional features. We tested the two methods on different test CLIR collections and the results show that the confidence measure outperforms the linear combination method.

1 Introduction

Cross Language Information Retrieval (CLIR) tries to determine documents written in a language from a query written in another language. Query translation is widely considered as the key problem in this task (Oard, 1998). In previous researches, various approaches have been proposed for query translation: using a bilingual dictionary, using an off-the-shelf machine translation system or using a parallel corpus. It is also found that when multiple translation resources are used, the translation quality can be improved, comparing to using only one translation resource (Xu, 2005). Indeed, every translation tool or resource has its own limitations. For example, a bilingual dictionary can suggest common translations, but they remain ambiguous – translations for different senses of the source word are mixed up. Machine translation systems usually employ sophisticated methods to determine the best translation sentence, for example, syntactic analysis and some semantic analysis. However, it usually output only one translation for a source word, while it is usually preferred that a source query word be translated by multiple words in order to produce a desired query expansion effect. In addition, the only word choice made by a machine translation system can be wrong. Finally, parallel corpora contain useful information about word translation in particular areas. One can use such a corpus to train a statistical translation model, which can then be used to translate a query. This approach has the advantage that few manual interventions are required to produce the statistical translation model. In addition, each source word can be translated by several related target words and the latter being weighted. However, among the proposed translation words, there may be irrelevant ones.

Therefore, one can take advantage of several translation resources and tools in order to produce better query translations. The key problem is the way to combine the resources.

A common method used in previous studies is to assign a weight to each resource. Then all the translation candidates are weighted and then combined linearly (Nie, 2000). However, this kind of combination assigns a single confidence score to all the translations from the same translation resource. In reality, a translation resource does not cover all the words with equal confidence. For some words, its translations can be accurate, while for some others, they are inappropriate. By using a linear combination, the relative order among the translation candidates is not changed. In practice, a translation with a low score can turn out to be a better translation when other information becomes available.

For example, the English word "nutritional" is translated into French by a statistical translation model trained on a set of parallel texts as follows:

{nutritive 0.32 (nutritious), alimentaire 0.21 (food)}.

We observe that the most common translation word "alimentaire" only takes the second place with lower probability than "nutritive". If these translations are combined linearly with another resource (say a BD), it is unlikely that the correct translation word "alimentaire" gain larger weight than "nutritive".

This example shows that we have to reconsider the relative weights of the translation candidates when another translation resource is available. The purpose of this reconsideration is to determine how reasonable a translation candidate is given all the information now available. In so doing, the initial ranking of translation candidates can be changed. As a matter of fact using the method of confidence measures that we propose in this paper, we are able to reorder the translation candidates as follows:

{alimentaire 0.38, nutritive 0.23, valeur 0.11 (value)}.

The weight of the correct translation "alimentaire" is considerably increased.

In this paper, we will propose to use a new method based on confidence measure to re-weight the translation candidates. In the re-weighting, the original weight according to each translation resource is only considered as one factor. The final weight is determined by combining all the available factors. In our implementation, the factors are combined in neural networks, which produce a final confidence measure for each of the translation candidates. This final weight is not a simple linear combination of the original weights, but a recalculation according to all the information available, which is not when each translation resource is estimated separately.

The advantages of this approach are twofold. On one hand, the confidence measure allows us to adjust the original weights of the translations and to select the best translation terms according to all the information. On the other hand, the confidence measures also provide us with a new weighting for the translation candidates that are comparable across different translation resources. Indeed, when we try to combine a statistical translation model with a bilingual dictionary, we had to assign a weight to a candidate from the bilingual dictionary. This weight is not directly compatible with the probability assigned in the former.

In the remaining sections of this paper, we will first describe the principle of confidence measure in section 2. In section 3, we will compare two methods to combine different translation resources: linear combination and confidence measure. Section 4 provides a description on how the parameters are tuned. Section 5 outlines the different steps for computing confidence measures. Finally, we present the results of our experiments on both English-French and English-Arabic CLIR. Our experiments will show that the method using confidence measure significantly outperforms the traditional approach using linear combination.

2 Confidence measure

Confidence measure is often used to re-rank or re-weight some outputs produced by separate means. For example, in speech recognition and understanding (Hazen et al., 2002), one tries to re-rank the result of speech recognition according to additional information using confidence measure. Gandrabur et al. (2003) used confidence measures in a translation prediction task. The goal is to re-rank the translation candidates according to additional information. Confidence measure is defined as the probability of correctness of a candidate. In the case of translation, given a candidate translation t_E for a source word t_F, the confidence measure is $P(correct \mid t_F, t_E, F)$, where F is a set of other features of the translation context (e.g. the POS-tag of the word, the previous translations words, etc.). In both applications, significant gains have been observed when using a confidence estimation layer within the translation models.

The problem of query translation is similar to general translation described in (Gandrabur et al. 2003). We are presented with several translation resources, each being built separately. Our goal now is to use all of them together. As we discussed earlier, we want to take advantage of the additional information (other translation resources as well as additional linguistic analysis on the query) in order to re-weight each of the translation candidates.

In previous studies, neural networks have been commonly used to produce confidence measures. The inputs to the neural networks are translation candidates from different resources, their original weights and various other properties of them (e.g. POS-tag, probability in a language model, etc.). The output of the neural networks is a confidence measure assigned to a translation candidate from a translation resource. This confidence measure is used to re-rank the whole set of candidates from all the resources.

In this study, we will use the same approach to combine different translation resources and to produce confidence measures.

The neural networks need to be trained on a set of training data. Such data are available in both speech recognition and machine translation. However, in the case of CLIR, the goal of query translation is not strictly equivalent to machine translation. Indeed, in query translation, we are not limited to the correct literal translations. Not literal translation words that are strongly related to the query are also highly useful. These latter related words can produce a desired query expansion effect in IR.

Given this situation, we can no longer use a parallel corpus as our training data as in the case of machine translation. Modifications are necessary. We will describe the modified way we use to create the training data in section 4. The informative features we use will be described n section 5.2.

3 General CLIR Problem

Assume a query Q_E written in a source language E and a document D_F written in a target language F, we would like to determine a score of relevance of D_F to Q_E. However, as they are not directly comparable, a form of translation is needed. Let us describe the model that we will use to determine its score.

Various theoretical models have been developed for IR, including vector space model, Boolean model and probabilistic model. Recently, language modeling is widely used in IR, and it has been show to produce very good experimental results. In addition, language modeling also provides a solid theoretical framework for integrating more aspects in IR such as query translation. Therefore, we will use it as our basic framework in this study.

In language modeling framework, the relevance score of the document D_F to the query Q_E is determined as the negative KL-divergence between the query's language model and the document's language model (Zhai, 2001a). It is defined as follows:

$$R(Q_E, D_F) \propto \sum_{t_F} p(t_F \mid Q_E) \log p(t_F \mid D_F) \qquad (1)$$

To avoid the problem of attributing zero probability to query terms not occurring in document D_F, smoothing techniques are used to estimate $p(t_F|D_F)$. One can use the Jelinek-Mercer smoothing technique which is a method of interpolating between the document and collection language models (Zhai, 2001b). The smoothed $p(t_F|D_F)$ is calculated as follows:

$$p(t_F \mid D_F) = (1-\lambda) p_{ML}(t_F \mid D_F) + \lambda p_{ML}(t_F \mid C_F) \qquad (2)$$

where $p_{ML}(t_F \mid D_F) = \frac{tf(t_F, D_F)}{|D_F|}$ and $p_{ML}(t_F \mid C_F) = \frac{tf(t_F, C_F)}{|C_F|}$ are the maximum likelihood estimates of a unigram language model based on respectively the given document D_F and the collection of documents C_F. λ is a parameter that controls the influence of each model.

In CLIR, the term $p(t_F \mid Q_E)$ in equation (1) representing the query model can be estimated as follows:

$$p(t_F \mid Q_E) = \sum_{q_E} p(t_F, q_E \mid Q_E) = \sum_{q_E} p(t_F \mid q_E, Q_E) p(q_E \mid Q_E)$$
$$\approx \sum_{q_E} p(t_F \mid q_E) p_{ML}(q_E \mid Q_E) \qquad (3)$$

where $p_{ML}(q_E \mid Q_E)$ is the maximum likelihood estimation: $p_{ML}(q_E \mid Q_E) = \frac{tf(q_E, Q_E)}{|Q_E|}$ and $p(t_F \mid q_E)$ is the translation model. Putting (3) in (1), we obtain the general CLIR score formula:

$$R(Q_E, D_F) \propto \sum_{t_F} \sum_{q_E} p(t_F \mid q_E) p_{ML}(q_E \mid Q_E) \log p(t_F \mid D_F) \quad (4)$$

In our work, we do not change the document model $p(t_F \mid D_F)$ from monolingual IR. Our focus will be put on the estimation of the translation model $p(t_F \mid q_E)$ - the translation probability from a source query term q_E to a target word t_F, in particular, when several translation resources are available.

Let us now describe two different ways to combine different translation resources for the estimation of $p(t_F \mid q_E)$: by linear combination and by confidence measure.

4 Linear Combination

The first intuitive method to combine different translation resources is by a linear combination. This means that the final translation model is estimated as follows:

$$p(t_F \mid q_E) = z_{q_E} \sum_i \lambda_i p_i(t_F \mid q_E) \qquad (5)$$

where λ_i is the parameter assigned to the translation resource *i* and z_{q_E} is a normalization factor so that $\sum_{t_F} p(t_F \mid q_E) = 1$. $p_i(t_F \mid q_E)$ is the probability of translating the source word q_E to the target word t_F by the resource *i*.

In order to determine the appropriate parameter for each translation resource, we use the EM algorithm to find values which maximize the log-likelihood LL of a set *C* of training data according to the combined model, i.e.:

$$LL(C) = \sum_{(f,e)\in C} p(f,e) \sum_{j=1}^{|f|} \log \sum_{k=1}^{n} \sum_{i=1}^{|e|} \lambda_k t_k(f_j \mid e_i) p(e_i) \qquad (6)$$

Where $(f, e) \in C$ is a pair of parallel sentences; $p(f,e) = \frac{\#(f,e)}{|C|}$ is the prior probability of the pair of sentences (*f, e*) in the corpus *C*, |*f*| is the length of the target sentence *f* and |*e*| is the length of the source sentence *e*. λ_k is the coefficient related to resource *k* that we want to optimize and *n* is the number of resources. $t_k(f_j|e_i)$ is the probability of translating the source word e_i with the target word f_j with each resource. $p(e_i)$ is the prior probability of the source word e_i in the corpus *C*. Note that the validation data set *C* on which we optimize the parameters must be different from the one used to train our baseline models.

The training corpora are as follows: For English-Arabic, we use the Arabic-English parallel news corpus[1]. This corpus consists of around 83 K pairs of aligned sentences. For English-French, we use a bitext extracted from two parallel corpora: The Hansard[2] corpus and the Web corpus (Kadri, 2004). It consists of around 60 K pairs of aligned sentences.

The component models for English-Arabic CLIR are: a STM built on a set of parallel Web pages (Kadri, 2004), another STM built on the English-Arabic United Nations corpus (Fraser, 2002), Ajeeb[3] bilingual dictionary and Almisbar[4] bilingual dictionary. For English-French CLIR, we use three component models: a STM built on Hansard corpus, another STM built on parallel Web pages and the Freedict[5] bilingual dictionary.

5 Using Confidence Measures

The question considered in confidence measure is: Given a translation *candidate*, is it correct and how confident are we on its correctness?

Confidence measure aims to answer this question. Given a translation candidate t_F for a source term q_E and a set *F* of other features, confidence measure corresponds to $p_i(C=1 \mid t_F, q_E, F)$. We can use this measure as an estimate of $p(t_F \mid q_E)$, i.e.:

$$p(t_F \mid q_E) = z_{q_E} \sum_i p_i(C=1 \mid t_F, q_E, F) \qquad (7)$$

where *F* is the set of features that we use. We will see several features to help determine the confidence measure of a translation candidate, for example, the translation probability, the reverse translation probability, language model features, and so on. We will describe these features in more detail in section 5.2.

In general, we can consider confidence measure as *P(C=1|X)*, given *X*— the source word, a translation and a set of features. We use a Multi Layer Perceptron (MLP) to estimate the probability of correctness *P(C=1|X)* of a translation. Neural networks have the ability to use input data of different natures and they are well-suited for classification tasks.

Our training data can be viewed as a set of pairs (*X,C*), where *X* is a vector of features relative to a translation[6] used as the input of the network, and *C* is the desired output (the correctness of the translation 0/1). The MLP implements a non-linear mapping of the input features by combining layers of linear transformation and non-linear transfer function. Formally, the MLP implements a discriminant function for an input X of the form:

[1] http://www.ldc.upenn.edu/
Arabic-English Parallel News Part 1 (LDC2004T18)

[2] LDC provides a version of this corpus: http://www.ldc.upenn.edu/.

[3] http://www.ajeeb.com/

[4] http://www.almisbar.com/

[5] http://www.freedict.com/

[6] By translation, we mean the pair of source word and its translation.

$$g(X;\theta) = o(V \times h(W \times X)) \qquad (8)$$

where θ ={W,V}, W is a matrix of weights between input and hidden layers and V is a vector of weights between hidden and output layers; *h* is an activation function for the hidden units which non-linearly transforms the linear combination of inputs $W \times X$; *o* is also a non-linear activation function but for the output unit, that transforms the MLP output to the probability estimate *P*(*C=1*|*X*). Under these conditions, our MLP was trained to minimize an objective function of error rate (Section 4.1).

In our experiments, we used a batch gradient descent optimizer. During the test stage, the confidence of a translation X is estimated with the above discriminant function *g*(*X; θ*); where *θ* is the set of weights optimized during the learning stage. These parameters are expected to correlate with the true probability of correctness *P*(*C=1*|*X*).

5.1 The objective function to minimize

A natural metric for evaluating probability estimates is the negative log-likelihood (or cross entropy CE) assigned to the test corpus by the model normalized by the number of examples in the test corpus (Blatz et al., 2003). This metric evaluates the probabilities of correctness. It measures the cross entropy between the empirical distribution on the two classes (correct/incorrect) and the confidence model distribution across all the examples $X^{(i)}$ in the corpus. Cross entropy is defined as follows:

$$CE = -\tfrac{1}{n}\sum_i \log P(C^{(i)} \mid X^{(i)}) \qquad (9)$$

where $C^{(i)}$ is 1 if the translation $X^{(i)}$ is correct, 0 otherwise. To remove dependence on the prior probability of correctness, Normalized Cross Entropy (NCE) is used:

$$NCE = (CE_b - CE)/CE_b \qquad (10)$$

The baseline CE_b is a model that assigns fixed probabilities of correctness based on the empirical class frequencies:

$$CE_b = -(n_0/n)\log(n_0/n) - (n_1/n)\log(n_1/n) \qquad (11)$$

where n_0 and n_1 are the numbers of correct and incorrect translations among *n* test cases.

5.2 Features

The MLP tends to capture the relationship between the correctness of the translation and the features, and its performance depends on the selection of informative features.

We selected intuitively seven classes of features hypothesized to be informative for the correctness of a translation.

Translation model index: an index representing the resource of translation that produced the translation candidate.

Translation probabilities: the probability of translating a source word with a target word. These probabilities are estimated with IBM model 1 (Brown et al., 1993) on parallel corpora. For translations from bilingual dictionaries, as no probability is provided, we carry out the following process to assign a probability to each translation pair (*e, f*) in a bilingual dictionary: We trained a statistical translation model on a parallel corpus. Then for each translation pair (*e,f*) of the bilingual dictionary, we looked up the resulting translation model and extracted the probability assigned by this translation model to the translation pair in question. Finally, the probability is normalized by the Laplace smoothing method:

$$p_{BD}(f \mid e) = \frac{p_{STM}(f \mid e) + 1}{\sum_{i=1}^{n} p_{STM}(f_i \mid e) + 1} \qquad (12)$$

Where *n* is the number of translations proposed by the bilingual dictionary to the word *e*.

Translation ranking: This class of features includes two features: The rank of the translation provided by each resource and the probability difference between the translation and the highest probability translation.

Reverse translation information: This includes the probability of translation of a target word to a source word. Other features measure the rank of source word in the list of translations of the target word and if the source word holds in the best translations of the target word.

Translation "Voting": This feature aims to know whether the translation is voted by more than one resource. The more a same translation is voted the more likely it may be correct.

Source sentence-related features: One feature measures the frequency of the source word in the source sentence. Another feature measures the number of source words in the source sentence that have a translation relation with the translation in question.

Language model features: We use the unigram, the bigram and the trigram language models for source and target words on the training data.

5.3 Training for confidence measures

The corpus used for training confidence is the same as the corpus for tuning parameters for the linear combination. It is a set of aligned sentences. Source sentences are translated to the target language word by word using baseline models. We translated each source word with the most probable[7] translations for the translation models and the best five translations provided by the bilingual dictionaries. Translations are then compared to the reference sentence to build a labeled corpus: a translation of a source word is considered to be correct if it occurs in the reference sentence. The word order is ignored, but the number of occurrences is taken into account. This metric fits well our context of IR: IR models are based on "bag of words" principle and the order of words is not considered.

We test with various numbers of hidden units (from 5 to 100). We used the NCE metric to compare the performance of different architectures. The MLP with 50 hidden units gave the best performance.

To test the performance of individual features, we experimented with each class of features alone. The best features are the translation "voting", language model features and the translation probabilities. The translation "voting" is very informative because it presents the translation probability attributed by each resource to the translation in question. The translation ranking, the reverse translation information, the translation model index and the source sentence-related features provide some marginally useful information.

6 CLIR experiments

The experiments are designed to test whether the confidence measure approach is effective for query translation, and how it compares with the traditional linear combination. We will conduct two series of experiments, one for English-French CLIR and another for English-Arabic CLIR.

6.1 Experimental setup

English-French CLIR: We use English queries to retrieve French documents. In our experiments, we use two document collections: one from TREC[8] and another from CLEF[9] (SDA). Both collections contain newspaper articles. TREC collection contains 141 656 documents and CLEF collection 44 013 documents. We use 4 query sets: 3 from TREC (TREC6 (25 queries), TREC7 (28 queries), TREC8 (28 queries)) and one from CLEF (40 queries).

English-Arabic CLIR: For these experiments, we use English queries to retrieve Arabic documents. The test corpus is the Arabic TREC collection which contains 383 872 documents. For topics, we use two sets: TREC2001 (25 queries) and TREC2002 (50 queries).

Documents and queries are stemmed and stopwords are removed. The Porter stemming is used to stem English queries and French documents. Arabic documents are stemmed using linguistic-based stemming method (Kadri, 2006). The query terms are translated with the baseline models (Section 4). The resulting translations are then submitted to the information retrieval process. We tested with different ways to assign weights to translation candidates: translations from each resource, linear combination and confidence measures.

When using each resource separately, we attribute the IBM 1 translation probabilities to our translations. For each query term, we take only translations with the probability $p(f|e) \geq 0.1$ when using translation models and the five best translations when using bilingual dictionaries.

6.2 Linear combination (LC)

The tuned parameters assigned to each translation resource are as follows:

English-Arabic CLR:
STM-Web: 0.29, STM-UN: 0.34,
Ajeeb BD: 0.14, Almisbar BD: 0.22.
English-French CLR:
STM-Web: 0.3588, STM-Hansard: 0.6408,
Freedict BD: 0.0003.

These weights produced the best log-likelihood of the training data.

[7] The translations with the probability $p(f|e) \geq 0.1$

[8] http://trec.nist.gov/

[9] http://www.clef-campaign.org/

For CLIR, the above combinations are used to combine translation candidates from different resources. The tables below show the CLIR effectiveness (mean average precision - MAP) of individual models and the linear combination.

Translation Model	TREC 2001	TREC 2002	Merged TREC 2001/2002
Monolingual IR	(0.33)	(0.28)	(0.31)
STM-Web	0.14 (42%)	0.04 (17%)	0.07 (25%)
STM-UN	0.11 (33%)	0.09 (34%)	0.10 (33%)
Ajeeb BD	0.27 (81%)	0.19 (70%)	0.22 (70%)
Almisbar BD	0.17 (51%)	0.16 (58%)	0.16 (54%)
Linear Comb.	0.24 (72%)	0.20 (71%)	0.21 (67%)

Table1. English-Arabic CLIR performance (MAP) with individual models and linear combination

Trans. Model	TREC6	TREC7	TREC8	CLEF
Monolingual IR	0.39	0.34	0.44	0.40
STM-Web	0.22 (56%)	0.17 (50%)	0.22 (50%)	0.29 (72%)
STM-Hansard	0.25 (64%)	0.24 (70%)	0.33 (75%)	0.30 (75%)
Freedict BD	0.17 (43%)	0.11 (32%)	0.13 (29%)	0.14 (35%)
Linear Comb.	0.26 (66%)	0.26 (76%)	0.36 (81%)	0.30 (75%)

Table2. English-French CLIR performance (MAP) with individual models and linear combination

We observe that the performance is quite different from one model to another. The low score recorded by the STMs for English-Arabic CLIR compared to the score of STMs for English-French CLIR is possibly due to the small data set on which the English-Arabic STMs are trained. A set of 2816 English-Arabic pairs of documents is not enough to build a reasonable STM. For English-Arabic CLIR, BDs present better performance than STMs because they cover almost all query terms and they provide multiple good translations to each query term. When combining all the resources, the performance is supposed to be better because we would like to take advantage of each of the models. However, we see that the combined model performs even worse than one of the models - Ajeeb BD for English-Arabic CLIR. This shows that the linear combination is not necessarily a good way to combine different translation resources.

An example of English queries is shown in Table 3: “What measures are being taken to develop tourism in Cairo?”. The Arabic translation provided by TREC to the word “measures” is: “إجراءات”. We see clearly that translations with different resources are different. Some resources propose inappropriate translations such as “مكيال” or “ميزان”. Even if two resources suggest the same translations, the weights are different. For this query, the linear combination produces better query translation terms than every resource taken alone: The most probable translations are selected from the combined list. However, this method is unable to attribute an appropriate weight to the best translation “إجراءات”; it is selected but ranked at third position with a weak weight.

Trans. model	Translation(s) of word “measures”
Ajeeb BD	تدبير 0.05 (measure), عيار 0.05 (caliber), قياس 0.05 (measurement), مقياس 0.05 (measurement), معيار 0.05 (standard), مكيال 0.05 (standard), ميزان 0.05 (balance)
Almisbar BD	إجراءات 0.05 (procedures), مقياس 0.03, قدر 0.03 (measurement), مقدار 0.03 (amount)
STM-UN	تدابير 0.69 (measures)
STM-Web	إجراءات 0.09
Linear Comb,	تدابير 0.61, مقياس 0.037, إجراءات 0.029, قياس 0.020

Table3. Translation examples

6.3 CLIR with Confidence Measures (CM)

In these experiments, we use confidence measures as weights for translations. According to these confidence measures, we select the translations with the best confidences for each query term. The following tables show the results:

Collection	TREC 2001	TREC 2002	TREC01-02
MAP of LC	0.2426	0.2032	0.2163
MAP of CM	0.2775(14.35%)	0.2052 (1%)	0.2290 (5.87 %)

Table4. Comparison of English-Arabic CLIR between linear combination and confidence measures

Collection	TREC6	TREC7	TREC8	CLEF
MAP of LC	0.2692	0.2630	0.3605	0.3071
MAP of CM	0.2988 (10.99%)	0.2699 (2.62%)	0.3761 (4.32%)	0.3230 (5.17 %)

Table5. Comparison of English-French CLIR between linear combination and confidence measures

In terms of MAP, we see clearly that the results using confidence measures are better than those obtained with the linear combination. The two-tailed t-test shows that the improvement brought by confidence measure over linear combination is statistically significant at the level *P*<0.05. This improvement in CLIR performance is attributed to the ability of confidence measure to re-weight each translation candidate. The final sets of translations (and their probabilities) are more reasonable than in linear combination. The tables below show some examples where we get a large improvement in average precision when using confidence measures to combine resources. The first example is the TREC 2001 query “What measures are being taken to develop tourism in Cairo?”. The translation of the query term “measures” to Arabic using the two methods is presented in table 6. The second example is the TREC6 query "Acupuncture". Table 7 presents the translation of this query term is to French using the two techniques:

Trans.Model	Translation(s) of term "measures"
Linear Comb.	تدابير 0.61, مقياس 0.037, إجراءات 0.029, قياس 0.020
Conf. meas.	إجراءات 0.51, قدر 0.10, قياس 0.06

Table6. Translation examples to Arabic

Trans.model	Translation(s) of term "Acupuncture"
Linear Comb.	Acupuncture 0.13 (acupuncture), sevrage 0.13 (severing), hypnose 0.13 (hypnosis)
Conf. meas.	Acupuncture 0.21, sevrage 0.17, hypnose 0.14

Table7. Translation examples to French

In the example of table 6, confidence measure has been able to redeem the best translation "إجراءات" and rescore it with a stronger weight than the other incorrect or inappropriate ones. The same effect is observed in the example of table 7. Confidence measure has been able to increase the correct translation "acupuncture" to a higher level than the other incorrect ones. These examples show the potential advantage of confidence measure over linear combination: The confidence measure does not blindly trust all the translations from different resources. It tests their validity on new validation data. Thus, the translation candidates are rescored and filtered according to a more reliable weight.

7 Conclusion

Multiple translation resources are believed to contribute in improving the quality of query translation. However, in most previous studies, only linear combination has been used. In this study, we propose a new method based on confidence measure to combine different translation resources. The confidence measure estimates the probability of correctness of a translation, given a set of features available. The measure is used to weight the translation candidates in a unified manner. It is also expected that the new measure is more reasonable than the original measures because of the use of additional features. Our experiments on both English-Arabic and English-French CLIR have shown that confidence measure is a better way to combine translation resources than linear combination. This shows that confidence measure is a promising approach to combine non homogenous resources and can be further improved on several aspects. For example, we can optimize this technique by identifying other informative features. Other techniques for computing confidence estimates can also be used in order to improve the performance of CLIR.

References

J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis and N. Ueffing. 2003. *Confidence estimation for machine translation*. Technical Report, CLSP/JHU 2003 Summer Workshop, Baltimore MD.

P. F. Brown, S. A. Pietra, V. J. Pietra and R. L. Mercer. 1993. *The mathematics of statistical machine translation: Parameter estimation*. Computational Linguistics, 19(2):263–311.

A. Fraser, J. Xu and R. Weischedel. 2002. *TREC 2002 Cross-lingual Retrieval at BBN*. TREC11 conference.

S. Gandrabur and G. Foster. 2003. *Confidence Estimation for Text Prediction*. Proceedings of the CoNLL 2003 Conference, Edmonton.

T. J. Hazen, T. Burianek, J. Polifroni and S. Seneff. 2002. *Recognition confidence scoring for use in speech understanding systems*. Computer Speech and Language, 16:49-67.

Y. Kadri and J. Y. Nie. 2004. *Query translation for English-Arabic cross language information retrieval*. Proceedings of the TALN conference.

Y. Kadri and J. Y. Nie. 2006. *Effective stemming for Arabic information retrieval*. The challenge of Arabic for NLP/MT Conference. The British Computer Society. London, UK.

J. Y. Nie, M. Simard and G Foster. 2000. *Multilingual information retrieval based on parallel texts from the Web*. In LNCS 2069, C. Peters editor, CLEF2000:188-201, Lisbon.

D. W. Oard and A. Diekema. 1998. *Cross-Language Information Retrieval*. In M. Williams (ed.), Annual review of Information science, 1998:223-256.

J. Xu and R. Weischedel. 2005. *Empirical studies on the impact of lexical resources on CLIR performance*. Information processing & management, 41(3):475-487.

C. Zhai and J. Lafferty. 2001a. *Model-based feedback in the language modeling approach to information retrieval*. CIKM 2001 Conference.

C. Zhai and J. Lafferty. 2001b. *A study of smoothing methods for language models applied to ad hoc information retrieval*. Proceedings of the ACM-SIGIR.