Computational Linguistics and Chinese Language Processing
Vol. 13, No. 1, March 2008, pp. 1-26

Exploring Shallow Answer Ranking Features in Cross-Lingual and Monolingual Factoid Question Answering

Cheng-Wei Lee[*+], Yi-Hsun Lee[+], and Wen-Lian Hsu[+]

Abstract

Answer ranking is critical to a QA (Question Answering) system because it determines the final system performance. In this paper, we explore the behavior of shallow ranking features under different conditions. The features are easy to implement and are also suitable when complex NLP techniques or resources are not available for monolingual or cross-lingual tasks. We analyze six shallow ranking features, namely, *SCO-QAT*, *keyword overlap*, *density*, *IR score*, *mutual information score*, and *answer frequency*. SCO-QAT (Sum of Co-occurrence of Question and Answer Terms) is a new feature proposed by us that performed well in NTCIR CLQA. It is a co-occurrence based feature that does not need extra knowledge, word-ignoring heuristic rules, or special tools. Instead, for the whole corpus, SCO-QAT calculates co-occurrence scores based solely on the passage retrieval results. Our experiments show that there is no perfect shallow ranking feature for every condition. SCO-QAT performs the best in C-C (Chinese-Chinese) QA, but it is not a good choice in E-C (English-Chinese) QA. Overall, Frequency is the best choice for E-C QA, but its performance is impaired when translation noise is present. We also found that passage depth has little impact on shallow ranking features, and that a proper answer filter with fined-grained answer types is important for E-C QA. We measured the performance of answer ranking in terms of a newly proposed metric EAA (Expected Answer Accuracy) to cope with cases of answers that have the same score after ranking.

[*] Department of Computer Science, National Tsing-Hua University, Taiwan, R.O.C, 101, Section 2, Kuang-Fu Road, Hsinchu, Taiwan, R.O.C.
[+] Institute of Information Science, Academia Sinica, Taiwan, R.O.C, 128 Academia Road, Section 2, Nankang, Taipei 115, Taiwan, R.O.C.
The author for correspondence is Wen-Lian Hsu.
E-mail: {aska, rog, hsu}@iis.sinica.edu.tw

[Received June 30, 2007; Revised October 8, 2007; Accepted January 12, 2008]

Keywords: Answer Ranking, Co-occurrence, CLQA, Question Answering, Shallow Method, SCO-QAT

1. Introduction

In recent years, question answering (QA) has become a key research area in several of the world's major languages, possibly because of the urgent need to deal with the information overload caused by the rapid growth of the Internet. Since 1999, many international question answering contests have been held at conferences and workshops, such as TREC[1], CLEF[2], and NTCIR[3]. Thus far, several languages – such as Bulgarian, Dutch, English, Finnish, French, German, Indonesian, Italian, Japanese, Portuguese, and Spanish – have been tested in monolingual or cross-lingual question answering tasks. In QA research, questions are usually classified into several categories, such as factoid questions, list questions, and definition questions, then dealt with by different techniques. Among these categories, factoid questions have been studied the most widely, and they are the focus of this paper.

There is usually exactly one answer, which is a noun or short phrase, for a factoid question. For example, "Who is the president of the United States?" is a factoid question because the name of the president is a noun, and there is only one current U.S. President. Factoid questions are usually classified into questions types, such as Q_PERSON, Q_LOCATION, Q_ORGANIZATION, Q_ARTIFACT, Q_TIME, and Q_NUMBER [Lee *et al*. 2007; Lee *et al*. 2005]. Although question types vary in different contests and different systems, the corresponding answer types can usually be recognized by named entity recognition (NER) techniques or simple rules.

A QA system is normally comprised of several modules. The answer ranking module implements the last step in answering a factoid question and determines the final performance. After candidate answers have been extracted from retrieved passages, the answer ranking module takes the question, the passages (or documents), and the candidate answers as input, ranks the candidate answers, and then outputs a ranked list of candidate answers. Although several answer ranking methods have been proposed, they can be generally categorized as either deep or shallow methods. A deep method uses complex NLP techniques and may require extensive rules, ontologies, or human effort, while a shallow method does not require much of these resources and is therefore cheaper to implement.

Although deep answer ranking methods have proven useful for English QA, as reported in [Cui *et al*. 2005; Harabagiu *et al*. 2005], the resources needed for such methods are usually

---

[1] Text REtrieval Conference (TREC). http://trec.nist.gov/

[2] Cross-Language Evaluation Forum (CLEF). http://www.clef-campaign.org/

[3] NTCIR (NII Test Collection for IR Systems) Project. http://research.nii.ac.jp/ntcir/

not available for some languages in monolingual or cross-lingual QA. In those cases, shallow ranking methods have to be used; however, to the best of our knowledge, very little research has been done on such methods. The situation is worse for cross-lingual tasks because most cross-lingual QA research has focused on the front-end modules, *i.e*., question processing and passage retrieval. Research on back-end modules, such as answer ranking, has received little attention in the cross-lingual QA domain.

In this paper, we attempt to fill this research gap by exploring the behavior of shallow ranking features under noise produced by other QA modules in both monolingual and cross-lingual situations. Herein, noise is defined in terms of the performance decrement of a QA module. For example, in the case of translation quality decrement, we say that we encounter translation noise and expect that the noise may impact the performance of some shallow ranking features. In addition to translation noise, we also consider passage retrieval noise and answer filter noise. We measure the influence of these types of noise by three performance metrics to determine which ranking feature is the most effective in dealing with each kind of noise.

Apart from considering widely used shallow ranking features, we propose a new ranking feature called SCO-QAT, which has been successfully applied to the ASQA2 system [Lee *et al*. 2007], and also achieved the best performance on the C-C and E-C subtasks in NTCIR-6 CLQA [Sasaki *et al*. 2007]. SCO-QAT is a co-occurrence based feature; however, unlike some co-occurrence features [Magnini *et al*. 2001], it does not need extra knowledge, word-ignoring heuristic rules, or special tools.

The remainder of this paper is organized as follows. Related works are discussed in Section 2. We introduce the SCO-QAT feature in Section 3. The evaluation metrics used are introduced in Section 4. The ASQA2 system used in our experiments is described in Section 5. We detail our experiment results and compare SCO-QAT with other shallow features in Section 6. Then, we present our conclusions in Section 7.

2. Related Work

Answer Ranking approaches can be divided in to deep and shallow methods. Deep approaches involve sophisticated tools or knowledge. The most advanced deep methods are logic-based and dependency-parser-based. The LCC team [Harabagiu *et al*. 2005] used an abductive inference method to evaluate the correctness of an answer according to the logic form of the question, the logic form of the sentence that supports the answer, and background knowledge from WordNet. The logic-based approach has achieved the best QA performance in TREC for several years.

Dependency-parser-based methods have also performed quite well on TREC tasks. The National University of Singapore team [Cui *et al*. 2005] used dependency relations identified by a dependency parser to select answer nuggets for factoid and list questions. The similarity between the question and the supporting passage is calculated by machine translation models. Shen [Shen *et al*. 2006] also used dependency relations, but incorporated them into a Maximum Entropy-based ranking model.

Although these deep approaches perform well on monolingual QA (about 0.7 accuracy), they are quite demanding in terms of linguistic resources and computational complexity. In cross-lingual or multilingual QA, it is usually impossible to employ deep approaches for some languages due to the lack of knowledge resources or tools. In contrast, approaches with shallow features are much more flexible when QA languages are changed. The following are some commonly used shallow approaches.

Surface patterns [Soubbotin and Soubbotin 2001] have been successful in the TREC QA Track, which uses string patterns to match questions with correct answers. However, from our perspective, if surface patterns are manually created, the method can not be regarded as "shallow", because it is likely labor intensive. Although there are some "shallow" variations [Geleijnse and Korst 2006; Ravichandran and Hovy 2002] that attempt to create surface patterns automatically/semi-automatically, they usually suffer from the low coverage problem, which means they can only be applied to a few questions.

Some approaches focus on local information, thus only take the *similarity* between a passage and the question into account when finding relevant answers. The simplest way to measure the similarity is by counting the ratio of question terms occurring in the answer passage, as has been reported [Cooper and Ruger 2000; Molla and Gardiner 2005; Zhao *et al*. 2005]. Kwok [Kwok and Deng 2006] and AnswerBus [Zheng 2002] adopt the IR score of the answer passage directly as a measure of similarity. Intuitively, the closeness of two terms may indicate a relation; therefore, some systems [Gillard *et al*. 2006; Lin *et al*. 2005; Lin *et al*. 2005; Sacaleanu and Neumann 2006; Tom´as *et al*. 2005] use features based on the distance between the answer and the question terms to obtain a better similarity measurement. Among these approaches, those of Lin *et al*. [Lin *et al*. 2005] and Roussinov *et al*. [Roussinov *et al*. 2004] incorporate the IDF value with term distances. The assumption is that, if the candidate answer is close to several keywords or question terms, it is more likely to be relevant.

Instead of utilizing local information, which only considers the question and a passage, *redundancy-based* features consider all the returned passages or the entire corpus. Clarke [Clarke *et al*. 2001] suggested that redundancy could be used as a substitute for deep analysis because correct answers may appear many times in high-ranking passages. Features using frequency or co-occurrence information are all regarded as redundancy-based. Several systems [Clarke *et al*. 2002; Cooper and Ruger 2000; Kwok and Deng 2006; Lin *et al*. 2005; Zhao *et al*.

2005; Zheng 2002] include answer frequency in their Answer Ranking components. A web-based co-occurrence shallow feature developed by Magnini *et al*. [Magnini *et al*. 2001] has been successfully applied on the TREC dataset. Magnini used three methods, *Pointwise Mutual Information*, *Maximal Likelihood Ratio*, and *Corrected Conditional Probability*, to measure the co-occurrence of each answer and the given question based on Web search results. However, to use Magnini's method, we also need some word-ignoring heuristic rules to remove search keywords when the number of returned web pages is insufficient.

3. The SCO-QAT Ranking Feature

Before comparing shallow ranking features, we define the SCO-QAT ranking feature that was applied successfully in the ASQA2 system at NTCIR-6. SCO-QAT relies on co-occurrence information about question terms and answer terms, and is therefore similar to Magnini's approach [Magnini *et al*. 2001]. However, unlike Magnini's approach, which utilizes the Web as a corpus to help answer questions posed on a local corpus, SCO-QAT uses passages retrieved by the passage retrieval module from the local corpus directly and does not use any word-ignoring rules.

The basic assumption of SCO-QAT is that, with good quality passages, the more often an answer co-occurs with question terms, the higher the probability that it is correct. Next, we describe the SCO-QAT function. Let the given answer be A and the given question be Q, where Q consists of a set, QT, of question terms $\{qt_1, qt_2, qt_3, \ldots\ldots, qt_n\}$. Based on QT, we define QC as a set of question term combinations, or more precisely $\{qc_i \mid qc_i$ is a subset of QT and qc_i is not empty$\}$. We also define a *freq(X)* function of a set X to indicate the number of retrieved passages in which all elements of X co-occur. The relation confidence is calculated as:

$$Conf(qc_i, A) = \begin{cases} \dfrac{freq(qc_i, A)}{freq(qc_i)}, & \text{if } freq(qc_i) \neq 0 \\ 0, & \text{if } freq(qc_i) = 0 \end{cases}. \quad (1)$$

Then, the SCO-QAT formula is defined as:

$$SCO\text{-}QAT(A) = \sum_{i=1}^{|QC|} Conf(qc_i, A). \quad (2)$$

For example, given a question Q consisting of three question terms {qt1, qt2, qt3} and a corresponding answer set {c1, c2}, the retrieved passages are presented as follows:

P1: qt1 qt2 c2

P2: qt1 qt2 qt3 c1

P3: qt1 qt2 c1

P4: qt1 c2

P5: qt2 c2

P6: qt1 qt3 c1 .

We use Equation (2) to calculate the candidate answer's SCO-QAT score as follows:

$$SCO\text{-}QAT(c1) = \frac{freq(qt1,c1)}{freq(qt1)} + \frac{freq(qt2,c1)}{freq(qt2)} + \frac{freq(qt3,c1)}{freq(qt3)} + \frac{freq(qt1,qt2,c1)}{freq(qt1,qt2)} + \frac{freq(qt1,qt3,c1)}{freq(qt1,qt3)} + \frac{freq(qt2,qt3,c1)}{freq(qt2,qt3)} + \frac{freq(qt1,qt2,qt3,c1)}{freq(qt1,qt2,qt3)}$$

$$= \frac{3}{5} + \frac{2}{4} + \frac{2}{2} + \frac{2}{3} + \frac{2}{2} + \frac{1}{1} + \frac{1}{1} = 5.77$$

$$SCO\text{-}QAT(c2) = \frac{2}{5} + \frac{2}{4} + \frac{0}{2} + \frac{1}{3} + \frac{0}{2} + \frac{0}{1} + \frac{0}{1} = 1.23 \quad .$$

Since the SCO-QAT score of c1 is higher than that of c2, c1 is considered a better answer candidate than c2.

The rationale behind SCO-QAT is that we try to use retrieved passages as a resource to look up question terms and locate the correct answer. When a set of question terms QT co-occurs with an answer A, we can infer that some kind of relation exists between the QT set and the answer A, which could be helpful for identifying correct answers. However, as this kind of relation is not always correct, we have to find a way to deal with noisy relations. To this end, we use the confidence score shown in Equation (1) to measure the goodness of a rule, which is similar to the method used for finding association rules. Then, we take the sum of the confidence scores of all the co-occurrences of all question term combinations to resolve the noisy rule problem. This technique is useful if the returned passages contain a lot of redundant information about the given question and the answer.

4. Evaluation Metrics

In this section, we describe the evaluation metrics used in this paper.

R-Accuracy and RU-Accuracy

Two metrics, R-Accuracy and RU-Accuracy, are used to measure QA performance in NTCIR CLQA. A QA system returns a list of ranked answer responses for each question, but R-accuracy and RU-accuracy only consider the correctness of the top-1 ranked answer response on the list. An answer response is a pair comprised of an answer and its source document. Each answer response is judged as Right, Unsupported, or Wrong, as defined in the

NTCIR-6 CLQA overview [Lee *et al*. 2007]:

"Right (R): the answer is correct and the source document supports it.

Unsupported (U): the answer is correct, but the source document cannot support it as a correct answer. That is, there is insufficient information in the document for users to confirm by themselves that the answer is the correct one.

Wrong (W): the answer "is incorrect."

Based on these criteria, the accuracy is calculated as the number of correctly answered questions divided by the total number of questions. R-accuracy means that only "Right" judgments are regarded as correct, while RU-accurakcy means that both "Right" and "Unsupported" judgments are counted. As R-accuracy only occurs a few times in this paper, we use "accuracy" to refer to RU-accuracy when the context is not ambiguous.

$$R-Accuracy = \frac{\text{the number of questions for which the top1 rank answer is Right}}{\text{number of questions}}$$

$$RU-Accuracy = \frac{\text{the number of questions for which the top1 rank answer is Right or Unsupported}}{\text{number of questions}}$$

Mean Reciprocal Rank (MRR)

We use MRR when we want to measure QA performance based on all the highest ranked correct answers, not only the top1 answer. MRR is calculated as follows:

$$MRR = \frac{1}{\text{number of questions}} \sum_{question_i} \begin{cases} \frac{1}{\text{the highest rank of correct answers}}, & \text{if a correct answer exists} \\ 0, & \text{if no correct answer} \end{cases}$$

Expected Answer Accuracy (EAA)

In addition to using the normal answer accuracy metrics, we propose a new metric called the Expected Answer Accuracy (EAA). We use EAA for cases where there are several top answers with the same ranking score.

The EAA score of a ranking method is defined as follows:

$$EAA = \frac{1}{\text{number of questions}} \sum_{question_i} \frac{\text{number of correct answers with top1 rank score}}{\text{number of answers with top1 rank score}}$$

Translation Cost

We use the "translation cost" metric to measure the cost of introducing the cross-lingual function to a QA system. It is calculated as follows:

$$TranslationCost = \frac{\text{accuracy of crosslingual QA - accuracy of monolingual QA}}{\text{accuracy of monolingual QA}}$$

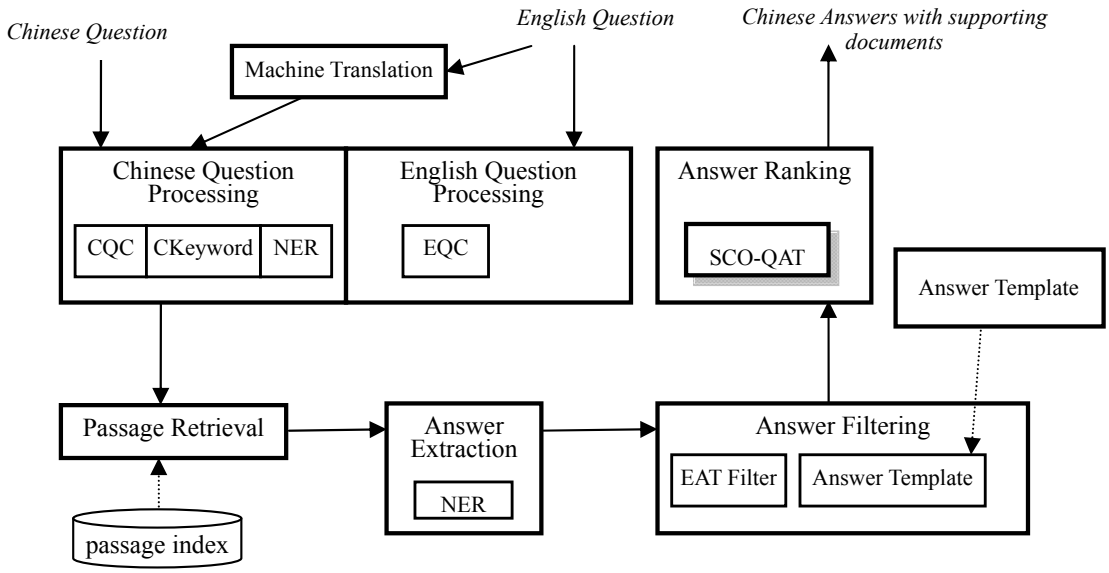

Figure 1. System architecture of ASQA2 for Chinese-Chinese and English-Chinese Factoid QA

5. The Testbed System: the ASQA2 Question Answering System

To evaluate answer ranking features, we chose the Academia Sinica Question Answering (ASQA) system as the testbed system for our experiment because it is modular and it performs well. Moreover, we can easily input different types of noise by adjusting the QA modules in ASQA. The system was developed by Academia Sinica[4] to deal with Chinese related QA tasks. The first version, ASQA1, can only deal with C-C QA, though. ASQA2, which is an extension of ASQA1, can deal with both C-C and E-C QA. We used ASQA1 in NTCIR-5 CLQA and ASQA2 in NTCIR-6 CLQA. NTCIR CLQA is the only QA contest in the world that focuses on Asian languages.

On the C-C and E-C subtasks in NTCIR-6 CLQA, ASQA2 achieved the best performance with 0.553 and 0.34 RU-Accuracy, respectively. The system consists of several modules, as shown in Figure 1. In Question Processing, ASQA2 uses SVMs (Support Vector Machines) and syntax rules to identify the input question type and infer the expected answer types. The type taxonomy has 6 coarse-grained and 62 fined-grained answer types. For passage retrieval, we use Lucene[5], an open source IR engine. The passage depth (the largest number of passages returned by the Passage Retrieval module) for each question is 100. Answers are then extracted from the returned passages by a fined-grained NER engine, and

[4] Academia Sinica, http://www.sinica.edu.tw

[5] Lucene, http://lucene.apache.org/

filtered by the Answer Filtering module according to the question type, answer type, and a mapping table that defines the types' compatibility. The final input for Answer Ranking is comprised of the question, the retrieved passages, and a set of filtered answers. Several answer ranking features are combined as a weighted sum. To deal with cross-lingual QA, ASQA2 adopts the *question translation* approach. Questions are translated with off-the-shelf machine translation engines.

Normally, a cross-lingual QA system is constructed by modifying some components of a monolingual system; however, since translation is involved, the approach often results in performance deterioration. The degree of performance deterioration is usually used with the accuracy metric to evaluate the effectiveness of a cross-lingual system. We define the performance deterioration in terms of the translation cost, which is defined in Section 4. Figure 2 shows the *translation cost* of systems in NTCIR-6 CLQA. When measuring the RU-Accuracy, the translation cost of ASQA2 ranks third, only slightly lower than the system in second place. Therefore, we consider that ASQA2 is an acceptable platform for our mono-lingual and cross-lingual experiments.

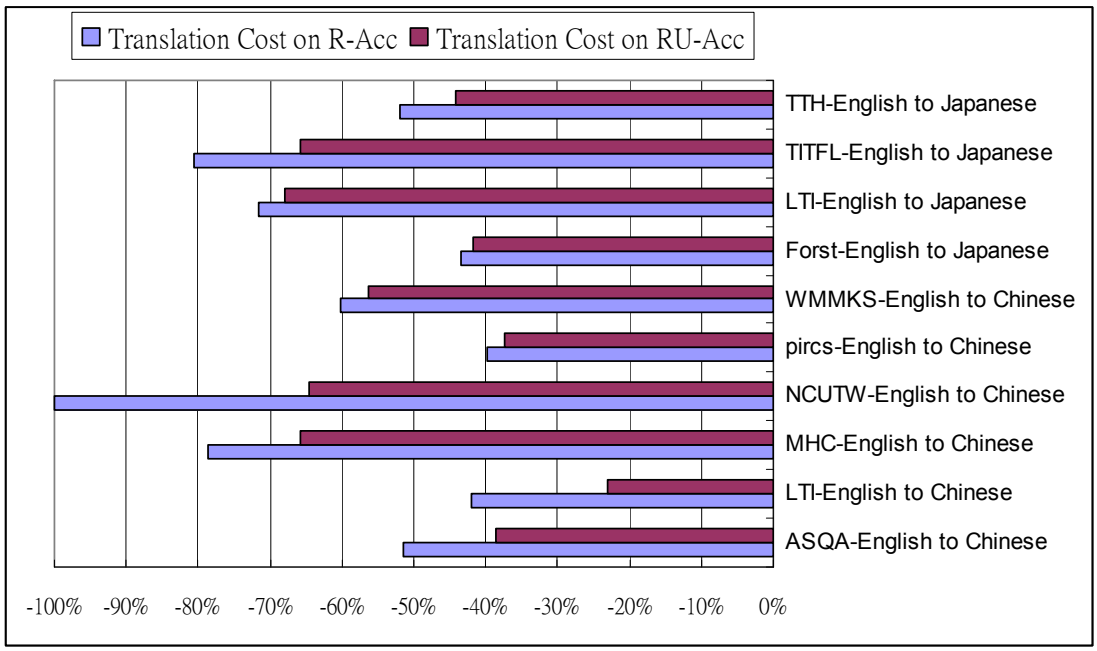

Figure 2. Translation costs of NTCIR-6 CLQA systems for factoid questions. The translation cost is calculated as the performance difference between cross-lingual and mono-lingual systems, divided by the mono-lingual performance.

According to the ASQA2 working notes [Lee *et al*. 2007], the system's success is attributable to three techniques: English question classification, answer template-based answer

filtering, and answer ranking with the SCO-QAT feature. When the answer template-based answer filter is applied, it removes all the candidates except the one it deems correct. As it is impossible to compare ranking methods when there is only one answer, we removed the answer template-based filter so that it would not influence our analysis of the answer ranking features.

6. Experiments

We conducted four experiments to explore the behavior of SCO-QAT and other shallow ranking features. In Experiment 1, we observed how shallow ranking features perform when a monolingual QA system is extended to a cross-lingual system. In Experiments 2, 3, and 4, we simulated situations where noise is introduced from the front-end modules and tried to determine which ranking feature is the most suitable under each kind of noise.

6.1 Variable Dependencies

Our testbed system is composed of several modules. Having described the system architecture in Section 5, we now elaborate on the dependencies between the experimental variables. First, we analyze the testbed system to identify several experimental variables and determine their interdependency, as shown in Figure 3. We are interested in the variables in bold font, as they will be used as independent or dependent variables in our experiments. The variables in gray font are not of interest because they are always controlled in the experiments. We provide details of the interdependency of the variables next.

In this study, we focus on the *Accuracy* and other QA performance metrics; therefore, they are always dependent variables. These performance metrics are directly influenced by three variables: the *ranking feature*, *passage quality*, and *answer quality*, since ranking features can use passages and answers. Furthermore, *passage quality* depends on the information retrieval model (*IR model*) used and the *passage depth* (the number of passages used for answer extraction). The greater the *passage depth*, the worse the passage quality is likely to be, which could result in more answers of progressively lower quality.

When ASQA switches from a monolingual to cross-lingual task, two variables are triggered: *translation* and *English question classification*. When translation is active, a translation engine has to be chosen to translate the question. Bad translation quality has a chain reaction effect because it leads to bad query quality, which leads to bad *passage quality* and bad *answer quality*. In ASQA, answer extraction is based on named entity recognition (NER) and answer filtering is based on the compatibility of the question type and the answer type. Therefore, NER and *question classification* are two more variables that could influence *answer quality*.

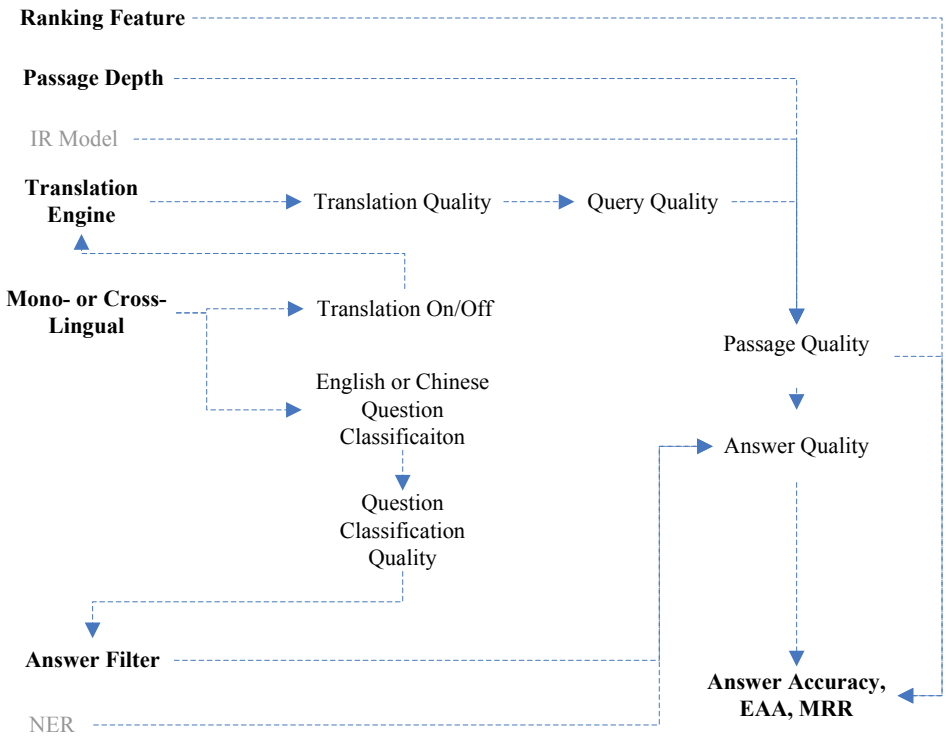

Figure 3. Dependencies of experimental variables based on the architecture of ASQA 2. When a variable at the tail of an arrow changes, it would have influence on the variable at the arrow head.

6.2 QA Datasets

We experimented on several QA datasets. A QA dataset is comprised of a set of questions, their answers, and the document IDs of supporting documents. The answers and supporting documents are regarded as the gold standard. We used the following six datasets from NTCIR5 and NTCIR6 for the CLQA Chinese-Chinese (CC) and English-Chinese (EC) subtasks: NTCIR5-CC-D200, NTCIR5-CC-T200, NTCIR5-EC-D200, NTCIR5-EC-T200, NTCIR6-CC-T150, and NTCIR6-EC-T150. The last item of a dataset name indicates the number of questions and the dataset's purpose, where T stands for "test" and D stands for "development". The CIRB40 corpus was used to compile the NTCIR5 CLQA datasets. It contains 901,446 Chinese newspaper news items published in 2000 and 2001. The corpus used for NTCIR6 CLQA was CIRB20, and it contains 249,508 Chinese newspaper news items published in 1998 and 1999.

Table 1. Datasets for experiments in this paper. Datasets created by NTCIR also has corresponding expanded datasets which consist of extra answer for post-hoc experiment. We postfix a "e" letter to the original name as the name of the expanded dataset name.

	corpus	question amount	creator	languages
NTCIR5-CC-D200	CIRB40	200	NTCIR	C-C
NTCIR5-CC-T200	CIRB40	200	NTCIR	C-C
NTCIR6-CC-T150	CIRB20	150	NTCIR	C-C
IASL-CC-Q465	CIRB40	465	Academia Sinica	C-C
1015				
NTCIR5-EC-D200	CIRB40	200	NTCIR	E-C
NTCIR5-EC-T200	CIRB40	200	NTCIR	E-C
NTCIR6-EC-T150	CIRB20	150	NTCIR	E-C
550				

According to Lin [Lin 2005], datasets created by QA evaluation forums are not suitable for post-hoc evaluation because the gold standard is not sufficiently comprehensive. This means we have to manually check all the extra answers not covered by the gold standard in order to derive more reliable experiment results. Since the number of questions in our experiments is quite large, it is not feasible for us to examine all the extra answers and their supporting documents. Therefore, we only use RU-accuracy to compare performances so that we do not have to check all the returned documents; only the answers are checked. These manually examined answers are then fed back to the datasets to form six expanded datasets: NTCIR5-CC-D200e, NTCIR5-CC-T200e, NTCIR5-EC-D200e, NTCIR5-EC-T200e, NTCIR6-CC-T150e, and NTCIR6-EC-T150e. In addition, we created the IASL-CC-Q465 dataset to increase the degree of confidence in our experiments. It was developed by three people using a program that randomly selected passages from the CIRB40 corpus, searched for relevant documents, and created questions from the collected documents. Finally, we had 1015 questions for the C-C task and 550 questions for the E-C task.

6.3 Experiment 1 – Single Shallow Features

Answer correctness features are usually combined in order to achieve the best performance. However, combining features in QA relies mostly on heuristic methods. Although some systems use machine learning approaches successfully for QA ranking, it is rare to see the same approach being applied to other QA work. This may be because QA feature combination methods are not mature enough to deal with the variability of QA systems, and the amount of

training data is not sufficient to train good models. Therefore, instead of combined features, we only studied the effect of single ranking features because we assume they are more reliable and can be easily applied to other systems or languages. Table 2 shows the experimental set-up.

Table 2. Experimental Set-up for Experiment 1 – Single Shallow Features

Independent Variables	Ranking Feature, Mono- or Cross-lingual
Dependent Variables	Accuracy, MRR, EAA
Controlled Variables	Passage Depth, Translation Engine, Answer Filter

Along with SCO-QAT, we tested the following widely used shallow features: *keyword overlap* (KO), *density*, *IR score* (IR), *mutual information score* (MI), and *answer frequency*. The *keyword overlap* feature represents the ratio of question keywords found in a passage, as used in [Cooper and Ruger 2000; Molla and Gardiner 2005; Zhao *et al*. 2005]. The *IR score* [Kwok and Deng 2006; Zheng 2002], which is provided by the passage retrieval module, is the score of the passage containing the answer. In ASQA2, the *IR score* is produced by the Lucene information retrieval engine[6]. *Density* is defined as the average distance between the answer and question keywords in a passage. There are several ways to calculate density. In this experiment, we simply adopt Lin's formula [Lin *et al*. 2005], which performed well in NTCIR-5 CLQA. The *mutual information score* is calculated by the PMI method used in [Magnini *et al*. 2001], and instead of being based on the Web, it is calculated based on the whole corpus.

The experiment results are listed in Table 3. SCO-QAT performs very well on C-C datasets, achieving 0.522 EAA for the NTCIR5-CC-D200e dataset, 0.515 for the NTCIR5-CC-T200e dataset, 0.546 for the IASL-CC-Q465 dataset, and 0.406 for the NTCIR6-CC-T150 dataset. Compared to other features, the differences are in the range 0.063~0.522 for EAA.

[6] We adopted Lucene 2.0.0, which uses Vector Space Model as the default method to calculate the IR score of a document. Detail information can be found in the Lucene API documentation: Class Similarity:http://lucene.apache.org/java/2_0_0/api/org/apache/lucene/search/Similarity.htmlClass DefaultSimilarity:
http://lucene.apache.org/java/2_0_0/api/org/apache/lucene/search/DefaultSimilarity.html

Table 3. The performance of single features. "Accuracy" is the RU-Accuracy, "MRR" is Top5 RU-Mean-Reciprocal-Rank scores, and "EAA" is the Expected Answer Accuracy. CC-ALL and EC-ALL are the respective combinations of all the CC and EC datasets.

	NTCIR5-CC-D200e			**NTCIR5-CC-T200e**		
	Accuracy	EAA	MRR	Accuracy	EAA	MRR
SCOQAT	**0.545**	**0.522**	**0.621**	**0.515**	**0.515**	**0.586**
KO	0.515	0.254	0.601	0.495	0.245	0.569
Density	0.375	0.368	0.501	0.390	0.380	0.479
Frequency	0.445	0.431	0.560	0.395	0.366	0.499
IR	0.515	0.425	0.598	0.495	0.420	0.569
MI	0.210	0.210	0.342	0.155	0.290	0.138
	IASL-CC-Q465			**NTCIR6-CC-T150**		
	Accuracy	EAA	MRR	Accuracy	EAA	MRR
SCOQAT	**0.578**	**0.546**	**0.628**	**0.413**	**0.406**	**0.495**
KO	0.568	0.247	0.618	0.367	0.130	0.476
Density	0.432	0.369	0.519	0.340	0.314	0.420
Frequency	0.413	0.406	0.486	0.340	0.343	0.431
IR	0.518	0.406	0.587	0.367	0.283	0.460
MI	0.138	0.124	0.280	0.167	0.142	0.281
	NTCIR5-EC-D200			**NTCIR5-EC-T200**		
	Accuracy	EAA	MRR	Accuracy	EAA	MRR
SCOQAT	0.250	0.240	0.349	0.185	0.187	0.265
KO	0.290	0.117	0.376	0.195	0.093	0.288
Density	0.190	0.186	0.294	0.180	0.177	0.245
Frequency	**0.300**	**0.297**	**0.394**	0.190	0.181	0.280
IR	0.295	0.262	0.385	**0.270**	**0.210**	**0.326**
MI	0.145	0.145	0.262	0.060	0.046	0.164
	NTCIR6-EC-T150					
	Accuracy	EAA	MRR			
SCOQAT	0.193	0.180	0.268			
KO	**0.220**	0.061	**0.292**			
Density	0.187	0.180	0.268			
Frequency	0.213	0.194	0.283			
IR	0.180	**0.265**	0.146			
MI	0.107	0.069	0.205			
	CC-ALL			**EC-ALL**		
	Accuracy	EAA	MRR	Accuracy	EAA	MRR
SCOQAT	**0.535**	**0.514**	**0.599**	0.211	0.204	0.296
KO	0.513	0.231	0.584	0.236	0.093	0.321
Density	0.399	0.363	0.493	0.185	0.181	0.269
Frequency	0.405	0.394	0.495	0.236	**0.227**	0.322
IR	0.491	0.424	0.538	**0.255**	0.212	**0.331**
MI	0.160	0.176	0.264	0.104	0.088	0.211

In addition to comparing single ranking features, we compared the SCO-QAT results with those of other participants in the NTCIR5 CLQA task (Table 4). As the other QA systems used combined features, this is a single- versus combined-feature comparison. In the NTCIR5 CLQA task [Sasaki *et al*. 2005], there were thirteen Chinese QA runs with an accuracy range of 0.105~0.445, and a mean of 0.315. It is impressive that ASQA2 with the single SCO-QAT feature achieved 0.515 accuracy[7], which was much better than the accuracy of ASQA1 [Lee *et al*. 2005], the best performing system in the NTCIR5 CLQA C-C subtask.

Table 4. Performance comparison of SCO-QAT (single feature) and the best systems at NTCIR5 and NTCIR6 CLQA (combined features)

Subtask	System	RU-Accuracy
NTCIR5 CC	Best Participant (ASQA1)	0.445
	ASQA2 with SCO-QAT only	0.515
NTCIR5 EC	Best Participant	0.165
	ASQA2 with SCO-QAT only	0.185
NTCIR6 CC	Best Participant (ASQA2 full version)	0.553
	ASQA2 with SCO-QAT only	0.413
NTCIR6 EC	Best Participant (ASQA2 full version)	0.340
	ASQA2 with SCO-QAT only	0.193

Although SCO-QAT still performs well on the E-C datasets, its performance is not as good as on the C-C datasets. After analyzing the failed cases of E-C QA, we found the major problem was that some translations introduced words not listed in the stop word list. For example, there were some English questions in NTCIR CLQA, such as "Who is in charge of Indonesia's cabinet in 2000?" After processing their Google translations, we identified improper keywords that were not on our stop word lists. For example, in the translation of the above question, "由誰負責的印尼內閣於 2000 年?", we found "由" and "於". Since SCO-QAT aggregates all co-occurrence scores, the effect of improper keywords is compounded. Although this problem could be solved by simply adding more stop words to the list, it should be noted that more new stop words may be introduced if the machine translation engine is changed. A better solution is to use the term-by-term translation approach because the stop word list can be controlled more easily.

Although *frequency* is the simplest of the shallow features, it performs surprisingly well. It even achieves the best performance on one E-C dataset (NTCIR5-EC-D200). This may be

[7] The 0.515 accuracy is based on NTCIR5-CC-T200e dataset. If based on the NTCIR5-CC-T200 dataset, the accuracy is 0.505

due to the effectiveness of the ASQA2 answer filtering module, the characteristics of the Chinese news corpus, or the way questions were created, which caused questions with high frequency answers to be selected. We cannot find any papers on the effect of applying the frequency feature only. Further investigation is, therefore, needed to explain the phenomenon.

The density feature measures the density of question terms around the answer based on the co-occurrence and distance information. Although it is widely used in QA systems, its performance is not as good as that of the IR score, which does not consider the distance information. This could be because the distance information is much noisier in QA that involves Chinese (*e.g.*, E-C and C-C).

We identified two types of errors caused by machine translations: wrong-term errors and synonym errors. Both types have a negative effect on the ranking features because the quality of the passages is often poor. The following is an example of a wrong term error. For the English question "Who is the director of the Chinese movie Crouching Tiger, Hidden Dragon?", the word "director" was translated by Google Translate to the wrong term "新任" in "誰是新任的中國電影臥虎藏龍?". Here, the semantics of "director" and "新任" are completely different. In cases like this, it is impossible to find good quality passages for ranking. Synonym errors occur when improper synonyms are introduced. For example, the English question "Who was Taiwan's Central Bank Governor with the longest tenure?" is translated to "誰是台灣的央行行長最長任期?" by Google. Although "行長" is the correct translation for mainland China, it is not the normal way to describe the head of a bank in Taiwan; therefore, a query with "行長" can not retrieve appropriate passages from Taiwanese news corpora (*e.g.*, CIRB40 and CIRB20).

6.4 Experiment 2 –Influence of Machine Translation Quality

To develop a cross-lingual QA system, a monolingual system is usually created first and then some modules are adjusted to meet cross-lingual requirements. There are two widely used approaches: question translation and term-by-term translation. In the question translation approach, the question is translated into the target language by machine translation. The translated question is then input to the monolingual system. In the term-by-term approach, questions are analyzed in the source language and split into several important terms, which are then translated by using a bilingual dictionary or other techniques.

Since ASQA2 adopts the question translation approach, we can control the translation quality intuitively using different machine translation engines. Noisy information introduced by a machine translation engine propagates down through the QA modules and results in wrong answers. We tested our system on two machine translation services (namely, Google

Translate and SYSTRAN[8]) to determine how the translation quality affects the answer ranking features. Table 5 shows the experimental set-up.

Table 5. Experimental Set-up for Experiment 2 – Influence of Machine Translation Quality

Independent Variables	Ranking Feature, Translation Engine
Dependent Variables	Accuracy, MRR, EAA
Controlled Variables	Passage Depth, Mono- or Cross-lingual, Answer Filter

We observe that Google's translation quality is better than that of SYSTRAN. In other words, the accuracy declines when Google Translate is replaced by SYSTRAN. The performance decrease ratio (calculated as the performance of using SYSTRAN divided by that of using Google) for each of the three E-C datasets is shown in Table 6. It seems to be difficult to predict the influence of the translation quality. If we only look at each dataset, the decrease ratio is quite unstable, ranging from 48.3% to 96.9% in terms of accuracy. However, when we consider the ratio based on all the datasets, it becomes more stable for all the ranking features. The standard deviation of the decrease in the accuracy ratio drops from more than 0.11 to 0.0655, which shows that the current datasets of NTCIR CLQA may be too small to be used with confidence in our experiments. Thus, it would be better to use all the EC datasets when comparing QA systems.

For the EC-ALL dataset, SCO-QAT yields a better performance decrease ratio in terms of accuracy and EAA, but not in terms of MRR. The Frequency feature still performs relatively well, because the frequency of an answer is less dependant on the translation quality.

[8] We used the Yahoo! BABEL FISH service, which is powered by SYSTRAN. The translations were obtained from Google and Yahoo in May 2007 and June 2007, respectively.

Table 6. Performance decrease ratio of shallow features on E-C QA when Google is replaced by SYSTRAN.

	(a) NTCIR5-EC-D200			(b) NTCIR5-EC-T200		
	Accuracy	EAA	MRR	Accuracy	EAA	MRR
SCOQAT	**80.0%**	79.2%	62.6%	59.5%	59.0%	67.4%
KO	69.0%	**83.5%**	76.2%	51.3%	58.7%	60.9%
Density	73.7%	75.3%	78.7%	61.1%	**59.8%**	68.2%
Frequency	73.3%	68.8%	77.5%	47.4%	47.4%	62.8%
IR	79.7%	80.5%	**78.9%**	35.2%	45.1%	49.3%
MI	48.3%	34.5%	66.8%	**91.7%**	72.2%	**79.0%**
Stdev.	*0.1173*	*0.1826*	*0.0697*	*0.1910*	*0.0980*	*0.0980*
	(c) NTCIR6-EC-T150			(d) EC-ALL		
	Accuracy	EAA	MRR	Accuracy	EAA	MRR
SCOQAT	82.8%	86.2%	83.1%	**74.1%**	**74.2%**	69.2%
KO	87.9%	71.9%	85.7%	68.5%	72.4%	73.6%
Density	89.3%	88.3%	85.7%	73.5%	73.3%	**77.1%**
Frequency	**96.9%**	**97.6%**	**91.3%**	71.5%	69.3%	76.2%
IR	66.7%	62.3%	71.3%	60.0%	64.3%	66.6%
MI	56.2%	51.0%	72.6%	59.6%	45.1%	71.8%
Stdev.	*0.1538*	*0.1762*	*0.0794*	*0.0655*	*0.1105*	*0.0403*

6.5 Experiment 3 –Influence of Passage Quality Introduced by Deep Passages

Passage depth, defined as the number of passages used for answer extraction and answer ranking, plays a critical role in a QA system. On the one hand, by increasing the passage depth we can obtain more relevant passages and, therefore, have a better chance of improving QA performance. On the other hand, increasing the passage depth also introduces more irrelevant passages. If a ranking feature can not handle the noise caused by deep passages, it can not benefit from additional relevant passages.

In this experiment, we increase the number of passages to evaluate the performance of shallow features when the number of irrelevant passages increases. The experimental setup is shown in Table 7.

Table 7. Experimental Set-up for Experiment 3 – Influence of Passage Quality Introduced by Deep Passages.

Independent Variables	Ranking Feature, Passage Depth, Mono-or Cross-lingual
Dependent Variables	Accuracy, MRR, EAA
Controlled Variables	Translation Engine, Answer Filter

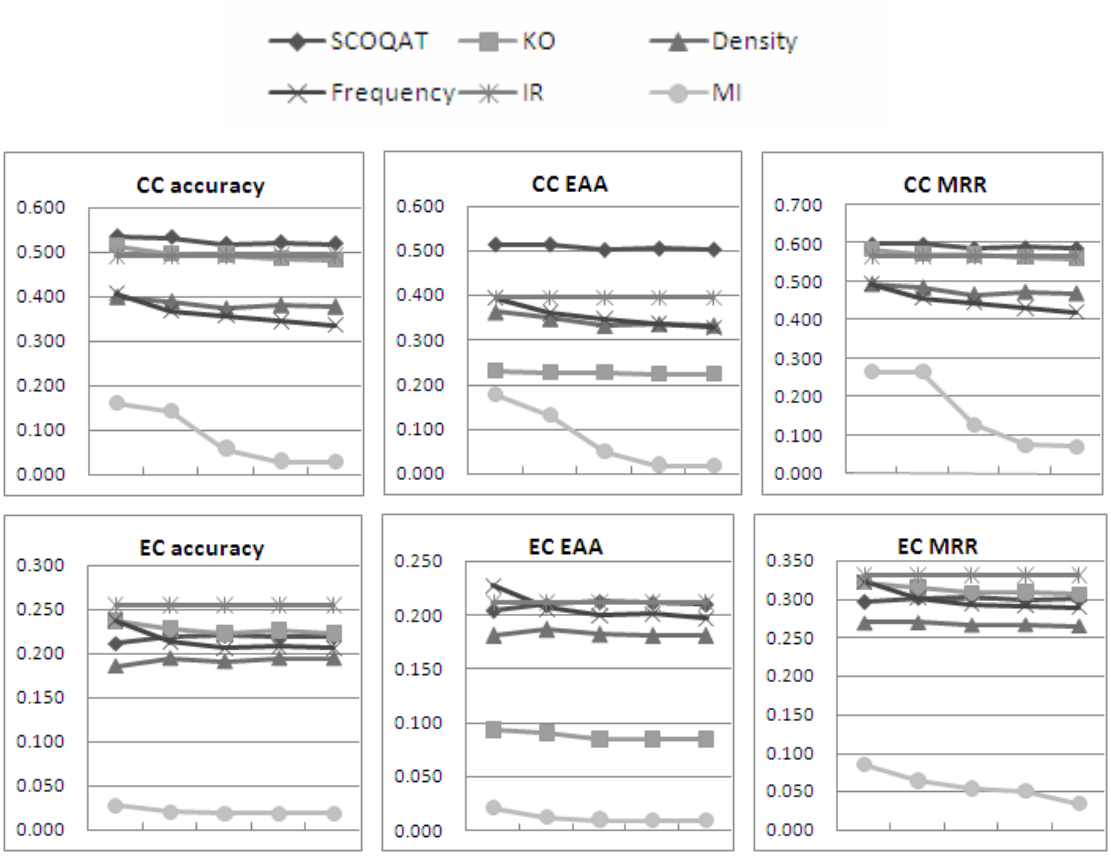

Figure 4. Single feature accuracy over 5 passage depth points (100, 200, 300, 400, 500) for all C-C and E-C datasets.

We observe the performance of all C-C and E-C datasets at five depth points between 100 and 500, as shown in Figure 4. We chose 100 as the starting depth because it is commonly adopted in QA systems as the document depth or passage depth. As expected, for both CC and EC situations, EAA declines when the passage depth increases. (The IR score ranking feature is an exception. It always remains the same because the passage IR score of an answer does not change when the passage depth increases). However, the decrease in EAA is not as high as we expected, which suggests that, with the exception of frequency and MI, shallow ranking features can handle deep passage noise.

Among the ranking features, *frequency* and *MI* are influenced by passage depth the most. In EC, while *frequency* is the best at depth 100 in terms of EAA, the latter decreases rapidly when the passage depth increases to 200, which is much more unreliable than in the CC situation. In other words, the *accuracy* feature is much more unreliable in EC. For *MI*, it not only performed worse than the other features in terms of EAA, but also decreased substantially when the depth increased. This suggests that *MI* may not be suitable for retrieved passages, although it has been applied successfully when using the Web as a corpus.

Some of the examples found confirm that the number of irrelevant passages increases when the number of passages increases. For example, when the number of passages is 100, the most frequent answer given to the Chinese question "西元 2000 年加入奧地利聯合政府的自由黨黨魁是誰？" (Who is the leader of Freedom Party joining the Austria coalition government in 2000?) is "海德" (Haider), which is correct. However, when the number increases to 200, the most frequent answer is "小澤一郎" (OZAWA Ichiro), which is incorrect. This causes *density* and the other shallow features to fail in this situation.

6.6 Experiment 4 –Influence of Answer Quality

As answer ranking is directly influenced by the answer quality, it is important to evaluate the ranking feature on answers of different quality. In this experiment, we adjusted the answer quality by changing the answer filter. The experimental set-up is detailed in Table 8.

Table 8. Experimental Set-Up for Experiment 4 – Influence of Answer Quality

Independent Variables	Ranking Feature, Mono- or Cross-lingual
Dependent Variables	Accuracy, MRR, EAA
Controlled Variables	Passage Depth, Translation Engine, Answer Filter

The Expected Answer Type filter (EAT filter) is a submodule of ASQA2 that eliminates answers deemed incompatible with the question type. For example, if the question type is Q_LOCATION_COUNTRY, only answers representing countries will be retained. It is common for QA systems to use this kind of filtering mechanism, but they differ in the granularity of the answer type system they use. With a good EAT filter, the quality of the input for the subsequent Answer Ranking module will be less noisy and easier to deal with.

By utilizing the ASQA2 answer-type system (*i.e.*, 6 coarse-grained and 62 fine-grained types), we can experiment with answer ranking features on different granularities. We built three EAT filters, namely, a DoNothing Filter, a Coarse-grained Filter[9], and a Fine-grained Filter. The DoNothing Filter does not filter out any answers; therefore, it may contain a lot of noisy information. The Coarse-grained Filter and Fine-grained Filter use coarse-grained and

[9] The Fine-grained filter was used in ASQA1 and ASQA2

fine-grained type information respectively.

The Fine-grained Filter is used in the single feature experiment described in Section 6.3. Here, we conduct the same single feature experiment with the other two noisier EAT filters. The results are shown in Table 9. As expected, the performance of every feature deteriorates with the noisy EAT filters. In the CC datasets, with the Coarse-grained Filter, SCO-QAT's EAA declines from 0.514 to 0.499 on the CC-ALL dataset, but it is still better than the other features. Even with the noisiest DoNothing Filter, SCO-QAT can still maintain a 71% decrease ratio for the CC-ALL dataset, thereby demonstrating its robustness. The calculation of decrease ratios in this section is similar to that in the "Influence of Machine Translation Quality" section. When speaking of Coarse-grained Filter, it is calculated as the performance of using Coarse-grained Filter divided by the performance of using Fine-grained Filter. When speaking of DoNothing Filter, the formula is the same except that the numerator is replaced with the performance of using DoNothing Filter.

Table 9(a). Performance and decrease ratio in CC QA when the Coarse-grained EAT filter is replaced by Fine-grained and DoNothing EAT filters.

Coarse-Grained (Decrease Ratio = Coarse-Grained / Fine-Grained)						
	(a) NTCIR5-CC-D200e			(b) NTCIR5-CC-T200e		
	Accuracy	EAA	MRR	Accuracy	EAA	MRR
SCOQAT	**0.515** (94%)	**0.492** (94%)	**0.594** (**96%**)	**0.5** (**97%**)	**0.498** (**97%**)	**0.56** (**96%**)
KO	0.475 (92%)	0.229 (90%)	0.564 (94%)	0.48 (**97%**)	0.224 (92%)	0.545 (**96%**)
Density	0.355 (**95%**)	0.35 (**95%**)	0.473 (95%)	0.345 (88%)	0.333 (88%)	0.441 (92%)
Frequency	0.41 (92%)	0.408 (**95%**)	0.524 (93%)	0.37 (94%)	0.344 (94%)	0.472 (95%)
IR	0.475 (92%)	0.392 (92%)	0.559 (94%)	0.465 (94%)	0.375 (89%)	0.539 (95%)
MI	0.035 (17%)	0.031 (15%)	0.089 (26%)	0.04 (26%)	0.034 (12%)	0.104 (75%)
	(c) IASL-CC-Q465			(d) NTCIR6-CC-T150		
	Accuracy	EAA	MRR	Accuracy	EAA	MRR
SCOQAT	**0.568** (**98%**)	**0.536** (**98%**)	**0.619** (**99%**)	**0.407** (98%)	**0.398** (98%)	**0.486** (98%)
KO	0.551 (97%)	0.232 (94%)	0.604 (98%)	0.367 (**100%**)	0.123 (95%)	0.468 (98%)
Density	0.406 (94%)	0.337 (91%)	0.498 (96%)	0.327 (96%)	0.301 (96%)	0.405 (97%)
Frequency	0.394 (95%)	0.385 (95%)	0.468 (96%)	0.34 (100%)	0.339 (**99%**)	0.43 (**100%**)
IR	0.508 (**98%**)	0.39 (96%)	0.576 (98%)	0.367 (100%)	0.269 (95%)	0.45 (98%)
MI	0.03 (22%)	0.02 (16%)	0.095 (34%)	0.06 (36%)	0.032 (23%)	0.124 (44%)
	(e) CC-ALL					
	Accuracy	EAA	MRR			
SCOQAT	**0.52** (**97%**)	**0.499** (**97%**)	**0.583** (**97%**)			
KO	0.495 (96%)	0.214 (93%)	0.564 (**97%**)			
Density	0.372 (93%)	0.333 (92%)	0.468 (95%)			
Frequency	0.384 (95%)	0.374 (95%)	0.474 (96%)			
IR	0.472 (96%)	0.369 (94%)	0.547 (**97%**)			
MI	0.037 (24%)	0.027 (16%)	0.1 (42%)			

Table 9 also shows the performance decrease ratio caused by inefficient EAT filters. It is calculated by dividing the performance score of a noisy EAT filter by that of the standard Fine-grained Filter. From this perspective, SCO-QAT is still the best CC feature, achieving 97% and 71% EAA decrease ratio with the Coarse-Grained Filter and DoNothing EAT filter, respectively.

Table 9(b). Performance and decrease ratio in CC QA when the Coarse-grained EAT filter is replaced by the Fine-grained and DoNothing EAT filters.

DoNothing (Decrease Ratio = DoNothing / Fine-Grained)						
	(f) NTCIR5-CC-D200e			(g) NTCIR5-CC-T200e		
	Accuracy	EAA	MRR	Accuracy	EAA	MRR
SCOQAT	**0.355** (65%)	**0.339** (65%)	**0.463** (**74%**)	**0.345** (**67%**)	**0.341** (**66%**)	**0.442** (**76%**)
KO	0.345 (**67%**)	0.082 (32%)	0.452 (75%)	0.315 (64%)	0.068 (28%)	0.414 (73%)
Density	0.16 (43%)	0.14 (38%)	0.16 (32%)	0.185 (47%)	0.179 (47%)	0.275 (57%)
Frequency	0.3 (**67%**)	0.285 (**66%**)	0.395 (71%)	0.23 (58%)	0.22 (60%)	0.331 (66%)
IR	0.32 (62%)	0.135 (32%)	0.43 (72%)	0.335 (68%)	0.152 (36%)	0.428 (75%)
MI	0.02 (10%)	0.018 (9%)	0.108 (32%)	0.015 (10%)	0.005 (2%)	0.036 (26%)
	(h) IASL-CC-Q465			(i) NTCIR6-CC-T150		
	Accuracy	EAA	MRR	Accuracy	EAA	MRR
SCOQAT	**0.428** (74%)	**0.406** (**74%**)	**0.52** (**83%**)	**0.293** (**71%**)	**0.295** (**73%**)	**0.374** (**76%**)
KO	0.426 (**75%**)	0.061 (25%)	0.513 (**83%**)	0.24 (65%)	0.034 (26%)	0.333 (70%)
Density	0.254 (59%)	0.179 (48%)	0.343 (66%)	0.153 (45%)	0.131 (42%)	0.246 (59%)
Frequency	0.288 (70%)	0.285 (70%)	0.356 (73%)	0.22 (65%)	0.223 (65%)	0.304 (70%)
IR	0.376 (73%)	0.211 (52%)	0.473 (80%)	0.24 (65%)	0.124 (44%)	0.331 (72%)
MI	0.013 (9%)	0.003 (2%)	0.04 (14%)	0.007 (4%)	0.001 (1%)	0.027 (10%)
	(j) CC-ALL					
	Accuracy	EAA	MRR			
SCOQAT	**0.377** (**70%**)	**0.364** (**71%**)	**0.472** (**79%**)			
KO	0.361 (**70%**)	0.063 (27%)	0.455 (78%)			
Density	0.207 (51%)	0.164 (45%)	0.279 (57%)			
Frequency	0.269 (66%)	0.263 (67%)	0.351 (71%)			
IR	0.337 (69%)	0.171 (44%)	0.435 (77%)			
MI	0.014 (9%)	0.006 (3%)	0.051 (19%)			

The decline in some features is caused by too many answers being collocated in the same passage. Without a proper EAT filter, a passage could contain the correct answer and other answers; or, at worst, contain several answers, none of which are compatible with the given question. For example, the first returned passage for the Chinese question "請問西元 2000 年 7 月美方派何人前往北京對 TMD 以及其他全球戰略佈局與中方展開對話？" (Who is the delegate of United States visiting Beijing to negotiate the TMD issue in July, 2000?) does not

contain any answers related to the PERSON type. Without a proper filter, wrong answers in the top-ranked passages would be sent to the answer ranking module. As a result, the IR score would not help us differentiate between the correct answer and incorrect ones.

Note that the decline in EC's performance is substantial when the DoNothing filter is applied. In the CC case, the decline in EAA for the SCO-QAT feature is 71%; however, in the EC case, it drops to 14%. This suggests that, in EC, information about the answer type is important, since it is more reliable than the shallow ranking features under noise introduced by translation.

Table 10. Performance and decrease ratio in EC QA when the Coarse-grained EAT filter is replaced by the Fine-grained and DoNothing EAT filters.

Coarse-Grained (Coarse-Grained / Fine-Grained)						
	(a) NTCIR5-EC-D200			(b) NTCIR5-EC-T200		
	Accuracy	EAA	MRR	Accuracy	EAA	MRR
SCOQAT	0.2 (80%)	0.1947 (81%)	0.3019 (86%)	0.17 (**91%**)	0.1702 (**91%**)	0.2431 (**91%**)
KO	**0.255** (**87%**)	0.102 (**87%**)	0.334 (**88%**)	0.155 (79%)	0.07 (75%)	0.2499 (86%)
Density	0.16 (84%)	0.1537 (82%)	0.2517 (85%)	0.15 (83%)	0.1442 (81%)	0.2183 (88%)
Frequency	**0.255** (85%)	**0.2559** (86%)	**0.3486** (**88%**)	0.16 (84%)	0.1608 (88%)	0.2509 (89%)
IR	0.25 (84%)	0.2262 (86%)	0.3359 (87%)	**0.23** (85%)	**0.1826** (87%)	**0.2966** (90%)
MI	0.02 (13%)	0.0175 (12%)	0.0944 (36%)	0.015 (25%)	0.0106 (23%)	0.0655 (39%)
	(c) NTCIR6-EC-T150			(d) EC-ALL		
	Accuracy	EAA	MRR	Accuracy	EAA	MRR
SCOQAT	0.1867 (96%)	0.1711 (95%)	0.2586 (96%)	0.1855 (**87%**)	0.1794 (87%)	0.2687 (**90%**)
KO	0.1867 (84%)	0.0591 (97%)	**0.2702** (92%)	0.2 (84%)	0.0787 (84%)	0.286 (89%)
Density	0.18 (96%)	0.1766 (98%)	0.2559 (95%)	0.1618 (**87%**)	0.1565 (86%)	0.2407 (89%)
Frequency	**0.1933** (90%)	**0.1769** (91%)	0.268 (94%)	0.2036 (86%)	**0.1998** (87%)	0.2911 (**90%**)
IR	0.18 (**100%**)	0.1449 (**99%**)	0.2598 (**98%**)	**0.2236** (**87%**)	0.1882 (**88%**)	**0.3009** (**90%**)
MI	0.0533 (49%)	0.0391 (56%)	0.1108 (53%)	0.0273 (26%)	0.0209 (23%)	0.0884 (41%)
DoNothing (DoNothing / Fine-Grained)						
	(e) NTCIR5-EC-D200			(f) NTCIR5-EC-T200		
	Accuracy	EAA	MRR	Accuracy	EAA	MRR
SCOQAT	**0.02** (**8%**)	0.0226 (9%)	0.1254 (36%)	**0.035** (**19%**)	0.035 (19%)	0.1206 (**46%**)
KO	**0.02** (7%)	**0.0232** (20%)	0.1385 (**37%**)	0.015 (8%)	0.02 (**21%**)	0.1207 (42%)
Density	0.015 (**8%**)	0.019 (10%)	0.1013 (35%)	0.02 (11%)	0.0225 (13%)	0.0934 (38%)
Frequency	**0.02** (7%)	0.0163 (5%)	0.1365 (35%)	0.01 (5%)	0.01 (6%)	0.1124 (40%)
IR	**0.02** (7%)	0.067 (**26%**)	**0.1397** (36%)	0.02 (7%)	**0.0357** (17%)	**0.1278** (39%)
MI	0 (0%)	0.0004 (0%)	0.0184 (7%)	0.005 (8%)	0.0003 (1%)	0.0199 (12%)
	(g) NTCIR6-EC-T150			(h) EC-ALL		
	Accuracy	EAA	MRR	Accuracy	EAA	MRR
SCOQAT	**0.0267** (14%)	0.0267 (15%)	0.1086 (**41%**)	**0.0273** (**13%**)	0.0282 (14%)	0.1191 (**40%**)
KO	0.02 (9%)	0.0136 (**22%**)	**0.1102** (38%)	0.0182 (8%)	**0.0194** (**21%**)	0.1243 (39%)
Density	0.02 (11%)	0.0184 (10%)	0.1061 (40%)	0.0182 (10%)	0.0201 (11%)	0.0997 (37%)
Frequency	0.02 (9%)	0.02 (10%)	0.1043 (37%)	0.0164 (7%)	0.015 (7%)	0.119 (37%)
IR	0.0133 (7%)	**0.0294** (20%)	0.1 (38%)	0.0182 (7%)	**0.0453** (**21%**)	**0.1245** (38%)
MI	**0.0267** (**25%**)	0.0041 (6%)	0.0464 (23%)	0.0091 (9%)	0.0013 (2%)	0.0266 (13%)

7. Conclusion

Sometimes, the resources needed to apply deep answer ranking approaches in a language are not available or the resource quality is not good enough. Hence, we conducted this research to help QA system designers choose shallow ranking features. We experimented on six shallow ranking features (SCO-QAT, keyword overlap, density, IR score, mutual information score, and answer frequency) under various types of noise caused by different QA modules in mono-lingual and cross-lingual situations.

We also proposed a novel answer ranking feature called SCO-QAT, which does not require extra knowledge or sophisticated tools. It is, therefore, easy to implement in QA systems and may be used on various languages. In this pilot study, when the ASQA2 system only used the SCO-QAT ranking feature, it outperformed all the systems in NTCIR5 CLQA. For example, on the NTCIR5-CC-T200e QA dataset, we achieved 0.515 RU-Accuracy with the SCO-QAT feature only. Even the E-C version also achieved a 0.05 improvement over the best system. SCO-QAT also performed well in NTCIR6 CLQA, where the host system, ASQA2, achieved the best performance in the C-C subtask and the E-C subtask.

To understand SCO-QAT better and to gain a deeper insight into shallow answer ranking features, we tested answer ranking features in various scenarios. We found that, although SCO-QAT performed very well in C-C QA, frequency seems the best choice for ranking in E-C QA in terms of EAA. However, the decrease in translation quality has a marked effect on the frequency of EAA, as shown by the fact that the EAA decrease ratio is 69.3%. In the same situation, SCO-QAT maintained a 74.2% EAA decrease ratio which was the best among the shallow ranking features. We also found that the noise introduced by passage depth does not impact much on ranking performance. This suggests that, if a long processing time is allowed, QA based on deep passages is a possible way to improve the performance when shallow features are used. In addition, answer-type-based filtering plays an important role, especially for E-C. When an extremely bad filter was used, the EAA decrease ratio in E-C for shallow ranking features was only 2%~21%, which shows a proper answer filter with fined-grained NER is critical to the success of an E-C system.

In our future research on shallow ranking features, we will address the following issues. We will introduce a question term weighting scheme for SCO-QAT; use a taxonomy or ontology to alleviate the synonym problem that arises when counting co-occurrences of answers and question terms; experiment with shallow features on a Web corpus; utilize more syntactic information to make co-occurrence information more reliable; and test shallow features on other languages.

Acknowledgments

This research was supported in part by the National Science Council of Taiwan under Center of Excellence Grant NSC 95-2752-E-001-001-PAE, the Research Center for Humanities and Social Sciences, Academia Sinica, and Thematic program of Academia Sinica under Grant AS 95ASIA02. We would like to thank the Chinese Knowledge and Information Processing Group (CKIP) in Academia Sinica for providing us with AutoTag for Chinese word segmentation.

REFERENCES

Clarke, C.L.A., G. Cormack, G. Kemkes, M. Laszlo, T. Lynam, E. Terra, and P. Tilker, "Statistical Selection of Exact Answers (MultiText Experiments for TREC 2002)," in *Proc. of TREC, 2002*, pp. 823-831.

Clarke, C.L.A., G.V. Cormack, and T.R. Lynam, "Exploiting redundancy in question answering," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 2001, pp. 358-365.

Cooper, R.J. and S.M. Ruger, "A Simple Question Answering System," in *Proc. of TREC*, 2000.

Cui, H., R. Sun, K. Li, M.Y. Kan, and T.S. Chua, "Question answering passage retrieval using dependency relations," in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 2005, pp. 400-407.

Geleijnse, G. and J. Korst, "Learning Effective Surface Text Patterns for Information Extraction," in *Proceedings of the EACL 2006 workshop on Adaptive Text Extraction and Mining (ATEM 2006)*, 2006, pp. 1-8.

Gillard, L., L. Sitbon, E. Blaudez, P. Bellot, and M. El-B`eze, "The LIA at QA@CLEF-2006," in *CLEF*, 2006.

Harabagiu, S., D. Moldovan, C. Clark, M. Bowden, A. Hickl, and P. Wang, "Employing Two Question Answering Systems in TREC 2005," in *Proceedings of the Fourteenth Text REtrieval Conference*, 2005.

Kwok, K.-L. and P. Deng, P., "Chinese Question-Answering:Comparing Monolingual with English-Chinese Cross-Lingual Results," in *Asia Information Retrieval Symposium*, 2006, pp. 244-257.

Lee, C.-W., M.-Y. Day, C.-L. Sung, Y.-H. Lee, T-J. Jiang, C-W. Wu, C-W. Shih, Y-R. Chen, .and W.-L. Hsu, "Chinese-Chinese and English-Chinese Question Answering with ASQA at NTCIR-6 CLQA," in *Proceedings of NTCIR-6 Workshop*, 2007, pp. 175-181.

Lee, C.W., C.W. Shih, M.Y. Day, T.H. Tsai, T.J. Jiang, C.W. Wu, C.L. Sung, Y.R. Chen, S.H. Wu, and W.L. Hsu, "ASQA: Academia Sinica Question Answering System for NTCIR-5 CLQA," in *Proceedings of NTCIR-5 Workshop Meeting* , 2005, Tokyo, Japan.

Lin, F., H. Shima, M. Wang, and T. Mitamura, "CMU JAVELIN System for NTCIR5 CLQA1," in *Proceedings of the 5th NTCIR Workshop*, 2005.

Lin, J., "Evaluation of resources for question answering evaluation," in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 392-399.

Lin, S.-J., M.-S. Shia, K.-H. Lin, J.-H. Lin, S. Yu, and W.-H. Lu, "Improving answer ranking using cohesion between answer and keywords," in *NTCIR Workshop*, 2005.

Magnini, B., M. Negri, R. Prevete, and H. Tanev, "Is it the right answer?: exploiting web redundancy for Answer Validation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2001, pp. 425-432.

Molla, D. and M. Gardiner, M., "AnswerFinder — Question Answering by Combining Lexical, Syntactic and Semantic Information," in *Australasian Language Technology Workshop (ALTW) 2004*, Sydney, Australia, pp. 9-16.

Ravichandran, D. and E. Hovy, "Learning Surface Text Patterns for a Question Answering System,"in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 41-47.

Roussinov, D., J. Robles, and Y. Ding, "Experiments with Web QA System and TREC2004 Questions," in *the proceedings of TREC conference*, November, 2004, pp. 16-19.

Sacaleanu, B. and G. Neumann, "DFKI-LT at the CLEF 2006 Multiple Language Question Answering Track," in *CLEF, 2006*.

Sasaki, Y., H.H. Chen, K. Chen, and C.J. Lin, "Overview of the NTCIR-5 Cross-Lingual Question Answering Task (CLQA1)," in P*roceedings of the Fifth NTCIR Workshop Meeting*, pp. 6-9.

Sasaki, Y., C.-J. Lin, K-H. Chen, and H.-H. Chen, "Overview of the NTCIR-6 Cross-Lingual Question Answering (CLQA) Task," in *Proceedings of NTCIR-6 Workshop*, 2007, Tokyo, Japan.

Shen, D., G. Saarbruecken, and D. Klakow, "Exploring Correlation of Dependency Relation Paths for Answer Extraction," in *Proceedings of ACL 2006*, 2006, Sydney, Australia, pp. 889-896.

Soubbotin, M.M. and S.M. Soubbotin, "Patterns of Potential Answer Expressions as Clues to the Right Answers," in *Proceedings of the Tenth Text REtrieval Conference (TREC 2001*), 2001, Gaithersburg, MD, pp. 134-143.

Tom´as, D., J.e.L. Vicedo, E. Bisbal, and L. Moreno, "Experiments with LSA for Passage Re-Ranking in Question Answering," in *CLEF*, 2005.

Zhao, Y., Z.M. Xu, Y. Guan, and P. Li, "Insun05QA on QA track of TREC2005," in *TREC*, 2005, Gaithersburg, MD.

Zheng, Z., "AnswerBus Question Answering System," in *Proceeding of Human Language Technology Conference*, 2002, San Diego, CA, pp. 24-27.

Computational Linguistics and Chinese Language Processing
Vol. 13, No. 1, March 2008, pp. 27-40

Two Approaches for Multilingual Question Answering: Merging Passages vs. Merging Answers

Rita M. Aceves-Pérez*, Manuel Montes-y-Gómez*,

Luis Villaseñor-Pineda*, and L. Alfonso Ureña-López+

Abstract

One major problem in multilingual Question Answering (QA) is the integration of information obtained from different languages into one single ranked list. This paper proposes two different architectures to overcome this problem. The first one performs the information merging at passage level, whereas the second does it at answer level. In both cases, we applied a set of traditional merging strategies from cross-lingual information retrieval. Experimental results evidence the appropriateness of these merging strategies for the task of multilingual QA, as well as the advantages of multilingual QA over the traditional monolingual approach.

Keywords: Multilingual Question Answering, Cross-Lingual Information Retrieval, Information Merging.

1. Introduction

Question Answering (QA) has become a promising research field whose aim is to provide more natural access to textual information than traditional document retrieval techniques [Laurent *et al*. 2006]. In essence, a QA system is a kind of search engine that responds to natural language questions with concise and precise answers. For instance, given the question "Where is the Popocatepetl Volcano located?", a QA system has to respond "Mexico", instead of returning a list of related documents to the volcano.

* Laboratory of Language Technologies, National Institute of Astrophysics, Optics and Electronics (INAOE). Luis Enrique Erro #1, Sta. María Tonantzintla, Puebla, Mexico.
Tel.: +52-222-2663100 ext: 8218 Fax: +52-222-2663152.
The author for correspondence is Manuel Montes-y-Gómez.
Email: mmontesg@inaoep.mx

+ Department of Computer Science, University of Jaén. Campus Las Lagunillas s/n, Edif D3, Jaén, Spain

[Received June 30, 2007; Revised October 8, 2007; Accepted January 12, 2008]

At present, due to the internet explosion and the existence of several multicultural communities, one of the major challenges to face this kind of system is *multilinguality*. In a multilingual scenario, it is expected that QA systems will be able to: (*i*) answer questions formulated in several languages, and (*ii*) look for answers in a number of collections in different languages.

There are two recognizable kinds of QA systems that allow management of information in various languages: cross-lingual QA systems and, strictly speaking, *multilingual QA systems*. The former addresses a situation where questions are formulated in a different language from that of the (single) document collection. The other, in contrast, performs the search over two or more document collections in different languages.

It is important to mention that both kinds of systems have some advantages over standard monolingual QA. They mainly allow users to access more information in an easier and faster way than monolingual systems. However, they also introduce additional issues due to the language barrier.

Generally speaking, a multilingual QA system can be described as an *ensemble* of several monolingual systems, where each one works on a different – monolingual – document collection. Under this schema, two additional tasks are required: first, the translation of incoming questions into all target languages, and second, the combination of relevant information extracted from different languages.

The first problem, namely, the translation of questions from one language to another, has been widely studied in the context of cross-language QA [Aceves-Pérez *et al*. 2007; Neumann *et al*. 2005; Rosso *et al*. 2007; Sutcliffe *et al*. 2005]. In contrast, the second task, the merging of information obtained from different languages, has not been specifically addressed in QA. Nevertheless, it is important to mention that there is significant work on combining capacities from several monolingual QA systems [Chu-Carroll *et al*. 2003; Ahn *et al*. 2004; Sangoi-Pizzato *et al*. 2005], as well as on merging multilingual lists of documents for cross-lingual information retrieval applications [Lin *et al*. 2002; Savoy *et al*. 2004].

In line with these previous works, in this paper we propose *two different architectures for multilingual question answering*. These architectures differ from each other by the way they handle the combination of multilingual information. Mainly, they take advantage of the pipeline architecture of monolingual QA systems (which includes three main modules, one for question classification, one for passage retrieval, and one for answer extraction) to achieve this combination at two different stages: after the passage retrieval module by mixing together the sets of recovered passages, or after the answer extraction module by directly combining all extracted answers. In other words, our first architecture performs the combination at *passage level*, whereas the second approach does it at *answer level*. In both cases, we applied a set of

well-known strategies for information merging from cross-lingual information retrieval, specifically, Round Robin, Raw Score Value (RSV), CombSUM, and CombMNZ [Lee *et al*. 1997; Lin *et al*. 2002; Savoy *et al*. 2004].

The contributions of this paper are two-fold. On the one hand, it represents – to our knowledge – the first attempt for doing "multilingual" QA. In particular, it proposes and compares two initial solutions to the problem of multilingual information merging in QA. In addition, this paper also provides some insights on the use of traditional ranking strategies from cross-language information retrieval into the context of multilingual QA.

The rest of the paper is organized as follows. Section 2 describes some previous works on information merging. Section 3 presents the proposed architectures for multilingual QA. Section 4 describes the procedures for passage and answer merging. Section 5 shows some experimental results. Finally, section 6 presents our conclusions and outlines future work.

2. Related Work

As we previously mentioned, a multilingual QA system has to consider, in addition to the traditional modules for monolingual QA, stages for question translation and information merging.

The problem of question translation has already been widely studied. Most current approaches rest on the idea of combining capacities of several translation machines. They mainly consider the selection of the best instance from a given set of translations [Aceves-Pérez *et al*. 2007; Rosso *et al*. 2007] as well as the construction of a new question reformulation by gathering terms from all of them [Neumann *et al*. 2005; Sutcliffe *et al*. 2005; Aceves-Pérez *et al*. 2007].

On the other hand, the problem of information merging in multilingual QA has not yet been addressed. However, there is some relevant related work on constructing ensembles of monolingual QA systems. For instance, [Ahn *et al*. 2004] proposes a method that performs a number of sequential searches over different document collections. At each iteration, this method filters out or confirms the answers found in the previous step. Chu-Carroll *et al*. [2003] describes a method that applies a general ranking over the five-top answers obtained from different collections. They use a ranking function that is inspired in the well-known RSV technique from cross-language information retrieval. Finally, Sangoi-Pizzato *et al*. [2005] uses various search engines in order to extract from the Web a set of candidate answers for a given question. It also applies a general ranking over the extracted answers; nevertheless, in this case the ranking function is based on the confidence of search engines instead that on the redundancy of individual answers.

Our proposal mainly differs from previous methods in that it not only considers the

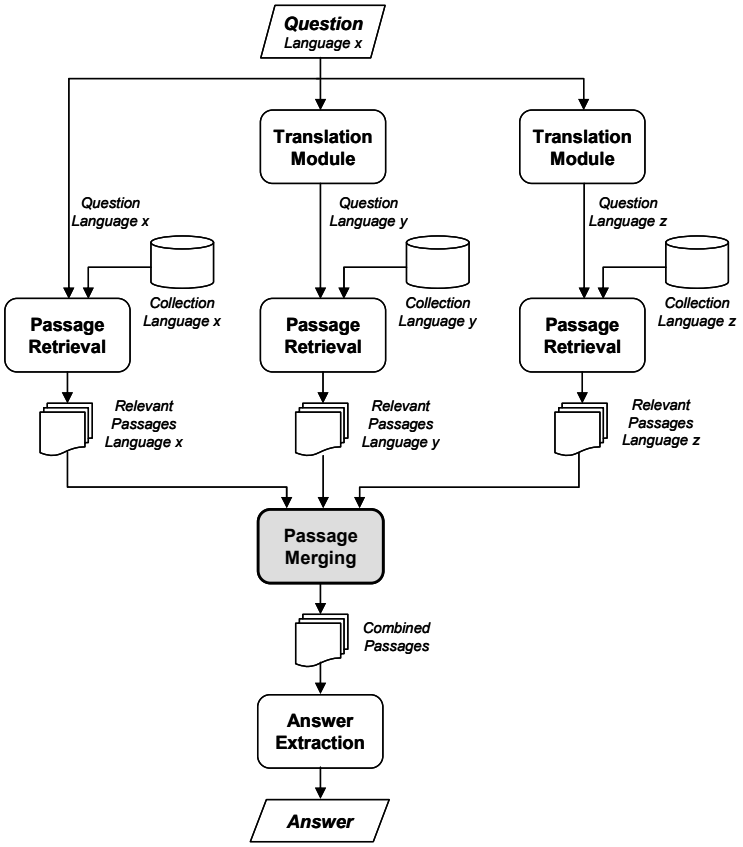

Figure 1. Multilingual QA based on passage merging

integration of answers but also takes into account the combination of passages. That is, it also proposes a method that carries out the information merging at an internal stage of the QA process. The proposed merging approach is similar in spirit to Chu-Carroll *et al*. [2003] and Sangoi-Pizzato *et al*. [2005] in that it also applies a general ranking over the information extracted from different languages. Like Chu-Carroll *et al*. [2003], it uses the RSV ranking function, although it also applies other traditional ranking strategies such as Round Robin, CombSUM and CombMNZ.

3. Two Architectures for Multilingual QA

The traditional architecture of a monolingual QA system considers three basic modules: (*i*) question classification, where the type of expected answer is determined; (*ii*) passage retrieval, where the passages with the greatest probability to contain the answer are obtained from the target document collection; and (*iii*) answer extraction, where candidate answers are ranked and the final answer recommendation of the system is produced. In addition, a multilingual QA system must include two other modules, one for question translation and another for

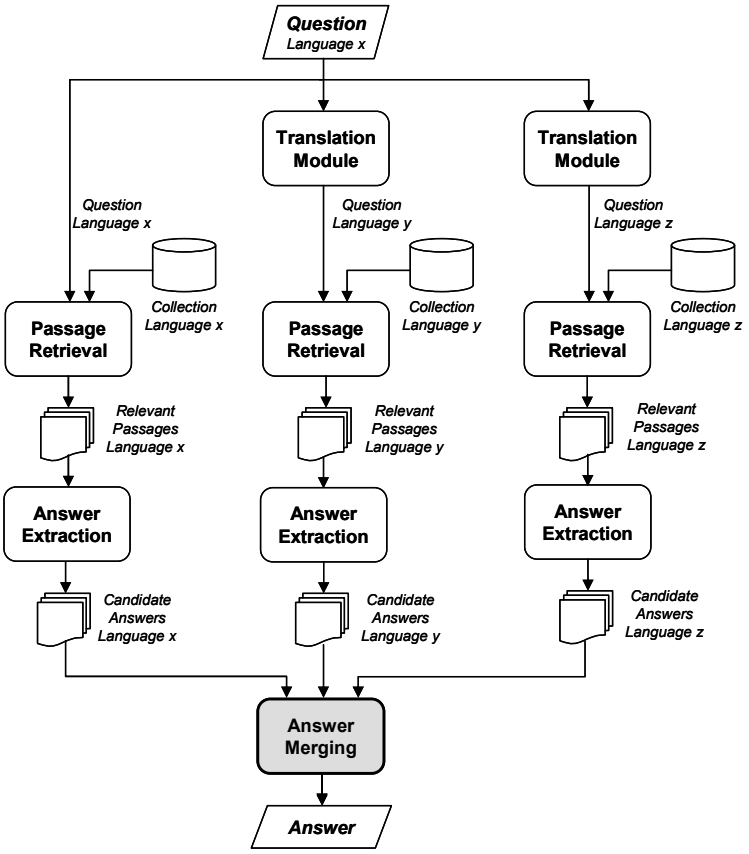

Figure 2. Multilingual QA based on answer merging

information merging. The purpose of the first module is to translate the input question to all target languages, whereas the second module is intended to integrate the information extracted from these languages into one single ranked list.

Figures 1 and 2 show two different architectures for multilingual QA. For the sake of simplicity, in both cases, we do not consider the module for question classification. On the one hand, Figure 1 shows a multilingual QA architecture that does the information merging at *passage level*. The idea of this approach is to perform in parallel the recovery of relevant passages from all collections (*i.e.*, from all different languages), then integrate these passages into one single ranked list, and then extract the answer from the combined set of passages. On the contrary, Figure 2 illustrates an architecture that achieves the information merging at *answer level*. In this case, the idea is to perform the complete QA process independently in all languages, and, after that, integrate the sets of answers into one single ranked list.

It is important to mention that merging processes normally rely on the translation of information to a common language. This translation is required for some merging strategies in order to be able to compare and rank the passages and answers extracted from different

languages.

The two proposed architectures have different advantages and disadvantages. For instance, doing the information merging at passage level commonly allows obtaining better translations for named entities (possible answers) since they are immersed in an extended context. On the other hand, doing the merging at answer level has the advantage of a clear (unambiguous) comparison of the multilingual information. In other words, comparing two answers (named entities) is a straightforward step, whereas comparing two passages requires the definition of a similarity measure and the determination of a criterion about how similar two different passages should be in order to be considered as equal. This previous problem is not present in monolingual QA ensembles, since in that case all individual QA systems search on the same document collection.

The following section introduces some of the most popular information merging strategies used in the task of cross-lingual information retrieval. It also describes the way these strategies are used within the proposed architectures for integrating passages and answers.

4. Merging Passages and Answers

4.1 Merging Strategies

Integrating information retrieved from different document collections or by different search engines is a longstanding problem in information retrieval. Researchers in this field have proposed several strategies for information merging; traditional ones are: Round Robin, RSV (Raw Score Value), CombSUM, and CombMNZ [Lee *et al*. 1997; Lin *et al.* 2002]. However, more sophisticated strategies have been proposed recently, such as the 2-step RSV [Martínez-Santiago *et al*. 2006], and the Z-score value [Savoy *et al*. 2004].

In this work, we mainly study the application of traditional merging strategies in the context of multilingual QA. The following paragraphs give a brief description of these strategies.

Round Robin. The retrieved information (in this case, passages or answers) from different languages is interleaved according to its original monolingual rank. In other words, this strategy takes one result in turn from each individual list and alternates them in order to construct the final merged output. The hypothesis underlying this strategy is the homogeneous distribution of relevant information across all languages. In our particular case, as described in Table 1, this restriction was fulfilled for almost 60% of test questions.

Raw Score Value (RSV). This strategy sorts all results (passages or answers) by their original score computed independently from each monolingual collection. Differing from Round Robin, this approach is based on the assumption that scores across different collections are comparable. Therefore, this method tends to work well when different collections are

searched by the same or very similar methods. In our experiments (refer to Section 5), this condition was fully satisfied since it was applied the same QA system for all languages.

CombSUM. In this strategy, the result scores from each language are initially (min-max) normalized. Afterward, the scores of duplicated results occurring in multiple collections are summed. In particular, we considered the implementation proposed by Lee *et al*. [1997]: we assigned a score of 21-*i* to the *i*-th ranked result from the top 20 of each language, this way, the top passage or answer was scored 20, the second one was scored 19, and so on. Any result not ranked in the top 20 was scored as 0. Finally, we added scores of duplicated results for all different monolingual runs and ranked these results in accordance to their new joint score. For instance, if an answer is ranked 3rd for one language, 10th for other one, and does not exist in a third language, then its score is (21-3) + (21-10) + 0 = 29.

CombMNZ. It is based on the same normalization as CombSUM, but also attempts to account for the value of multiple evidence by multiplying the sum of the scores (CombSUM-value) of a result by the number of monolingual collections in which it occurs. Therefore, it can be said that CombSUM is equivalent to averaging, whereas CombMNZ is equivalent to weighted averaging. Using the same example as for the CombSUM strategy, the answer's score is in this case 2 × ((21-3) + (21-10) + 0) = 58.

It is important to point out that Round Robin and RSV strategies take advantage of the complementarity among collections (when answers are extracted from only one language), whereas ComSUM and CombMNZ also take into account the redundancies of answers (the repeated occurrence of an answer in several languages).

4.2 Merging Procedures

Given several sets of relevant passages obtained from different languages, the procedure for passage merging considers the following two basic steps:

1. Translate all passages into one common language. This translation can be done by means of any translation method or online translation machine. However, we suggest translating all passages into the original question's language in order to avoid translation errors in at least one passage set.

It is important to clarify that translation is only required by the CombSUM and CombMNZ strategies. Nevertheless, all passages should be translated to one common language before entering the answer extraction module.

2. Combine the sets of passages according to a selected merging strategy. In the case of using the Round Robin or RSV approaches, the combination of passages is straightforward. In contrast, when applying CombSUM or CombMNZ, it is necessary to determine the occurrence of a given passage in two or more collections. Given that it is practically

impossible to obtain exactly the same passage from two different collections, it is necessary to define a criterion about how similar two different passages should be in order to be considered as equal. In particular, we measure the similarity of two passages by the Jaccard function (calculated as the cardinality of their vocabulary intersection divided by the cardinality of their vocabulary union) and consider them as equal only if their similarity is greater than a given specified threshold (empirically, we set the threshold value to 0.5).

The procedure for answer merging is practically the same as that for passage merging. It also includes one step for answer translation and another step for answer combination. However, the combination of answers is much simpler than the combination of passages, since they are directly comparable. In this case, the application of all merging strategies is straightforward.

5. Evaluation

5.1 Experimental Setup

The following paragraphs describe the data and tools used in the experiments.

Languages. We considered three different languages: Spanish, Italian, and French.

Search Collections. We used the document sets from the QA@CLEF evaluation forum. In particular, the Spanish collection consists of 454,045 news documents, the Italian set has 157,558, and the French one contains 129,806.

Test questions. We selected a subset of 170 factoid questions from the MultiEight corpus of CLEF. From all these questions at least one monolingual QA system could extract the correct answer. Table 1 shows answer's distributions across all languages.

Table 1. Distribution of questions by source language

	Answers in:						
	SP	**FR**	**IT**	**SP, FR**	**SP, IT**	**FR, IT**	**SP, FR, IT**
Questions	37	21	15	20	25	23	29
Percentage	21%	12%	9%	12%	15%	14%	17%

It is important to note that this set of questions covers all types of currently-evaluated factoid questions; therefore, it is possible to formulate some accurate conclusions about the appropriateness of the proposed architectures.

Monolingual QA System. We used the passage retrieval and answer extraction components of the TOVA question answering system [Montes-y-Gómez *et al*. 2005]. Its selection was mainly supported by its competence in dealing with all the considered languages. Indeed, it obtained the best precision rate for Italian and the second best for both Spanish and

French in the CLEF-2005 evaluation exercise.

Translation Machine. The translation of passages and answers was done using the Systran online translation machine (www.systranbox.com). On the other hand, questions were manually translated in order to avoid mistakes at early stages and therefore focus the evaluation on the merging phase.

Merging strategies. As we mentioned in the previous section, we applied four traditional merging strategies, namely, Round Robin, RSV, CombSUM, and CombMNZ.

Evaluation Measure. In all experiments, we used the precision as the evaluation measure. It indicates the general proportion of correctly answered questions. In order to enhance the analysis of results, we show the precision at one, three, and five positions.

Baseline. We decided to use the results from the best monolingual system (the Spanish system in this case) as a baseline. In this way, it is possible to reach conclusions about the advantages of multilingual QA over the standard monolingual approach.

5.2 Experimental Results

The objectives of the experiments were twofold: first, to compare the performance of both architectures; and second, to study the applicability and usefulness of traditional merging strategies in the problem of multilingual QA. Additionally, these experiments allowed us to analyze the advantages of multilingual QA over the traditional monolingual approach.

The first experiment considered information merging at passage level. In this case, the passages obtained from different languages were combined, and the 20 top-ranked were delivered to the answer extraction module. Table 2 shows the precision results obtained using all merging strategies as well as the precision rates of the best monolingual run.

From Table 2, it is clear that merging strategies relying on the complementarity of information (such as Round Robin and RSV) obtain better results than those also considering its redundancy (*e.g*. CombSUM and CombMNZ). We hypothesize that this behavior was mainly produced by three different factors: (*i*) the impact of translation errors on the CombSUM and CombMNZ strategies[1]; (*ii*) the complexity of assessing the redundancy of passages, *i.e*., the complexity of correctly deciding whether two different passages should be considered as equal; and (*iii*) the large number of questions (42%) that have an answer in just one language.

[1] We do not have an exact estimation of the translation errors for this task, but we suppose they are very abundant. This supposition is based on current reports from cross-lingual QA [Vallin *et al*. 2005] which indicate severe reductions – as high as 60% – in precision results as a consequence of unsatisfactory question translations.

Table 2. Precision results of the passage merging approach

Merging Strategy	*Precision at*:		
	1st	3rd	5th
Round Robin	0.41	0.57	0.65
RSV	**0.45**	**0.65**	**0.66**
CombSUM	0.40	0.54	0.64
CombMNZ	0.40	0.54	0.63
Best Monolingual	**0.45**	0.57	0.64

The second experiment achieved information merging at answer level. In this experiment, we considered the 10 top-ranked answers from each monolingual QA system. Table 3 shows the results obtained using all different merging strategies.

Table 3. Precision results of the answer merging approach

Merging Strategy	*Precision at*:		
	1st	3rd	5th
Round Robin	**0.45**	**0.68**	0.74
RSV	0.44	0.61	0.69
CombSUM	0.42	0.66	**0.75**
CombMNZ	0.42	0.62	0.70
Best Monolingual	0.45	0.57	0.64

The results of Table 3 are encouraging. They show that all merging strategies achieved high performance levels, improving baseline results at the third and fifth positions by more than 7% and 8%, respectively. Once again, these results indicate that simple strategies outperformed complex ones. However, they do not necessarily mean that Round Robin and RSV are better than CombSum and CombMNZ, instead they only express that the former methods are less sensitive to translation errors.

Comparing the results of both architectures, it is easy to observe that merging answers obtained better precision rates than merging passages. It seems that this situation is because the combination of answers is easier than the combination of passages; therefore, the first one allows to better taking advantage of both the complementarity as well as the redundancy of information. This phenomenon is more evident in the performance of CombSUM and CombMNZ; in the case of passage merging, their results were always below the baseline, and were – on average – 6% below the best precision rate, whereas, in answer merging, they were only 3% below the best result.

In addition, the fact that RSV was the best strategy for passage merging and Round Robin for answer merging shows, on the one hand, the pertinence of the passage scores against the low confidence of the answer scores, and on the other hand, the homogeneous distribution of the answers in all languages (from Table 1: 65% of the questions has an answer –at the first 20 positions– in Spanish, 55% in French and 55% in Italian).

6. Conclusions

The problem of cross-lingual QA has been widely studied; nevertheless – to our knowledge – there are no specific solutions to the related problem of multilingual QA. This paper focused on this new direction. It proposed *two different architectures for multilingual QA*. One of them performs information merging at passage level, whereas the other does it at answer level.

A secondary contribution of our work, but not necessarily less important, is the study of the *usefulness of traditional ranking strategies* from cross-language information retrieval into the context of multilingual QA.

The presented experimental results allowed us to reach the following conclusions:

A multilingual QA system may help respond to a larger number of questions than a traditional monolingual QA system. Considering that practical QA systems supply lists of candidate answers instead of isolated responses, our results demonstrated that, using a simple multilingual QA approach, it was possible to answer up to 10% more questions than using a traditional monolingual system.

Merging answers seems to be more convenient than merging passages. This assertion is mainly supported by the fact that it is more difficult to observe and compute the information redundancy at passage level than at answer level. In addition, the results of passage merging will inevitably be affected by the (quality of the) answer extraction module, whereas the results of answer merging are the actual output.

Translation errors directly affect the performance of some merging strategies. It seems that merging strategies such as CombSUM and CombMNZ are more relevant than the rest (simple ones, such as Round Robin and RSV). However, our results demonstrate that they are more sensitive to translation mistakes.

Finally, in order to improve the results of multilingual QA we plan to investigate the following issues:

1. Using different criteria to evaluate the similarity between passages. In particular, we consider that this action can have an important influence on the performance of strategies based on the information redundancy, such as CombSUM and CombMNZ.
2. Using ensemble methods for improving the translation of passages and answers. We plan to work with methods that combine the capacities of several translation machines by selecting

the best instance from a given set of translations or by constructing a new translation reformulation by gathering terms from all of them.

3. Using new merging strategies. In particular, we are considering applying graph and probabilistic based ranking techniques. We believe these kinds of techniques will help develop more robust multilingual merging strategies.

Acknowledgements

This work was done under partial support of CONACYT (Project Grant 43990), SNI-Mexico, and the Human Language Technologies Laboratory at INAOE. We also want to thanks to the CLEF organization as well as the EFE agency for the resources provided.

References

Aceves-Pérez, R., M. Montes-y-Gómez, and L. Villaseñor-Pineda, "Enhancing Cross-Language Question Answering by Combining Multiple Question Translations," In *Proceedings of the 8th Internactional Conference in Computational Linguistics and Intelligent Text Processing CICLing-2007*, 2007, Mexico City, Mexico, pp. 485-493.

Ahn, D., V. Jijkoun, K. Müller, M. de Rijke, S. Schlobach, and G. Mishne, "Making Stone Soup: Evaluating a Recall-Oriented Multi-stream Question Answering System for Dutch," In *Proceedings of the 5th Workshop of the Cross-Language Evaluation Forum CLEF 2004,* 2004, Bath, UK, pp. 423-434.

Chu-Carroll, J., K. Czuba, A. J. Prager, and A. Ittycheriah, "In Question Answering, Two Heads are Better than One," In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology HLT-NAACL 2003*, 2003, Edmonton, Canada, pp. 24-31.

Laurent, D., P. Séguéla, and S. Nègre, "QA better than IR?," In *Proceedings of the Workshop on Multilingual Question Answering MLQA-2006*, 2006, Trento, Italy, pp. 1-8.

Lee, J., "Analysis of Multiple Evidence Combination," In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1997, Philadelphia, Pennsylvania, United States, pp. 267-276.

Lin, W. C., and H. H. Chen, "Merging Mechanisms in Multilingual Information Retrieval," In *Proceedings of the Third Workshop of the Cross-Language Evaluation Forum CLEF 2002*, 2002, Rome, Italy, pp. 175-186.

Martínez-Santiago, F., L. A. Ureña-López, and M. Martín-Valdivia, "A Merging Strategy Proposal: The 2-step Retrieval Status Value Method," *Information Retrieval*, 9(1), 2006, pp. 71-93.

Montes-y-Gómez, M., L. Villaseñor-Pineda, M. Pérez-Coutiño, J. M. Gómez-Soriano, E. Sanchis-Arnal, and P. Rosso, "A Full Data-Driven System for Multiple Language Question Answering," In *Proceedings of the 6th Workshop of the Cross-Language Evalution Forum CLEF 2005*, 2005, Vienna, Austria. pp. 420-428.

Neumann, G., and B. Sacaleanu, "Experiments on Cross-Linguality and Question-Type Driven Strategy Selection for Open-Domain QA," In *Proceedings of the 6th Workshop of the Cross-Language Evalution Forum CLEF 2005*, 2005, Vienna, Austria, pp. 429-438.

Rosso, P., D. Buscaldi, and M. Iskra, "Web-based Selection of Optimal Translations of Short Queries," *Procesamiento de Lenguaje Natural*, 38, 2007, pp.49-52.

Sangoi-Pizzato, L. A., and D. Molla-Aliod, "Extracting Exact Answers using a Meta Question Answering System," In *Proceedings of the Australasian Language Technology Workshop*, 2005, Sidney, Australia, pp. 105-112.

Savoy, J., and P. Y. Berger, "Selection and Merging Strategies for Multilingual Information Retrieval," In *Proceedings of the 5th Workshop of the Cross-Language Evaluation Forum CLEF 2004*, 2004, Bath, UK, pp. 27-37.

Sutcliffe, R., M. Mulcahy, I. Gabbay, A. O'Gorman, K. White, and D. Slatter, "Cross-Language French-English Question Answering Using the DLT System at CLEF 2005," In *Proceedings of the 6th Workshop of the Cross-Language Evalution Forum CLEF 2005*, 2005, Vienna, Austria, pp. 502-509.

Vallin, A., B. Magnini, D. Giampiccolo, L. Aunimo, C. Ayache, P. Osenova, A. Peñas, M. de Rijke, B. Sacaleanu, D. Santos, and R. Sutcliffe, "Overview of the CLEF 2005 Multilingual Question Answering Track," In *Proceedings of the 6th Workshop of the Cross-Language Evalution Forum CLEF 2005*, 2005, Vienna, Austria, pp. 307-331.

Computational Linguistics and Chinese Language Processing
Vol. 13, No. 1, March 2008, pp. 41-60

Cross-Lingual News Group Recommendation Using Cluster-Based Cross-Training

Cheng-Zen Yang*, Ing-Xiang Chen*, and Ping-Jung Wu*

Abstract

Many Web news portals have provided clustered news categories for readers to browse many related news articles. However, to the best of our knowledge, they only provide monolingual services. For readers who want to find related news articles in different languages, the search process is very cumbersome. In this paper, we propose a cross-lingual news group recommendation framework using the cross-training technique to help readers find related cross-lingual news groups. The framework is studied with different implementations of SVM and Maximum Entropy models. We have conducted several experiments with news articles from Google News as the experimental data sets. From the experimental results, we find that the proposed cross-training framework can achieve accuracy improvement in most cases.

Keywords: Cross-Lingual News Group Mapping, Cross-Training, Semantic Overlapping, Mapping Recommendation

1. Introduction

As the Web becomes an abundant source of news information, it also becomes an important medium for people to learn recent tidings. To provide readers a convenient way of viewing a news event described by different news agencies, many Web news portals, such as AltaVista News and Google News, cluster news articles according to their relevance with consistent user interfaces. With such news clustering services, readers could easily acquire more details of an interesting news event from numerous reports. Ideally, they can simply click through an entry link to browse many related news reports without need of a cumbersome searching procedure. Nevertheless, if the news event is originally reported by foreign news agencies, the readers usually find that there are only few translated news articles and can only acquire an overview

* Dept. of Computer Sci. and Eng., Yuan Ze University, 135 Yuan-Tung Rd., Chungli, 320, Taiwan.
Tel.: +886-3-4638800 ext: 2361 Fax: +886-3-4638850.
E-mail: {czyang,sean,pjwu}@syslab.cse.yzu.edu.tw

[Received June 30, 2007; Revised October 8, 2007; Accepted January 12, 2008]

of the news event. If they want to find more related foreign news stories, they may generally get frustrated due to the following two reasons. First, the translated news articles seldom provide as much information as the original news articles. Second, the translation may add more interpretations that can mislead in the searching direction. The following example illustrates these situations.

This news story, reported in BBC News [2006], is a good example to show these problems. The title of its English version is "First impressions count for web" and the article contains 15 paragraphs mainly focused on the impressions in a 20th of a second after first sight [BBC News 2006]. However, the title of its Chinese news story is "好網頁還需要讓讀者一見鍾情" and may be translated into "Good web pages need to let readers fall in love at first sight", which includes additional semantic information related to love. In addition, the Chinese news article has only 7 paragraphs. When readers read the Chinese news article (the source document) and want to find more information from (for example) English news articles (target documents), they will most likely search for the news article entitled with "fall in love at first sight" and find nothing related. Apparently, the readers cannot easily find the English news article. Additionally, the amount of information of the source news article may not be equal to that of the corresponding target news article. In this example, the amount of information of the translated Chinese article is much less than that of the original English article. The scant amount of translated information will perplex the readers in other searching operations. These observations suggest the need of a cross-lingual news recommendation framework for readers to get a broader view to a news event.

To address the recommendation issue for cross-lingual news groups, the simplest approach is to directly translate the source news article and find the related news group in another language. Unfortunately, the quality of translation and the amount of news information highly influence the recommendation results. Readers may get translated results of poor quality. For instance, using Google Translation (http://www.google.com/translate_t?hl=zh-TW) to translate the Chinese news title of the above example gets "Readers need to make a good website was love at first sight". As many Web news portals have provided monolingual cluster-based news browsing interfaces, the quality of cross-lingual news group recommendation can be improved if the cluster information of the source documents is exploited. Such exploration of cluster information has been studied recently in many applications, such as Web catalog integration [Agrawal and Srikan 2001; Tsay *et al*. 2003; Sarawagi *et al*. 2003; Zhang and Lee 2004a; Zhang and Lee 2004b; Chen 2005] and title generation [Tseng *et al*. 2006].

In this paper, we propose a cross-lingual news group recommendation framework using the cross-training approach from recent Web taxonomy integration techniques [Sarawagi *et al*. 2003] to find the possible semantic corresponding relationships between news groups of

different languages. With the cross-training approach, the framework explores the implicit clustering information from the source news groups and the target news groups by learning the group features alternately. Then, the framework utilizes the implicit clustering information to improve the mapping accuracy between news groups of different languages.

Such a framework has two major advantages. First, it will save considerable news searching effort resulting from the cumbersome searching procedure in which readers need to query different monolingual news portals in a trial-and-error manner. Second, it mitigates the translation inaccuracy to provide readers a broader panorama of news events from different aspects.

The cross-training framework has been implemented in Support Vector Machines (SVM) and Maximum Entropy (ME) classifiers. We have also conducted experiments to investigate the accuracy improvement of the cross-training approach with a 21-day data set containing English and Chinese news articles collected from Google News. In the experiments, we measured the accuracy performance for different approaches. The experimental results show that the cross-training approach can benefit the mapping accuracy in most cases.

The rest of the paper is organized as follows. In Section 2, we present the problem definitions and briefly review previous related research on Web catalog integration. Section 3 elaborates the proposed cross-training framework. Section 4 describes our experiments in which English news and Chinese news articles from Google News were used as the data sets. Section 5 concludes the paper and discusses future directions.

2. Problem Statement and Related Research

For the recommendation problem of clustered news groups in different languages, we assume that the recommendation process deals with two Web news catalogs in two different languages to find the best semantically correlated relationships between the two news catalogs. We also assume that readers browse one news catalog and want to find related news articles in another news catalog of another language for the sake of simplicity. The catalog browsed by readers is the source S in which the news articles (source documents) are written in language L_s and have been classified into m event clusters $S_1, S_2,\ldots, S_m$. The other is the target catalog T in which the news articles (target documents) are written in language L_t and have been also classified into n clusters $T_1, T_2,\ldots, T_n$. The terms of the documents of each cluster comprise the feature space of the corresponding news event.

In the recommendation process, therefore, the objective of the framework is to discover all possible cluster-to-cluster mapping relationships between S and T, and report these relationships to the readers for recommendation. For the sake of simplicity in discussion, we only consider the best mapping relationships in this paper, *i.e.*, given a source catalog S_i, the

best corresponding target catalog T_j ($S_i \rightarrow T_j$) is identified in this study. Ideally, if both news clusters S_i and T_j focus on the same news event, the news articles in both clusters should have semantic overlap, as shown in Figure 1. Generally, the mapping relationships are one-to-one and symmetric. However, in our observations, one-to-many situations indeed have occurred because more than one target cluster is overlapped by the same source cluster. Furthermore, source documents will be translated in L_t first, and the quality of the feature space of the translated source documents may be hindered due to the poor translation process. These factors may make the symmetric relationships asymmetric. Therefore, the reverse mappings ($T_j \rightarrow S_i$) are separately considered.

Generally, the cluster-to-cluster mapping discovery problem can be viewed as a generalization of the Web catalog integration problem on a coarse-grained basis. In the Web catalog integration problem, the objective of the integration process is to classify the documents in the source catalog into the target catalog with the enhancement of the implicit source information. In recent years, there have been many approaches proposed for the general catalog integration problem. For example, the Naïve Bayes approaches [Agrawal and Srikan 2001; Tsay *et al*. 2003], the SVM-based approaches [Sarawagi *et al*. 2003; Zhang and Lee 2004a; Zhang and Lee 2004b; Chen 2005], and the Maximum Entropy approach [Wu *et al*. 2005] have shown that the integration improvement can be effectively achieved.

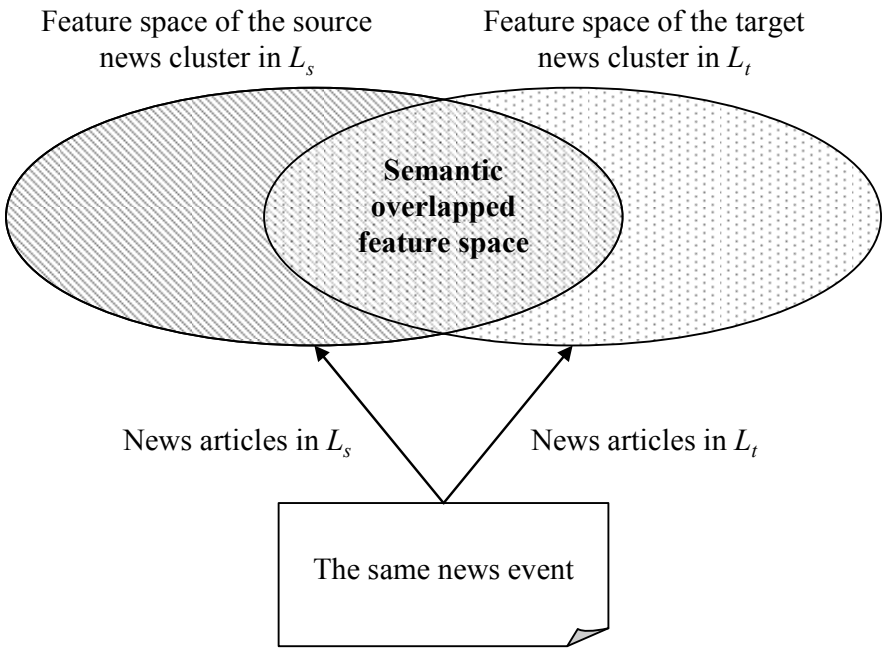

Figure 1. The relation of the news event and the correspondent news clusters in L_s and L_t.

Some enhancement approaches, however, may not be suitable for the cross-lingual cluster-to-cluster mapping discovery problem. For example, the topic restriction approach proposed in Tsay *et al*. [2003] requires that the testing target clusters are the clusters containing common documents from the source cluster. Nonetheless, in the cross-lingual cluster-to-cluster mapping discovery problem, there cannot be such a common subset. The enhanced Naïve Bayes (ENB) approach proposed in [Agrawal and Srikan 2001] exploits the implicit source catalog information to enhance the integration accuracy performance. However, due to the diversity of news articles and the translation variety, the iterative algorithm may introduce many false-positive mappings to twist the overlapped space into a larger one. The shrinkage approach adopted in Wu *et al*. [2005] also needs to be adapted because the news clusters are usually not hierarchically organized.

Our recommendation framework uses the cross-training approach adapted from the cross-training (CT) approach proposed in [Sarawagi *et al*. 2003]. The CT approach is a semi-supervised learning strategy. The idea behind CT is that a better classifier can be built with the assistance of another catalog that has semantic overlap. The overlapped document set is fully-labeled and partitioned into a development set and a test set where the development set is used to tune the system performance and the test set is used to evaluate the system. Through the cross-training process, the implicit information in the source taxonomy is learnt, and more source documents can be accurately integrated into the target taxonomy.

The proposed framework utilizes the CT approach to first obtain the potential mapping relationships from the reverse mappings ($T_j \rightarrow S_i$) through a learning process. The extracted information then is used to augment the feature space in the next learning phase. Finally, the mappings from S_i to T_j are explored in a classification process.

3. Cross-Training for Mapping Discovery

The main design principle of the cross-training framework is that the implicit mapping relationships are extracted through the first learning phase on reverse mappings. In this phase, the strength of each possible mapping is identified and ranked. For each S_i, the framework can find the most possibly corresponding T_j. Before the second learning phase, the feature space of each T_j is expanded with the discovered mapping information. Then, the augmented classifiers are used to identify the mapping relationships from S_i to T_j, and give the recommendations.

3.1 The Processing Flow

Figure 2 depicts the processing flow in the cross-training framework. Without loss of generality, we use English and Chinese here as two language representatives for L_s and L_t to explain our bilingual recommendation process in this paper.

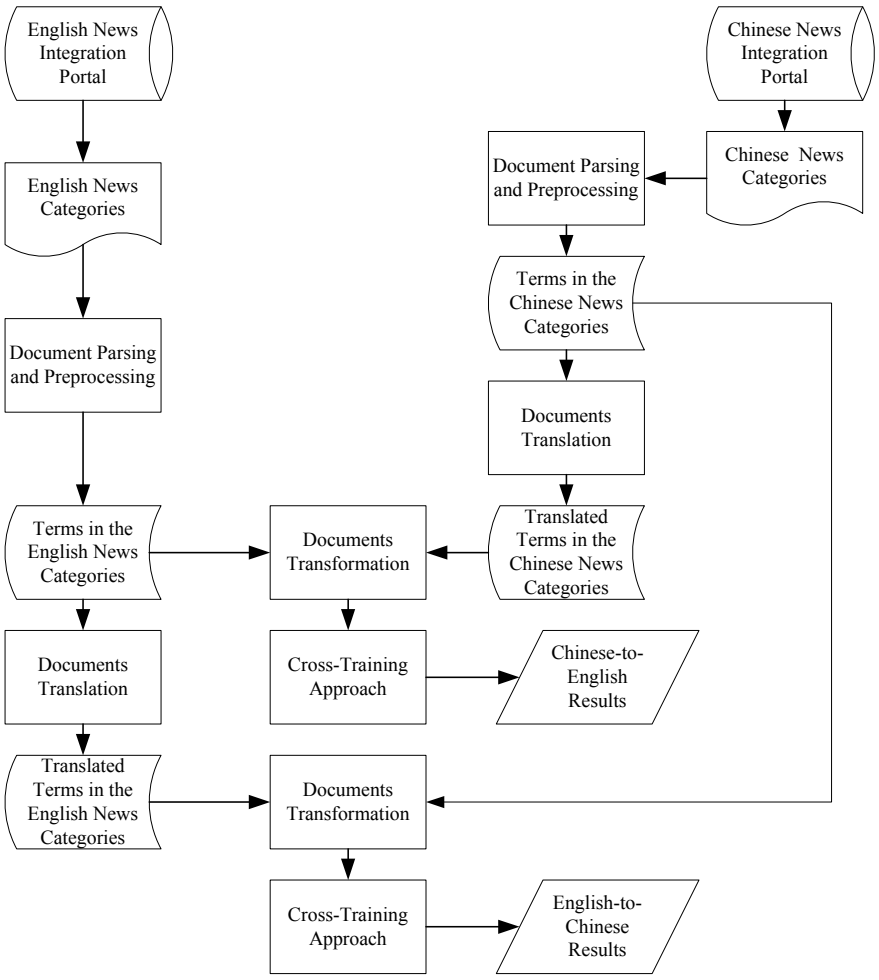

Figure 2. The processing flow for bilingual news group recommendation in the cross-training framework.

In the framework, the classification system first retrieves English and Chinese news articles from news portals, say Google News or Yahoo! News. These news articles have been usually clustered well in the news portals. The framework then performs parsing and preprocessing on each news cluster to get its feature space. The preprocessor parses the Web news, and eliminates stopwords [Fox 1992] and HTML tags. After the preprocessing, the source news groups are translated into the target language. For example, if a reader wants to find the possible English news groups for a designated Chinese news group, the English news articles are in the source news groups and will be translated into Chinese. After the translation process, all the source and target news groups are prepared as the data sets for further cross-training operations.

A debate may arise about whether the framework should re-cluster the news articles after the translation process. Since the translation process may introduce semantic variety into the news clusters, re-clustering the news articles may produce clusters with better semantic integrity for the following recommendation process. Nonetheless, the observations in Chen *et al*. [2003] show that the re-clustering process can contrarily reduce the quality of the original semantic integrity. Therefore, the proposed framework will not re-cluster the news articles.

3.2 Parsing and Preprocessing

As each Web news article is composed of plain text and HTML tags, it needs to be parsed first to extract useful information. For simplicity sake, the document parsing procedure is currently designed in a conservative manner by ignoring the HTML tags and extracting only the plain text.

Both Chinese and English news articles are then preprocessed. There are four steps for English news articles: (1) tokenization, (2) stopword removal, (3) stemming, and (4) generation of term-frequency vectors. As there is no word boundary in Chinese sentences, the Chinese articles need to be segmented first [Nie and Ren 1999; Nie *et al*. 2000; Foo and Li 2004]. We use a hybrid approach proposed by Tseng [2002], which can achieve a high precision rate and a considerably good recall rate by considering unknown words. The hybrid approach combines the longest match dictionary-based segmentation method and a statistical-based approach which is a fast keyword/key-phrase extraction algorithm. With this hybrid approach, each sentence is scanned sequentially and the longest matched words based on the dictionary entries are extracted. This process is repeated until all characters are scanned.

3.3 Translation and Transformation

After preprocessing, the Chinese and English news articles in each category are tokenized. Then, the Chinese news documents are translated. The translation can be based on a bilingual dictionary or a well-trained machine translation system. In the translation, we adopt a straightforward word expansion method. Each Chinese word is simply translated to a set of English terms listed in a bilingual dictionary or derived from a machine translation system. The same procedure is also applied to the English news articles. Currently, the translation process does not consider the word choice disambiguation problem when there are several candidates for each word. The translation quality is not further addressed using different translation technologies. Nonetheless, it can be found that the proposed cross-training approach achieves around 90% accuracy performance in top-1 ranking.

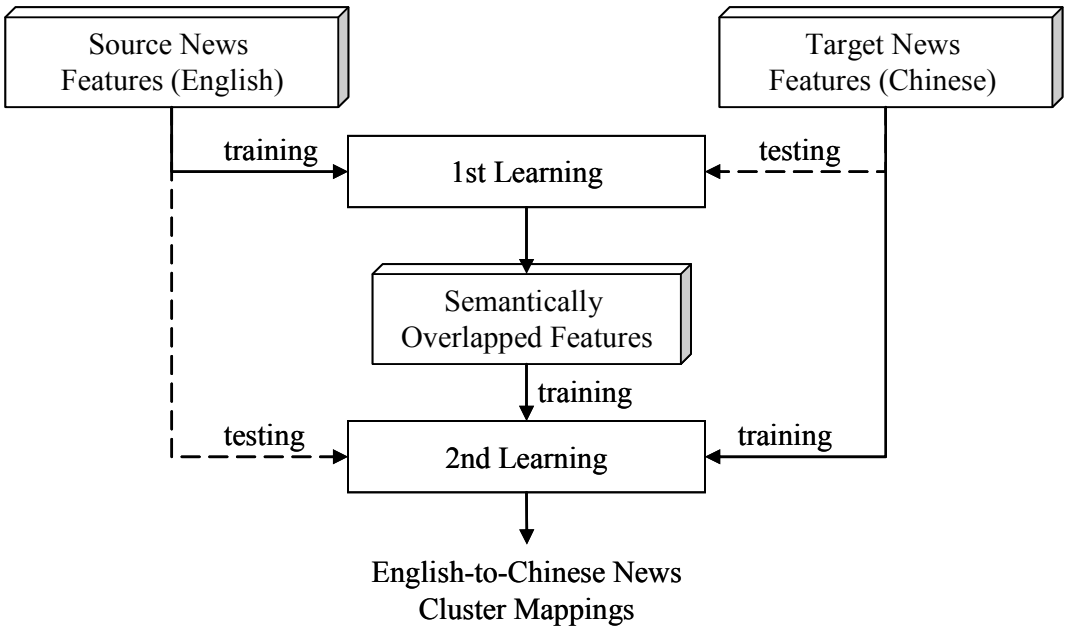

Figure 3. The basic concept of the cross-training process.

Finally, each news article is converted to a feature vector. For each index term in the feature vector, a weight is associated with the term to express the importance. In the current design, the weight of each term is calculated by $TF_x / \sum TF_i$, where i denotes the number of the stemmed terms in each news article.

3.4 The Cross-Training Process

Previous studies on the general Web catalog integration problem show that, if a source document can be integrated into a target category, there must be a sufficiently large semantic overlap between them [Agrawal and Srikan 2001; Sarawagi *et al*. 2003; Tsay *et al*. 2003; Zhang and Lee 2004a; Zhang and Lee 2004b; Wu *et al*. 2005; Yang 2006]. For the cluster-to-cluster mapping discovery problem, this observation is also an important basis. If an English news category can be associated with a Chinese news category, this mapping must be concluded from a situation in which the semantically overlapped feature space is sufficiently large.

3.4.1 Learning to Extract the Implicit Information

The cross-training process is incorporated mainly for exploring the overlapped feature space. Figure 3 illustrates a cross-training process in which there are two learning phases. In the first phase, the source news clusters are used as the training data sets to train m classifiers, and the target news clusters are used as the testing data sets to extract the implicit mapping

information. The m classifiers then calculate the mapping scores (Sc_{ij}) for n target news clusters to predict the strengths of the semantic overlaps.

Since SVM and ME are studied in the framework implementations, the mapping score Sc_{ij} of $T_j \rightarrow S_i$ can be defined as either the ratio at which the target documents in T_j are classified into the source news cluster S_i or the average weight derived from the classifier. For example, if the classification scheme used in the framework is SVM, the mapping score Sc_{ij} can be calculated by either Eq. (1) where N_{Tj} is the news documents of the target cluster T_j or Eq. (2) which is the average of the distance from each document to the hyperplane. This average can be viewed as the discriminative characteristic of all documents to the classifier.

$$Sc_{ij} = \frac{\#\text{of } N_{Tj} \text{ classified in } S_i}{\#\text{of } N_{Tj}} \tag{1}$$

$$Sc_{ij} = \frac{\sum w_i x_i + b}{\#\text{of } N_{Tj}} \tag{2}$$

Basically, Equation (1) represents a voting scheme in which the predicted rank of a target cluster T_j depends on the number of the positively classified news articles in T_j. Equation (2) represents a weighting scheme in which the predicted rank of T_j depends on the average of the total distance to the hyperplane. For each source cluster, the target cluster with the highest mapping score is qualified as the potential candidate that may have the accurate $S_i \rightarrow T_j$ mapping relationship in the second learning phase. The reason the mapping scores are considered in an asymmetric way is that the cross-training approach will adjust the feature vectors back and forth in each learning iteration. Other mapping discovery approaches may provide efficient schemes to consider both mapping scores of $S_i \rightarrow T_j$ and $T_j \rightarrow S_i$ as an integrated scoring method. This has been left for our future study.

3.4.2 Learning to Find the Corresponding Mappings

The implicit information explored in the first learning phase is then used as the prediction information in the second learning phase. The cross training process can be continued until the results converge. The category information of the corresponding source cluster, say S_i, for the previously discovered candidate target cluster, say T_j, is inserted into the feature space of T_j. The category information can be category identifiers or the category title words. For example, we used identifiers starting from 1000001 to 1000040 for categories in the current experiments.

Figure 4 depicts the detailed process of concatenating the predicted implicit information to the ordinary feature vectors of the target cluster in the cross-training approach. In the figure, F_T is a feature vector for the term features of the target news articles, L_T is a feature vector for the label features (category information) of the target cluster, and the *test output* contains the

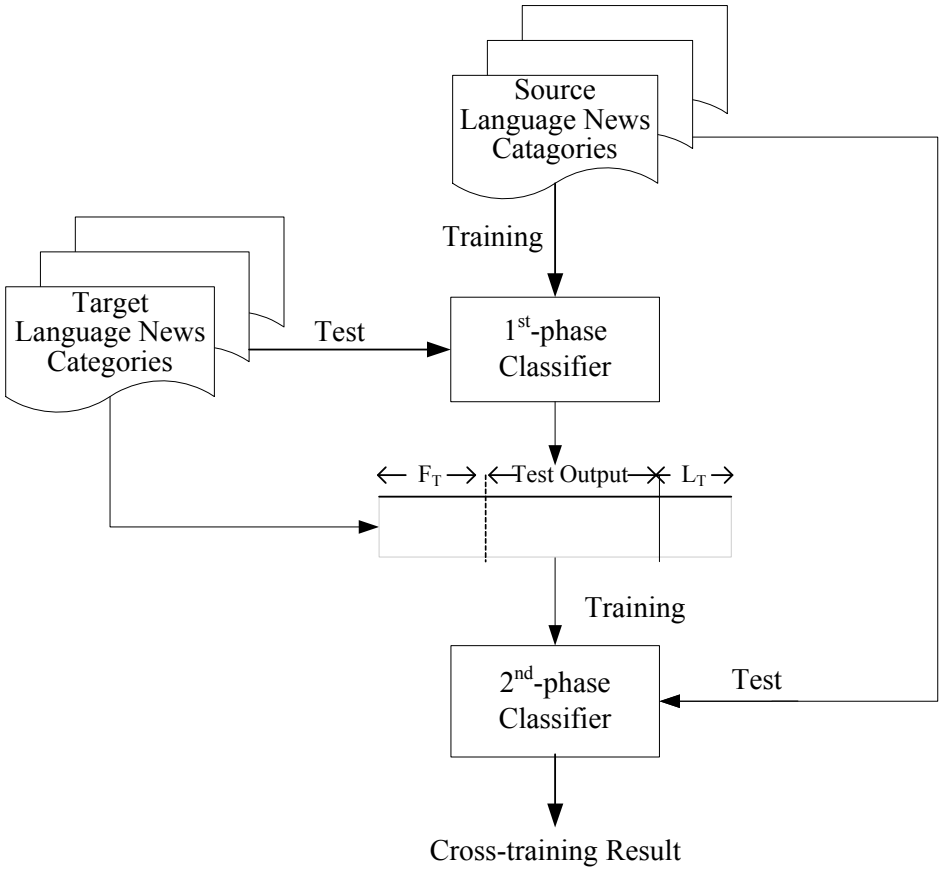

Figure 4. Adding the predicted implicit information in the cross-training process.

label features of the predicted source clusters. With the predicted mapping information, the discriminative power of the classifiers of the second phase can be enhanced.

For controlling the discriminative power of the added semantically-overlapped implicit label information, as in Sarawagi *et al*. [2003], the ordinary feature weights in the augmented target vectors are scaled by a factor of f, and the weight of each label attribute by a factor of $1 - f$. The parameter f is used to decide the relative weights of the label and term features and can be tuned for different application environments. In the current experiments, the results show that the best f value ranges from 0.02 to 0.05. The small f values show that the augmented information should not be overemphasized in the cross-training process. This observation for factoring is consistent with previous studies [Sarawagi *et al*. 2003; Chen *et al*. 2004].

Finally, the second-phase classifiers are trained with the augmented target vectors. The recommended source news groups of the target news groups are calculated using the same mapping scoring method.

4. Experiments

We have implemented the cross-training framework in SVM and ME classifiers. To rank the predictive corresponding target clusters, we implemented the voting scheme in the cross-training framework of SVM (SVM-VCT) and ME (ME-VCT), and the weighting scheme with SVM (SVM-WCT). As stated in Section 3.4.1, Equation (1) was used to rank the target clusters in SVM-VCT and ME-VCT. Equation (2) was used in the weighting scheme SVM-WCT. We also implemented the voting scheme and the weighting scheme in SVM (SVM-V and SVM-W) for comparison. In the experiments, an English news catalog and a Chinese news catalog from Google News were used as the representatives to demonstrate the classification performance of the proposed cross-training framework. We measured the accuracy performance at top-1, top-3, and top-5 ranks. The details of the experiments are presented as follows.

4.1 The Experimental Environment

The framework is currently implemented in Java. The segmentation corpus is based on the Academia Sinica Bilingual Wordnet 1.0 published by the Association for Computational Linguistics and Chinese Language Processing (ACLCLP) [Sinica BOW 2005]. We used SVM^{light} (version 5.00) [Joachims 2002] as the SVM tool with a linear kernel, and the maximum-entropy toolkit (version 20041229) [Zhang 2004] as the Maximum Entropy model kernel.

The bilingual word lists published by Linguistic Data Consortium (LDC) were used as the bilingual dictionaries. The Chinese-to-English dictionary ver. 2 (ldc2ce) has about 120,000 records, and the English-to-Chinese dictionary (ldc2ec) has about 110,000 records. In ldc2ce and ldc2ec, each entry is composed of a single word and several translated words separated by slashes without any indication of the importance. Therefore, the translated words are treated equally in our experiments. In the translation, each word in the source document was replaced with these translated words. The translation quality issue is not addressed in depth because we want to follow the normal reader behaviors. Furthermore, the implicit semantic information embedded in each category of news articles may mitigate the poorness of the translation.

4.2 Data Sets

In our experiments, two news portals were chosen as the bilingual news sources: Google News U.S. version for English news and Google News Taiwan version for Chinese news. Both the Chinese and English news articles were retrieved from the world news category from May 10 to May 23, 2005 and from October 21 to October 27, 2007. The experiments were performed on the data set of each day. Twenty news categories were collected per day. All the English

news articles were translated into Chinese with the bilingual dictionaries. The size of the English-to-Chinese data set is 454.5 Mbytes. All the Chinese news articles were also translated. The size of the Chinese-to-English data set is 341.5 Mbytes. The 21-day data sets contain 36,548 English news articles and 8,224 Chinese news articles.

In the experiments, the mapping relations between the Chinese and English news reports were first identified by three graduate students manually and independently. The mapping between an English news category and a Chinese news category is recognized if at least two students have the same mapping identification. These manually-identified mapping relations were used to evaluate the accuracy performance of the bilingual classification systems. We found that there were 122 identified mappings in the Chinese-to-English recommendation task and 123 identified mappings in the English-to-Chinese recommendation task. The difference existed because an English category was identified that was to be mapped to two Chinese categories. The data sets collected currently cannot significantly reveal the influences of one-to-many situations. In our future work plan, more news categories need to be collected to verify our scheme for one-to-many cases.

The experiments were conducted in two ways: finding the related Chinese news groups from the English news groups (Chinese-to-English) and finding the related English news groups from the Chinese news groups (English-to-Chinese). Here, we take the Chinese-to-English recommendation process as the example to present the experimental details. The English-to-Chinese recommendation process was conducted in a similar manner.

In the Chinese-to-English experiments, each Chinese news catalog was first used as the training set in the first learning phase. To find a corresponding Chinese category (S_i) of an English target category (T_j), the news articles in S_i were all used as the positive training examples, and the news articles in the other Chinese news categories (S_k, $k \neq i$) were randomly selected as the negative training examples. Then, all mapping scores between English categories and Chinese categories were measured based on the first-phase classification results. The English category with the highest mapping score was considered as the possibly mapped category.

In the second learning phase of the Chinese-to-English experiments, the category information of the previously identified English cluster was concatenated to the corresponding Chinese cluster. Then, the English categories were used as the training set to train the second-phase classifiers. The augmented source Chinese categories were classified to calculate the mapping scores for each English news category. Finally, we measured the accuracy performance for each day using the correct mappings at the top-1, top-3, and top-5 recommendation ranks by the following equation:

$$\text{Accuracy} = \frac{\text{Number of the correctly discovered mapping in } S \to T}{\text{Total number of the correct mapping in } S \to T}, \quad (3)$$

which is similar to Agrawal and Srikan [2001]. Accuracy, rather than precision or recall, is used because the recommendation process is performed on a cluster-to-cluster basis. The error rate is the complement of the accuracy. In the English-to-Chinese experiments, the roles of two catalogs were switched.

4.3 Results and Discussion

Table 1. Experimental results of the correctly discovered Chinese-to-English mappings in the top-1 recommendation lists.

Day	Tagged Mappings	SVM-V	SVM-VCT	SVM-W	SVM-WCT	ME	ME-VCT
1	7	0	6	5	5	4	7
2	6	1	6	6	4	3	6
3	6	0	5	4	5	3	6
4	6	1	6	6	5	5	6
5	6	0	6	6	5	3	5
6	6	2	3	6	6	3	3
7	3	1	3	1	0	1	3
8	6	2	5	6	3	2	6
9	5	1	5	3	4	1	5
10	5	1	5	4	4	2	3
11	4	0	4	2	2	2	3
12	7	3	7	4	5	2	7
13	4	3	4	4	3	4	4
14	8	5	7	7	6	4	5
15	10	5	10	10	9	9	9
16	8	2	8	7	7	6	7
17	5	1	5	4	4	4	5
18	6	1	6	4	6	5	6
19	2	0	2	1	1	2	2
20	5	2	5	5	4	5	5
21	7	1	7	7	5	5	7
Total	122	32	115	102	93	75	110
Avg. Acc.		26.23%	94.26%	83.61%	76.23%	61.48%	90.16%

Table 1 lists the experimental results of the correctly discovered Chinese-to-English mappings at the top-1 recommendation lists identified by different approaches. Table 2 lists the correctly discovered Chinese-to-English mappings at the top-3 recommendation lists. From these tables, we can notice that the cross-training approach significantly improves the voting approaches in SVM-V and ME to find correct mappings in the top-1 recommendation results. In addition, it improves SVM-V, SVM-W, and ME entirely to find correct mappings in the top-3 recommendation results. Here, the scaling factor f is 0.05. When f ranged from 0.02 to 0.05, we attained similar results.

Table 2. Experimental results of the correctly discovered Chinese-to-English mappings in the top-3 recommendation lists.

Day	Tagged Mappings	SVM-V	SVM-VCT	SVM-W	SVM-WCT	ME	ME-VCT
1	7	0	6	7	7	5	7
2	6	1	6	6	6	6	6
3	6	0	5	6	6	5	6
4	6	1	6	6	6	5	6
5	6	0	6	6	6	4	6
6	6	3	3	6	6	4	4
7	3	1	3	3	3	1	3
8	6	2	5	6	6	3	6
9	5	2	5	4	5	3	5
10	5	1	5	5	5	3	3
11	4	1	4	3	3	3	3
12	7	3	7	7	7	5	7
13	4	3	4	4	4	4	4
14	8	5	7	8	8	7	7
15	10	5	10	10	10	9	9
16	8	2	8	7	7	6	7
17	5	1	5	5	5	4	5
18	6	1	6	5	6	5	6
19	2	0	2	2	2	2	2
20	5	2	5	5	5	5	5
21	7	1	7	7	7	6	7
Total	122	35	115	118	120	95	114
Avg. Acc.		28.69%	94.26%	96.72%	98.36%	77.87%	93.44%

From Table 1, it is noticeable that SVM-W outperformed SVM-WCT. The reason the cross-training approach cannot benefit the accuracy performance is because adding more features changes the characteristics of the hyperplanes learned by SVM, thereby affecting the distance summation results in Eq. (2). Therefore, some correct mappings were ranked at the second rank in the recommendation lists for SVM-WCT but at the top rank for SVM-W. For the top-3 recommendation lists as shown in Table 2, SVM-WCT outperformed SVM-W and got the best accuracy performance.

Table 3. Experimental results of the correctly discovered English-to-Chinese mappings in the top-1 recommendation lists.

Day	Tagged Mappings	SVM-V	SVM-VCT	SVM-W	SVM-WCT	ME	ME-VCT
1	7	0	4	7	5	6	5
2	6	3	6	6	4	6	6
3	6	0	5	6	6	6	5
4	6	0	4	6	5	5	6
5	6	0	2	5	5	6	6
6	6	2	5	5	5	4	6
7	3	0	2	2	2	3	2
8	6	1	5	6	3	5	5
9	5	1	3	4	3	3	4
10	5	1	4	5	4	2	5
11	4	1	4	2	1	1	2
12	7	2	7	6	5	5	6
13	4	0	3	3	3	3	4
14	8	1	8	7	6	7	7
15	10	5	10	9	10	8	10
16	8	0	6	6	4	5	5
17	5	2	5	4	5	2	2
18	7	6	7	7	7	6	6
19	2	1	2	2	1	2	2
20	5	2	5	5	5	3	5
21	7	3	7	7	6	5	7
Total	123	31	104	110	95	93	106
Avg. Acc.		25.20%	84.55%	89.43%	77.24%	75.61%	86.18%

Table 3 and Table 4 list the experimental results of the correct English-to-Chinese mappings in the top-1 and top-3 recommendation lists, respectively. From these two tables, we can see that the cross-training approach significantly improved SVM-V and ME in finding the correct mappings in the top-1 recommendation results. From the top-3 recommendation results, we can observe that SVM-V is highly improved by the cross-training approach. SVM-W and SVM-WCT has the same results and both achieve the best performance. Although the cross-training approach cannot benefit ME more in the top-3 results as in the Chinese-to-English experiments, the performance of ME-VCT is comparable to ME.

Table 4. Experimental results of the correctly discovered English-to-Chinese mappings in the top-3 recommendation lists.

Day	Tagged Mappings	SVM-V	SVM-VCT	SVM-W	SVM-WCT	ME	ME-VCT
1	7	0	4	7	7	7	6
2	6	3	6	6	6	6	6
3	6	0	5	6	6	6	6
4	6	0	4	6	5	6	6
5	6	0	2	5	6	6	6
6	6	2	5	6	6	5	6
7	3	0	2	3	3	3	2
8	6	2	6	6	6	6	5
9	5	1	3	4	4	3	4
10	5	1	4	5	5	5	5
11	4	1	4	3	3	2	3
12	7	2	7	6	6	6	6
13	4	1	3	4	4	4	4
14	8	1	8	8	8	8	8
15	10	5	10	10	10	10	10
16	8	0	6	7	7	7	6
17	5	2	5	5	5	5	3
18	7	6	7	7	7	6	6
19	2	1	2	2	2	2	2
20	5	2	5	5	5	5	5
21	7	6	7	7	7	7	7
Total	123	36	105	118	118	115	112
Avg. Acc.		29.27%	85.37%	95.93%	95.93%	93.50%	91.06%

Table 5 lists the experimental results of the correct Chinese-to-English and English-to-Chinese mappings in the top-5 recommendation lists. Here, we omit the details of the correct mappings of each day and only show the total results. The top-5 results are very similar to the top-3 results.

Table 5. Experimental results of the correctly discovered Chinese-to-English and English-to-Chinese mappings in the top-5 recommendation lists.

(a) Results of Chinese-to-English mappings

Day	Tagged Mappings	SVM-V	SVM-VCT	SVM-W	SVM-WCT	ME	ME-VCT
Total	122	37	115	119	120	103	117
Avg. Acc.		30.33%	94.26%	97.54%	98.36%	84.43%	95.90%

(b) Results of English-to-Chinese mappings

Day	Tagged Mappings	SVM-V	SVM-VCT	SVM-W	SVM-WCT	ME	ME-VCT
Total	123	36	105	120	120	117	113
Avg. Acc.		29.27%	85.37%	97.56%	97.56%	95.12%	91.87%

Other improvements can still be introduced in the recommendation framework. For example, unknown name entity recognition (NER) and transliteration processing are two important issues for cross-lingual processing. Improvements to the quality of machine translation in the framework should further enhance the accuracy performance.

5. Conclusion

As the amount of news information explosively grows over the Internet, on-line Web news services have played an important role in delivering news information to people. Although these Web news portals have provided readers with clustered monolingual news services, cross-lingual news clustering services are still in great demand.

In this paper, we propose a cross-lingual news group recommendation framework with the cross-training approach to get high accuracy performance in finding the mapping relationships between two news catalogs in different languages. From the experimental results, we can find that the proposed cross-training recommendation framework comprehensively has the superior accuracy performance. Among all approaches, SVM-WCT can achieve the best accuracy in the top-3 and top-5 recommendation lists for both Chinese-to-English and English-to-Chinese.

There are still many research issues left for our future study. For example, feature weighting plays an important role in system performance. Meaningful features should be explored and employed for integration. In addition, we only consider the accuracy rate of correct mappings in current experiments. The correct rejection rate needs to be further studied for independent source/target categories. Furthermore, the scoring method can be discussed to find whether there are other better approaches to discover the correct mapping. In addition, a filtering scheme needs be discussed to screen out incorrect mapping recommendations (negative mappings) for practical use. One of the most challenging issues is how to translate new words which are created daily due to the rapidly changing Web. A better automatic bilingual translation system is needed to fulfill the requirements of effective term translation for the NER problem and the transliteration problem.

Acknowledgement

The authors would like to thank the National Science Council of the Republic of China, Taiwan, for partially supporting this research under Contract No. NSC 95-2745-E-155-008. The authors would also like to express many thanks to the anonymous reviewers for their precious suggestions for this paper.

References

Agrawal, R. and R. Srikan, "On Integrating Catalogs," in *Proceedings of the 10th International Conference on World Wide Web*, 2001, pp. 603-612.

Chen, H.-H., J.-J. Kuo, and T.-C. Su, "Clustering and Visualization in a Multi-lingual Multidocument Summarization System," in *Proceedings of 25th European Conference on Information Retrieval Research*, 2003, pp. 266-280.

Chen, I.-X., C.-H. Shih, and C.-Z. Yang, "Web Catalog Integration using Support Vector Machines," in *Proceedings of the 1st Workshop on Intelligent Web Technology (IWT 2004)*, Taipei, Taiwan, 2004, pp. 7-13.

Chen, I.-X., J.-C. Ho, and C.-Z. Yang, "An Iterative Approach for Web Catalog Integration with Support Vector Machines," in *Proceedings of 2nd Asia Information Retrieval Symposium (AIRS 2005)*, 2005, pp. 703-708.

Foo, S. and H. Li, "Chinese Word Segmentation and Its Effect on Information Retrieval," *Information Processing and Management*, 40(1), 2004, pp. 161-190.

Fox, C., "Lexical Analysis and Stop Lists", *Information Retrieval: Data Structures and Algorithms*, Chapter 7, Frakes, W. and Baeza-Yates, R., (eds.), Prentice-Hall, 1992, pp. 102-130.

Nie, J.Y. and F. Ren, "Chinese Information Retrieval: Using Characters or Words," *Information Processing and Management*, 35(4), 1999, pp. 443-162.

Nie, J.Y., J. Gao, J. Zhang, and M. Zhou, "On the Use of Words and N-grams for Chinese Information Retrieval," in *Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages*, 2000, pp. 141-148.

Sarawagi, S., S. Chakrabarti, and S. Godbole, "Cross-training: Learning Probabilistic Mappings between Topics," in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003, pp. 177-186.

Sinica BOW, The Academia Sinica Bilingual Wordnet. Ver. 1.0, The Association for Computational Linguistics and Chinese Language Processing, 2005.

Tsay, J.-J., H.-Y. Chen, C.-F. Chang, and C.-H. Lin, "Enhancing Techniques for Efficient Topic Hierarchy Integration," in *Proceedings of the 3rd International Conference on Data Mining (ICDM'03)*, 2003, pp. 657-660.

Tseng, Y.-H., "Automatic Thesaurus Generation for Chinese Documents," *Journal of the American Society for Information Science and Technology*, 53(13), 2002, pp. 1130-1138.

Tseng, Y.-H., C.-J. Lin, H.-H. Chen, and Y.-I Lin, "Toward Generic Title Generation for Clustered Documents," in *Proceedings of the 3rd Asia Information Retrieval Symposium (AIRS 2006)*, 2006, Singapore, pp. 145-157.

Wu, C. W., T. H. Tsai, and W. L. Hsu, "Learning to Integrate Web Taxonomies with Fine-Grained Relations: A Case Study Using Maximum Entropy Model," in *Proceedings of 2nd Asia Information Retrieval Symposium (AIRS 2005)*, 2005, pp. 190-205.

Yang, C.-Z., C.-M. Chen, and I.-X. Chen, "A Cross-Lingual Framework for Web News Taxonomy Integration," in *Proceedings of the 3rd Asia Information Retrieval Symposium (AIRS 2006)*, 2006, Singapore, pp. 270-283.

Zhang, D. and W. S. Lee, "Web Taxonomy Integration using Support Vector Machines," in *Proceedings of the 13th International Conference on World Wide Web*, 2004a, pp. 472-481.

Zhang, D. and W. S. Lee, "Web Taxonomy Integration through Co-Bootstrapping," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004b, pp. 410-417.

Online Resources

Altavista News, http://www.altavista.com/news/default.

BBC News, "First impressions count for web." English version is available at http://news.bbc.co.uk/2/hi/technology/4616700.stm; Chinese version is available at http://news.bbc.co.uk/chinese/trad/hi/newsid_4610000/newsid_4618500/4618552.stm, 2006.

Google News, http://news.google.com/.

Google Translation, http://www.google.com/translate_t?hl=zh-TW.

Joachims, T., SVM^{light}, version 5.0, http://svmlight.joachims.org/, 2002.

Linguistic Data Consortium, http://projects.ldc.upenn.edu/Chinese/LDCch.htm.

Yahoo! News, http://news.yahoo.com/.

Zhang, L., The Maximum Entropy model toolkit, version 20041229, http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html, 2004.

Computational Linguistics and Chinese Language Processing
Vol. 13, No. 1, March 2008, pp. 61-90

Web-Based Query Translation for English-Chinese CLIR

Chengye Lu*, Yue Xu*, and Shlomo Geva*

Abstract

Dictionary-based translation is a traditional approach in use by cross-language information retrieval systems. However, significant performance degradation is often observed when queries contain words that do not appear in the dictionary. This is called the Out of Vocabulary (OOV) problem. In recent years, Web mining has been shown to be one of the effective approaches for solving this problem. However, the questions of how to extract Multiword Lexical Units (MLUs) from the Web content and how to select the correct translations from the extracted candidate MLUs are still two difficult problems in Web mining based automated translation approaches.

Most statistical approaches to MLU extraction rely on statistical information extracted from huge corpora. In the case of using Web mining techniques for automated translations, these approaches do not perform well because the size of the corpus is usually too small and statistical approaches that rely on a large sample can become unreliable. In this paper, we present a new Chinese term measurement and a new Chinese MLU extraction process that work well on small corpora. We also present our approach to the selection of MLUs in a more accurate manner. Our experiments show marked improvement in translation accuracy over other commonly used approaches.

Keywords: Cross-Language Information Retrieval, CLIR, Query Translation, Web Mining, OOV Problem, Term Extraction

1. INTRODUCTION

As more and more documents written in various languages become available on the Internet, users increasingly wish to explore documents that were written in either their native language

* Faculty of Information Technology, School of Software Engineering and Data Communications, Queensland University of Technology, Brisbane, QLD 4001, Australia
E-mail: {c.lu,yue.xu,s.geva}@qut.edu.au

[Received June 30, 2007; Revised October 8, 2007; Accepted January 12, 2008]

or some language other than English. Cross-language information retrieval (CLIR) systems allow users to retrieve documents written in more than one language through queries written in a different language. This is a helpful end-user feature. Obviously, translation is needed in the CLIR process; either translating the query into the document language, or translating the documents into the query language. The common approach is to translate the query into the document language using a dictionary. Dictionary-based translation has been adopted in cross-language information retrieval because bilingual dictionaries are widely available, dictionary-based approaches are easy to implement, and the efficiency of word translation with a dictionary is high. However, due to the vocabulary limitation of dictionaries, very often the translations of some words in a query cannot be found in a dictionary. This problem is called the Out of Vocabulary (OOV) problem. Very often, the OOV terms are proper names or newly created words. Even using the best dictionary, the OOV problem is unavoidable. As input queries are usually short, query expansion does not provide enough information to help recover the missing words. Furthermore, in many cases, it is exactly the OOV terms that are the crucial words in the query. For example, a query "***SARS, CHINA"*** may be entered by a user in order to find information about ***SARS*** in China. However, ***SARS*** is a newly created term and may not be included in a dictionary published only a few years ago. If the word ***SARS*** is left out of the translated query, it is most likely that the user will be unable to find any relevant documents. Moreover, a phrase cannot always be translated by translating each individual word in the phrase. For example, an idiom is a phrase and should not be translated by combining translations of the individual words because the correct translation may be a specific word which is not the combination of individual word translations of the original phrase.

Another problem with the dictionary-based translation approach is the translation disambiguation problem. The problem is more serious for a language which does not have word boundaries, such as Chinese. Translation disambiguation refers to finding the most appropriate translation from several choices in the dictionary. For example, the English word STRING has over 20 different translations in Chinese, according to the Kingsoft online dictionary (www.kingsoft.com). One approach is to select the most likely translation [Eijk 1993] – usually the first one offered by a dictionary. However, even if the choices are ordered based on some criteria and the most likely *a-priori* translation is picked, in general, such an approach has a low probability of success. Another solution is to use all possible translations in the query with the OR operator. However, while this approach is likely to include the correct translation, it also introduces noise into the query. This can lead to the retrieval of many irrelevant documents which is, of course, undesirable. [Jang *et al.* 1999] and [Gao *et al.* 2001] report that this approach has precision that is 50% lower than the precision that is obtained by human translation.

In this paper, we present a Web-based approach to term extraction and translation selection. Specifically, we introduce a statistics-based approach to extracting terms and a translation disambiguation technique to improve the precision of the translations. The remainder of this paper is structured as follows: in Section 2, we present the existing approaches to query translation; in Section 3 we present our approach. Experimental evaluation and results discussion are presented in Section 4 and Section 5, respectively. Finally, we conclude the paper in Section 6.

2. PREVIOUS WORK

2.1 Translation

Dictionary-based query translation is one of the conventional approaches in CLIR. The appearance of OOV terms is one of the main difficulties arising with this approach. In very early years, OOV terms were not translated at all, leaving out the original terms in the translated query. This approach may significantly limit retrieval performance. In this section, several existing approaches to OOV translation are reviewed.

2.1.1 Transliteration

Proper names, such as personal names and place names, are a major source of OOV terms because many dictionaries do not include such terms. It is common for foreign names to be translated word-by-word based on phonetic pronunciations. In this manner, a name in one language will be pronounced similarly in another language – this is called transliteration. Such translation is usually done by a human when a new proper name is introduced from one language to another language.

Some researchers [Paola *et al*. 2003; Yan *et al*. 2003] have applied the rule of transliteration to automatically translate proper names. Basically, the transliteration will first convert words in one language into phonetic symbols, then convert the phonetic symbols into another language. Some researchers have found that transliteration is quite useful in proper name translation [Paola *et al*. 2003; Yan *et al*. 2003]. However, transliteration is useful in only a few language pairs. When dealing with language pairs for which there are many phonemes in one language that are not present in the other, such as Chinese and English, the problem is exacerbated. There are even more problems when translating English to Chinese. First, as there is no standard for name translation in Chinese, different communities may translate a name in different ways. For example, the word “Disney” is translated as “迪斯尼” in mainland China but is translated as“迪士尼” in Taiwan. Both are pronounced similarly in Chinese, but use different Chinese characters. Even a human interpreter would have difficulty in unambiguously choosing which character should be used. Second, at times, the Chinese

translation only uses some of the phonemes of the English names. For example, the translation of "American" is "美国" which only uses the second syllable of "American". Finally, the translation of a name is not limited to only using translation but also to transliteration. Sometimes, the translation of a proper name may even use a mixed form of translation and transliteration. For example, the translation of "New Zealand" in mainland China is "新西兰", where "新" is the translation of "New" and "西兰" is the transliteration of "Zealand".

2.1.2 Parallel Text Mining

Parallel text is a text in one language together with its translation in another language. The typical way to use parallel texts is to generate translation equivalence automatically, without using a dictionary. It has been used in several studies [Eijk 1993; Kupiec 1993; Smadja *et al*. 1996; Nie *et al*. 1999] on multilingual related tasks such as machine translation or CLIR.

The idea of parallel text mining is straightforward. Since parallel texts are texts in two languages, it should be possible to identify corresponding sentences in two languages. When the corresponding sentences have been correctly identified, it is possible to learn the correspondence translation of each term in the sentence using statistical information since the term's translation will always appear in the corresponding sentences. Therefore, an OOV term can be translated by mining parallel corpora. Many researchers have also reported that parallel text mining based translation can significantly improve the CLIR performance [Eijk 1993; Kupiec 1993; Smadja *et al*. 1996; Nie *et al*. 1999].

In the very early stages, parallel text based translation approaches were word-by-word based and only domain specific noun terms were translated. In general, these approaches [Eijk 1993; Kupiec 1993] first align the sentences in each corpus, then noun phrases are identified by a part-of-speech tagger. Finally, noun terms are mapped using simple frequency calculations. In such translation models, phrases, especially verb phrases, are very hard to translate. As phrases in one language may have different word order in another language, phrases cannot be translated on a word-by-word basis. This problem in parallel text based translation is called the collocation problem.

Some later approaches [Smadja *et al.* 1996; Nie *et al*. 1999] started to use more complex strategies such as statistical association measurement or probabilistic translation models to solve the collocation problem. Smadja *et al.* [Smadja *et al*. 1996] proposed an approach that can translate word pairs and phrases. In particular, they used a statistical association measure of the Dice coefficient to deal with the problem of collocation translation. Nie *et al*. [Nie *et al*. 1999] proposed an approach based on a probabilistic model that demonstrates another approach to solving the collocation problem. Using parallel texts, their translation model can return $p(t|S)$, which is the probability of having the term t of the target language in the translation of the source sentence S. As the probability model does not consider the order and

the position of words, collocation is no longer a problem.

Some of the advantages of the parallel text based approaches include the very high accuracy of translation without using bilingual dictionaries and the extraction of multiple transitions with equivalent meaning that can be used for query expansion. However, the sources of parallel corpora tend to be limited to some particular domain and language pairs. Currently, large-scale parallel corpora are available only in the form of government proceedings, *e.g.* Canadian parliamentary proceedings in English and French, or Hong Kong government proceedings in Chinese and English. Obviously, such corpora are not suitable for translating newly created terms or domain-specific terms that are outside the domains of the corpora. As a result, current studies of parallel text based translation are focusing on constructing large-scale parallel corpora in various domains from the Web.

2.1.3 Web Mining

Web mining for automated translation is based on the observation that there are a large number of Web pages on the Internet that contain parallel text in several languages. Investigation has found that when a new English term, such as a new technical term or a proper name, is introduced into Chinese, the Chinese translation to this term and the original English term very often appear together in literature publications in an attempt to avoid misunderstanding. Some earlier studies have already addressed the problem of extracting useful information from the Internet using Web search engines such as Google and Yahoo. These search engines search for English terms on pages in a certain language, *e.g*., Chinese or Japanese. The results of Web search engines are normally a long, ordered list of document titles and summaries to help users locate information. Mining the result lists can help find translations to the unknown query terms. Some studies [Cheng *et al*. 2004; Zhang *et al*. 2004] have shown that such approaches are rather effective for proper name translation.

Generally, Web-based translation extraction approaches consist of three steps:

1. Web document retrieval: use a Web search engine to find the documents in the target language that contain the OOV term in the original language and collect the text (*i.e*. the summaries) in the result pages returned from the Web search engine.
2. Term extraction: extract the meaningful terms in the summaries where the OOV term appears and record the terms and their frequency in the summaries. As a term in one language could be translated to a phrase or even a sentence, the major difficulty in term extraction is the identification of correct MLUs in the summaries (refer to Section 2.2 for the definition of MLUs).
3. Translation selection: select the appropriate translation from the extracted words. As the previous steps may produce a long list of terms, translation selection has to find the correct translation from the extracted terms.

The existing term extraction techniques in the second step fall into two main categories: approaches that are based on lexical analysis or dictionary-based word segmentation, and approaches that are based on co-occurrence statistics. When translating Chinese text into English, Chinese terms should be correctly detected first. As there are no word boundaries in Chinese text, the mining system has to perform segmentation of the Chinese sentences to find the candidate words. The quality of the segmentation greatly influences the quality of the term extraction because incorrect segmentation of the Chinese text may break the correct translation of an English term into two or more words so that the correct word is lost. The translation selection in the third step also suffers from the problem that selection of the most frequent word or the longest word, which is the more popular techniques, does not always produce a correct translation. The term extraction and translation selection problems will be further addressed in subsequent sections.

2.2 Term Extraction

Term extraction is mainly the task of finding MLUs in the corpus. The concept of MLU is important for applications that exploit language properties, such as Natural Language Processing (NLP), information retrieval and machine translation. An MLU is a group of words that always occur together to convey a specific meaning. For example, compound nouns like *Disneyland*, phrasal verbs like *take into account*, adverbial locutions like *as soon as possible*, and idioms like *cutting edge* are MLUs. In most cases, it is necessary to extract MLUs rather than individual words from a corpus because the meaning of an MLU is not always the combination of individual words in the MLU. The meaning of the MLU 'cutting edge' is not the combination of the meaning of individual words, 'cutting' and 'edge'.

Finding MLUs from the summaries returned by a search engine is important in Web mining based automated translation. If only words are extracted from the summaries, the following process may not be able to find the correct translation since the translation might be a phrase rather than a word. For Chinese text, a word consisting of several characters is not explicitly delimited since Chinese text contains sequences of Chinese characters without spaces between them. Chinese word segmentation is the process of marking word boundaries. The Chinese word segmentation is actually similar to the extraction of MLUs in English documents as the MLU extraction in English documents also needs to mark the lexical boundaries between MLUs. Therefore, term extraction in Chinese documents can be considered as Chinese word segmentation. Many existing systems use lexicon-based or dictionary-based segmentation techniques to determine word boundaries in Chinese text. However, in the case of Web mining for automated translation, as an OOV term is an unknown term to the system, the dictionary-based segmenters usually cannot correctly identify the OOV terms in the sentence. Therefore, the translation of an OOV term cannot be found in

a later process. Some researchers have suggested approaches that are based on co-occurrence statistics model for Chinese word segmentation to avoid this problem [Chen *et al*. 2000; Maeda *et al*. 2000; Gao *et al*. 2001; Pirkola *et al*. 2001].

2.2.1 Mutual Information and its Variations

One of the most popular statistics-based extraction approaches is to use mutual information [Chien 1997; Silva *et al*. 1999]. Mutual information is defined as:

$$MI(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)} = \log_2 \frac{Nf(x,y)}{f(x)f(y)}, \quad (1)$$

The mutual information measurement quantifies the distance between the joint distribution of terms *X* and *Y* and the product of their marginal distributions. When using mutual information in Chinese segmentation, *x, y* are two Chinese characters; *f(x)*, *f(y)*, *f(x,y)* are the frequencies that *x* appears, *y* appears, and *x* and *y* appear together, respectively; *N* is the size of the corpus. A string *XY* will be judged as a term if the *MI* value is greater than a predefined threshold.

Chien [Chien 1997] suggests a variation of the mutual information measurement called significance estimation to extract Chinese keywords from corpora. The significance estimation of a Chinese string is defined as:

$$SE(c) = \frac{f(c)}{f(a) + f(b) - f(c)}, \quad (2)$$

where *c* is a Chinese string with *n* characters; *a* and *b* are the two longest composed substrings of *c* with length *n-1*; *f* is the function to calculate the frequency of a string. Two thresholds are predefined: *THF* and *THSE*. This approach identifies a Chinese string as an MLU by the following steps. For the whole string *c*, if *f(c)>THF*, *c* is considered a Chinese term. For the two *(n-1)*-substrings *a* and *b* of *c*, if *SE(c)>=THSE*, both *a* and *b* are not a Chinese term. If *SE(c)<THSE*, and *f(a)>>f(b)* or *f(b)>>f(a)* , *a* or *b* is a Chinese term, respectively. Then, for each *a* and *b*, the method is recursively applied to determine whether their substrings are terms.

2.2.2 Local Maxima Based Approaches

All mutual information based approaches require tuning the thresholds for generic use. Silva and Lopes suggest an approach called Local Maxima to extract MLU from corpora without using any predefined threshold [Silva *et al*. 1999]. The equation used in Local Maxima is known as SCP and is defined as follows:

$$SCP(s) = \frac{f(s)^2}{\frac{1}{n-1}\sum_{i=1}^{n-1} f(w_1 \ldots w_i) f(w_{i+1} \ldots w_n)}, \quad (3)$$

where S is an n-gram string and $w_1,...,w_i$ is the substring of S. A string is judged as an MLU if the SCP value is greater than or equal to the SCP value of all the substrings of S and also greater than or equal to the SCP value of its antecedent and successor. The antecedent of S is an (n-1)-gram substring of S. The successor of S is a string where S is its antecedent.

Although Local Maxima should be a language-independent approach, Jenq-Haur Wang *et al*. [Cheng *et al*. 2004] found that it does not work well in Chinese word extraction. They introduced context dependency (CD) used together with the Local Maxima. The new approach is called SCPCD. The rank for a string uses the function:

$$SCPCD(s) = \frac{LC(s)RC(s)}{\frac{1}{n-1}\sum_{i=1}^{n-1} freq(w_1...w_i)freq(w_{i+1}...w_n)}, \quad (4)$$

where S is the input string, $w_1..w_i$ is the substring of S, and *LC()* and *RC()* are functions to calculate the number of unique left or right adjacent characters of S. A string is judged as a Chinese term if the SCPCD value is greater than or equal to the SCPCD value of all the substrings of S.

In summary, statistics-based approaches are widely used in Chinese term extraction for translation. The main reason is that Chinese terms are required to be extracted from search engine result pages. However, search engine results are usually partial sentences, which makes the traditional Chinese word segmentation hard to apply in this situation.

Current statistics-based approaches still have weaknesses. Web pages returned from a search engine are used for search engine based OOV term translation. In most cases, only a few hundred of the top result snippets on the result pages are used for translation extraction. Consequently, the corpus size for search engine based approaches is quite small. In a small collection, the frequencies of strings very often are too low to be used in the approaches reviewed. In Section 3.2, we will describe our approach to addressing this difficulty in detail.

3. Web-Based Query Translation

Our approach is based on earlier work by Chen [Chen *et al*. 2000] and Zhang [Zhang *et al*. 2004]. Both approaches submit English queries (usually an English term) to a Web search engine and the top returned results (*i.e*., summaries in Chinese) are segmented into a word list. Each of the words in the list is then assigned a rank calculated from term frequency. The word with the highest rank in the word list is selected as the translation of the English term. However, observations have shown that the correct translation does not always have the highest frequency, even though it very often has a high frequency. The most appropriate translation is not necessarily the term with the highest rank.

As in previous work, we adopted the same idea of finding the OOV term's translation

through a Web search engine. However, our approach differs in term ranking and selection strategy. The aim of our approach is to find the most appropriate translation from the word list regardless of term frequency, which is the basic measurement used in previous work. Our approach combines translation disambiguation technology and Web-based translation extraction technology. Web-based translation extraction usually returns a list of words in a target language. As those words are all extracted from the result snippets returned by the Web search engine, it is reasonable to assume that these words are relevant to the English terms that were submitted to the Web search engine. If we assume all these words are potential translations of the English terms, we can apply the translation disambiguation technique to select the most appropriate word as the translation of the English terms.

Our translation extraction approach contains three major modules: collecting Web document summaries, word extraction, and translation selection. For easier understanding, we will use an example of finding the translation to the term "Stealth Fighter" to demonstrate our approach.

3.1 Collecting Web Document Summaries

First, we collect the top 100 document summaries returned from Google that contain both English and Chinese words. The English queries entered into Google will be enclosed in double quotation marks to ensure Google only returns results with the exact phrase. Sample document summaries are shown in Figure 1.


PChome Online 網路家庭-下載 - [Translate this page]
其中包括了隱形戰機(Stealth fighter)、Su-27、F-16、Sr-71、Glider、X-29等。安裝好後，到控制台中的顯示器內容設定螢幕保護裝置為"3D-Terrain Flight"就可以了。然後，在其設定值中設定所要出現的"戰鬥機種"(總共有六種機型) 及螢幕的視野 ...
toget.pchome.com.tw/intro/desktop_ssaver/desktop_ssaver_military/5317.html - 16k - Cached - Similar pages

SOC GAMING > [分享]Allegiance - [Translate this page]
Stealth Fighter(隱形戰機): 隱形戰機是一種靈活度極差的戰機,因此短距離的纏鬥決不是Stealth Fighter的強項,Stealth Fighter的威力在於其裝載的Hunter Misslie,一種長距離且高殺傷力的飛彈, Stealth Fighter ...
www.socgame.com.tw/bbs/lofiversion/index.php/t56968.html - 15k - Supplemental Result - Cached - Similar pages

蜻蜓的俱樂部- Yahoo!奇摩部落格 - [Translate this page]
1991年一月十七日凌晨，F-117A隱形戰機（stealth fighter），由沙烏地阿拉伯基地 ... 《詳全文》. 回應 (0) 引用 (0). 米格-29支點系列戰機. 分類：特種部隊. 2006/11/29 15:18. MIG-29的進氣道具保護裝置，注意滑行時機翼下方進氣道的保 ... 《詳全文》 ...
tw.myblog.yahoo.com/jw!LXskEjiWHwXt3jX2TYlqG8BQ2w--/archive?l=f&id=32&page=3 - 32k - Supplemental Result - Cached - Similar pages

Figure 1. Three sample document summaries for "Stealth Fighter"

Figure 1 shows that *Stealth Fighter* and its translation in Chinese 隱形戰機 always appear together. The Chinese translation of *Stealth Fighter* appears either before or after the English words. In the example summaries shown in Figure 1, the translation and the English term "Stealth Fighter" are highlighted in red.

Although the query submitted to Google is asking for Chinese documents, Google may still return some documents purely in English. Therefore, we need to filter out the documents that are written in English only. The documents that contain both the English terms and Chinese characters are kept. Also, all the html tags are removed, and only the plain text is kept.

Second, from the document summaries returned by the search engine, we collect the sentences in the target language; for example, we can collect three Chinese sentences from the three sample document summaries in Figure 1. Each sentence must contain the English term and the Chinese characters before and after the term. From the summaries given in Figure 1, we get the following Chinese sentences shown in Figure 2.

Stealth Fighter(隱形戰機): 隱形戰機是一種靈活度極差的戰機,
因此短距離的纏鬥決不是 Stealth Fighter 的強項,
Stealth Fighter 的威力在於其裝載的 Hunter Misslie
一種長距離且高殺傷力的飛彈,

其中包括了隱形戰機(Stealth fighter)、Su-27、F-16、Sr-71、Glider、X-29 等。
安裝好後，到控制台中的顯示器內容設定螢幕保護裝置為"3D-Terrain Flight"就可以了。

1991 年一月十七日凌晨，F-117A 隱形戰機（stealth fighter），由沙烏地阿拉伯基地

Figure 2. Sample output of Chinese string collection

3.2 Word/Phrase Extraction

In order to calculate the statistical information of the Chinese terms, the Chinese sentences have to be correctly segmented. The term extraction approaches reviewed in Section 2.2 have been widely used on large corpora. However, in our experiments, the performance of those approaches is not always satisfactory for search engine based OOV term translation approaches.

In this section, we describe a term extraction approach specifically designed for search engine based translation extraction, which uses term frequency change as an indicator to determine term boundaries and also uses the similarity comparison between individual

character frequencies instead of terms to reduce the impact of low term frequency in small collections. Together with the term extraction approach, we also describe a bottom-up term extraction algorithm that can help to increase the extraction quality.

3.2.1 Frequency Change Measurement

The approaches mentioned in Section 2 use a top-down approach that starts with examining the whole sentence and then examining substrings of the sentence to extract MLUs until the substring becomes empty. We propose using a bottom-up approach that starts with examining the first character and then examining super strings. Our approach is based on the following observations for small document collections:

Observation 1: In a small collection of Chinese text, such as a collection of Web pages returned from a search engine, the frequencies of the characters in an MLU are similar. This is due to the nature of the sample: in a small collection of text, there are a small number of MLUs and the characters appearing in one MLU may not appear in other MLUs. We also found that some different MLUs with similar meanings very often share similar characters and those characters are unlikely to be used in other unrelated MLUs. For example, 戰機 (Fighter Aircraft) and 戰鬥機 have the same meaning in Chinese. They share similar Chinese characters. Therefore, although the term frequency is low in a small collection, the individual characters of the term might still be relatively high and also have similar frequencies. The high frequency can help in term extraction.

Observation 2: When a correct Chinese term is extended with an additional character, the frequency of the extended term very often drops significantly.

According to Observation 1, the frequencies of a term and each character in the term should be similar. We propose to use the root mean square error (RMSE) given in Equation (5) to measure the similarity between the character frequencies.

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \ . \tag{5}$$

For a given Chinese character sequence with n characters, x_i is the frequency of each character in the sequence and $\bar{x}$ is the average frequency of all the characters in the sequence. Although the frequency of a string is low in small corpora, the frequencies of Chinese characters still have relatively high values. According to Observation 1, if a given sequence is an MLU, the characters in the sequence should have a similar frequency, in other words, σ should be small.

If the frequencies of all the characters in a Chinese sequence are equal, then σ = 0. Since σ represents the average frequency deviation from the mean of individual characters in the sequence, according to Observation 1, in an MLU, the longer substring of that MLU will have smaller average frequency error.

According to Observation 1, an MLU can be identified by Equation 5. However, as Equation 5 only measures the frequency similarity between individual characters, any character combinations may be identified as MLUs if their frequencies are similar, even when they are not occurring together. To avoid this problem, we introduce sequence frequency f(S) into the formula. With this addition, if the characters are not occurring together, they will not be considered as a sequence, causing f(S) = 0. Thus, any character combination can be identified if it appears as a sequence in the corpus.

Finally, we combine the sequence frequency and the RMSE measurement. We designed the following equation to measure the possibility of S being a term:

$$R(S) = \frac{f(S)}{\sigma + 1} = \frac{f(S)}{1 + \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \tag{6}$$

where S is a Chinese sequence; f(S) is the frequency of s in the corpus. We use σ +1 as the denominator instead of using σ to avoid 0 denominators.

Let S be a Chinese sequence with n characters; S= a1a2….an. And S' is a substring of S with length n-1; S' = a1a2….an-1. According to Observation 1, if S is an MLU, we should have f(S) ≈ f(S'), and the longer S is, the smaller σ should be. Therefore, in the case where S' is a substring of S with length n-1, we would have $\sigma < \sigma'$. As a result we will have R(S)>R(S'). Consider another case where S' is a substring of S and S' is an MLU while S is not. In other words, S adds an additional character to an MLU. In this case, we will have f(S) <f(S') and the frequency of the additional character makes σ larger, so $\sigma > \sigma'$ and R(S) <R(S').

In summary, for a string S and its substring S', the one with higher R value would most likely be an MLU. Table 1 gives the R value of each possible term in a Chinese sentence chosen from a small collection of summaries returned from a search engine: "隱形戰機/是/一種/靈活度/極差/的/戰機" ("/" indicates the lexicon boundary given by a human).

Table 1. Chinese strings and R(S)

String *S*	*R(S)*
隱形	26.00
隱形戰	0.94
戰機	2.89
戰機是	0.08
一種	0.44
一種靈	0.21
靈活	2.00
靈活度	2.00
靈活度極	1.07
極差	0.8
極差的	0.07
戰機	2.89

This example clearly shows that, if a Chinese MLU has an additional character, its R value will be significantly smaller than the R value of the MLU. For example, 一種, 靈活, and 靈活度 are valid MLUs, but 一種靈 and 靈活度極 are not.

In Table 1, we have:

R(一種)=0.44 > R(一種靈)=0.21, R(靈活)=R(靈活度)=2.00>R(靈活度極)=1.07,

which shows the R value drop from 一種 to 一種靈, and from 靈活 and 靈活度 to 靈活度極.

This example indicates that it is reasonable to segment the Chinese sentence at the positions where the string's R value drops greatly. For the example sentence, it would be segmented as: "隱形/戰機/是/一種/靈活度/極差/的/戰機" by the proposed method. The only difference between the human segmented sentence and the automatic segmented sentence is that "隱形戰機" (Stealth Fighter) is segmented into two words "隱形" (Stealth) and "戰機" (Fighter) by the proposed method. However, this is still an acceptable segmentation because those two words are meaningful words in Chinese and have the same meaning as the combination of the two words.

3.2.2 A Bottom-Up Term Extraction Strategy

As mentioned in Section 3.1, the top-down strategy is to first check whether the whole sentence is an MLU, then reduce the sentence size by 1 and recursively check sub-sequences.

It is reported that over 90% of meaningful Chinese terms consist of less than 4 characters [Wu 2004], and, on average, the number of characters in a sentence is much larger than 4. Obviously, a whole sentence is unlikely to be an MLU. Therefore, checking the whole sentence for an MLU is unnecessary. In this section, we describe a bottom-up strategy that extracts terms starting from the first character in the sentence. The basic idea is to determine the boundary of a term in a sentence by examining the frequency change (*i.e.*, the change of the R value defined in Equation (6)) when the size of the term is increasing. If the R value of a term with size n+1 drops compared with its largest sub term with size n, the sub term with size n is extracted as an MLU. For example, in Table 1, there is a big drop between the R value of the third term "靈活度" (2.00) and its super term "靈活度極" (1.07). Therefore, "靈活度" is considered as an MLU.

The following algorithm describes the bottom-up term extraction strategy:

Algorithm BUTE(s)

Input: s=a1a2….an is a Chinese sentence with n Chinese characters

Output: M, a set of MLUs

Check each character in s, if it is a stop character such as 是(is, are)，的(of) ，了…, remove it from s. After removing all stop characters, s becomes a1a2….am, m≤n.

Let b=1, e=1, and M=φ

Let t1= aba2….ae, t2= aba2….a(e+1).

If R(t1) >R(t2), then M=M∪(t1), b=e+1.

e=e+1, if e+1>m, return M, otherwise go to step 3.

Once a sequence is identified as an MLU, the algorithm BUTE will not check its subsequences for other possible MLUs (*i.e.*, b=e+1 in step 3 makes it so the next valid checkable sequence doesn't contain t1, which was just extracted as an MLU). However, when using the bottom-up strategy, some longer terms might be missed when the longer term contains several shorter terms. As shown in our example, "隱形戰機" (Stealth Fighter) consists of two terms "隱形" and "戰機". When using bottom-up strategy, "隱形戰機" would not be extracted because the composite term has been segmented into two terms. To avoid this problem, we set up a fixed number ω which specifies the maximum number of characters to be examined before reducing the size of the checkable sequence. The modified algorithm is given below:

Algorithm BUTE-M(s)

Input: s=a1a2….an is a Chinese sentence with n Chinese characters

Output: M, a set of MLUs

Check each character in s, if it is a stop character such as 是，了，的…, remove it from s. After removing all stop characters, s becomes a1a2….am, m≤n.

Let b=1, e=1, First-term = true, and M= φ

Let t1= aba2….ae, t2= aba2….a(e+1).

If R(t1) >R(t2),

then M:=M ∪ {t1)

If First-term = true

then first-position:= e and First-term:= false

If e-b+1 ≥ ω

then e:=first-position, b:=e+1, First-term:=true.

e=e+1, if e+1>m, return M, otherwise go to step 3

In algorithm BUTE-M, the variable first-position gives the ending position of the first identified MLU. Only when ω characters have been examined will the first identified MLU be removed from the next valid checkable sequence, otherwise the current sequence is still being checked for a possible MLU even if it contains an extracted MLU. Therefore, not only will the terms "隱形" and "戰機" be extracted, but also the longer term "隱形戰機" (Stealth Fighter) will be extracted.

3.3 Translation Selection

At this point, we have a list of translation candidates for the query term. The final step is to find the correct translation from the candidate list.

As we have described in another paper [Lu *et al*. 2007], the traditional translation selection approaches select the translation based on word frequency and word length [Chen *et al*. 2003; Zhang *et al*. 2004]. We have proposed an approach to determining the most appropriate translation from the extracted word list using the documents in the collection dataset regardless of term frequency. Using this approach, even a low-frequency word might be selected. Our experiments in that paper show that in some cases, the most appropriate translation can be a word with low frequency.

First, we retrieve the documents that contain each candidate translation from the collection. Then, we calculate the frequency of each candidate translation in the collection.

For instance, suppose we have an English query with three terms A,B,C and A1,A2…, B1,B2…, and C1,C2… are the candidate translations for A, B, and C, respectively, and the frequency of A1, A2, …, B1, B2, …, C1,C2… in the collection is f(A1), f(A2),… f(B1), f(B2)…, and so on. Second, we retrieve the documents that contain all the possible combinations of the candidate translations and calculate the frequencies. For example, the frequency of combination A1B1C1 is f(A1B1C1), A1B2C1 is f(A1B2C1), and A1B2C3 is f(A1B2C3)… and so on. Finally, we calculate the co-occurrence of all the possible combinations using the following equation:

$$C(x_1x_2x_3...x_n) = \log_2 \frac{N^{n-1} f(x_1x_2x_3...x_n)}{f(x_1)f(x_2)f(x_3)...f(x_n)}, \quad (7)$$

where x_i is a candidate translation for the ith query term, $f(x_i)$ is the frequency of word x_i appearing in the corpus, $x_1x_2…x_n$ is a combination of the candidate translation, $f(x_1x_2x_3...x_n)$ is the frequency that $x_1x_2…x_n$ appears in the corpus. N is the size of the corpus.

For the example query with three terms A, B, C, the co-occurrence of three candidate translation $A_1B_1C_1$ is calculated by:

$$C(A1B1C1) = \log_2 \frac{N^2 f(A1B1C1)}{f(A1)f(B1)f(C1)}. \quad (8)$$

The translation combination with the highest total correlation value C is selected as the correct translation for that query.

4. Experiments

We have conducted experiments to evaluate our proposed query translation approach. The Web search engine used in the experiments was Google.

4.1 Test Set

Queries, document collection, and relevance judgments provided by NTCIR (http://research.nii.ac.jp/ntcir/) are used in the experiments. The NTCIR6 Chinese test document collection was used as our test collection. The articles in the collection are news articles published on United Daily News (udn), United Express (ude), MingHseng News (mhn), and Economic Daily News (edn) in 2000-2001, for a total of 901,446 articles.

Queries used in the experiments are from NTCIR5 and NTCIR6 CLIR tasks. Altogether, there are 100 queries created by researchers from Taiwan, Japan, and Korea. NTCIR provided both English queries and corresponding Chinese queries. The Chinese queries are translated by human translators and are, thus, correct translations of the corresponding English queries.

In our experiments, English queries are extracted from English description fields by human experts. The corresponding Chinese translations are transcribed from the Chinese title fields by humans.

Yahoo's online English-Chinese dictionary (http://tw.dictionary.yahoo.com/) is used in the experiments. We first translate the English queries using the Yahoo's online English-Chinese dictionary. The terms that could not be translated by the online dictionary were used as the input queries to evaluate the performance of our proposed Web-based query translation approach. There are 108 OOV terms that cannot be translated by the online dictionary and, therefore, are used in the experiments.

4.2 Retrieval System

The documents were indexed using a character-based inverted file index. In the inverted file, the indexer records each Chinese character, its position in the document, and the document ID. Chinese phrase is determined by each Chinese character position and document ID. Only when character positions are consecutive and have the same document ID will the character sequence be considered as a phrase in the document. English words and numbers in the document are also recorded in the inverted file.

The retrieval model that is used in the system is an extended Boolean model with tf-idf weighting schema which is used in GPX by [Geva 2006]. Document rank for a query Q is calculated by the equation below:

$$D_{rank} = n^5 \sum tf_i * idf_i$$

Here, *n* is the number of the unique query terms in the document. *tfi* is the frequency of the ith term in the document and *idfi* is the inverse document frequency of the ith term in the collection. This equation can ensure two things: first, the more unique query terms that match in a document, the higher rank the document has. For example, the document that contains five unique query terms will always have higher rank than the document that contains four query terms, regardless of the query terms frequency in the document; second, when documents contain the same number of unique terms, the score of a document will be determined by the sum of query terms' *tf-idf*, as traditional information retrieval does.

We do not employ relevance feedback in the retrieval system. Also, all the retrieval results are initial search results without query expansion.

4.3 Experiment Design

We designed two sets of experiments to evaluate our approach. The first set of experiments was designed to evaluate the effectiveness of term extraction for OOV translation, and the second set of experiments was designed to evaluate the effectiveness of translation selection

for OOV translation.

4.3.1 Experiment Set 1

In this experiment, we compared the performance of our proposed translation extraction approach (denoted as SQUT) with the approaches reviewed in Section 2.2, including the Mutual Information method (denoted as MI), the approach introduced by [Chien 1997] (denoted as SE), the Local Maxima method introduced by [Silva *et al*. 1999] (denoted as SCP), and the approach introduced by [Cheng *et al*. 2004] (denoted as SCPCD).

The OOV term is translated via the following steps:

Send the OOV term as a query to Google; from the result pages returned from Google, use the five different term extraction approaches to produce five Chinese term lists.

If a Chinese word can be translated to an English word using a dictionary, the English word must not be an OOV word. This means, the Chinese word must not be a translation of the queried English OOV word. Therefore, for each term list obtained in Step 1, remove the terms which can be translated to English by Yahoo's online dictionary. After this step, only OOV terms remain.

Select the top 20 terms in the new term list as translation candidates. Select the final translation from the candidate list using our translation selection approach described in 3.3.

Finally, we have five sets of OOV translations produced by the five approaches, respectively. A sample of the translation is given in Appendix 1.

Translation accuracy will be determined by human experts. Chinese queries will be used as reference only. As we were using the same corpus and the same translation selection approach, the difference in translation accuracy is the result of using different term extraction approaches. Thus, we can claim that the approach with the higher translation accuracy has higher extraction accuracy.

4.3.2 Experiment Set 2

This experiment is to retrieve Chinese documents for a given English query. The following experiments were conducted:

1. Mono: in this run, we use the original Chinese queries form NTCIR5. Only the title field is used and the Chinese terms are segmented by a human. This run provides the baseline result for comparison with all other runs.
2. IgnoreOOV: in this run, the English queries are translated using the online Yahoo

English-Chinese dictionary with the disambiguation technology proposed in 3.3. If a translation is not found in the dictionary, the query will keep the original English word.

3. SimpleSelect: similar to IgnoreOOV, English queries are translated using the online Yahoo English-Chinese dictionary with disambiguation technology. If a term cannot be translated by the dictionary, it will be translated by the proposed Web mining based approach. However, in the translation selection step, the longest and the highest frequency string were selected as its translation. This run simulates the previous Web translation selection approaches.
4. TQUT: like SimpleSelect, except that the translation for the "missing word" is selected with the disambiguation technology that is discussed in 3.3. Actually, TQUT uses the same translation technology as SQUT which we used in Experiment Set 1. We named it TQUT here simply to distinguish the concept that TQUT is an information retrieval task while SQUT is a translation task.

Although NTCIR gives 190 queries, only 100 of them have relevance judgments. Therefore, we are only able to evaluate the retrieval performance using those 100 queries in Experiment Set 2.

5. Results and Discussion

5.1 Experiment Set 1

For the 108 OOV terms, using the five different term extraction approaches, we obtained the translation results shown in Table 2. As we were using the same corpus and the same translation selection approach, the difference in translation accuracy is the result of different term extraction approaches. Thus, we can claim that the approach with the higher translation accuracy has higher extraction accuracy.

As we can see from Table 2, below, SQUT has the highest translation accuracy. SCP and SCPCD provided similar performance. The approaches based on Mutual Information provided lowest performance.

Table 2. OOV translation accuracy

	Correct	Accuracy (%)
MI	48	44.4
SE	58	53.7
SCP	73	67.6
SCPCD	74	68.5
SQUT	84	77.8

5.1.1 Mutual Information Based Approaches

In the experiment, the MI based approaches were unable to determine the Chinese term boundaries well. The term lists produced by the MI based approaches contain a huge number of partial Chinese terms. It is quite often the case that partial Chinese terms were chosen as the translation of OOV terms. Some partial Chinese terms selected by our system are listed in Table 3.

Table 3. Some Extracted terms by MI

OOV Terms	Extracted terms	Correct terms
Embryonic Stem Cell	胚胎幹細	胚胎幹細胞
consumption tax	費稅	消費稅
Promoting Academic Excellence	卓越發	卓越發展計畫

The performance of the Mutual Information based term extraction approaches, such as MI and SE, is affected by many factors. These approaches rely on predefined thresholds to determine the lexicon boundaries. Those thresholds can only be adjusted experimentally. Therefore, they can be optimized in fixed corpora. However, in OOV term translation, the corpus is dynamic. It is almost impossible to optimize thresholds for general use. As a result, the output quality is not guaranteed.

In addition, Mutual Information based approaches seem unsuitable in Chinese term extraction. As there are no word boundaries between Chinese words, the calculation of MI values in Chinese is based on Chinese characters but not words as in English. On average, a high school graduate in the U.S. has a vocabulary of 27,600 words [Salovesh 1996]. Unless stemming or lemmatizing is used, the number of English word variations in a corpus is much greater. In contrast, the cardinality of the commonly used Chinese character set is under 3000. Due to the small set of Chinese characters, Chinese characters have much higher frequencies than English words. This means that one Chinese character could be used in many MLUs while an English word will have a much lower probability of being used in Multiple MLUs. As a result, an English MLU will have much higher MI value than a Chinese MLU. The subtle difference in MI values between MLUs and non-MLUs in Chinese makes the thresholds hard to tune for general use.

Some filtering techniques are used in SE to minimize the affect of thresholds. In our experiment, there is a 17.2% improvement in translation accuracy. Obviously, the improvement comes from the higher quality of extracted terms. However, the limitation of thresholds is not avoidable.

5.1.2 Local Maxima Based Approaches

Without using thresholds, Local Maxima based approaches have much better flexibility than the MI based approaches in various corpora, achieving higher translation accuracy in our experiment. Comparing the two, the SCP approach tries to extract longer MLUs while the SCPCD approach tries to extract shorter ones. The translation of "Autumn Struggle", "Wang Dan", "Masako" and "Renault" are all 2-character Chinese terms. SCPCD can extract the translation with no problem while SCP always has difficulty with them. As over 90% of the Chinese terms are short terms, this is a problem for SCP in Chinese term extraction. Conversely, SCPCD has difficulty in extracting long terms. Overall, the two Local Maxima based approaches have similar performance. However, since most of the translation of OOV terms are long terms in our experiment, SCP's performance is a little better than that of SCPCD.

Local Maxima based approaches use string frequencies in the calculation of $\frac{1}{n-1}\sum_{i=1}^{n-1} f(w_1...w_i)f(w_{i+1}...w_n)$. In a small corpus, the frequency of a string becomes very low, which makes the calculation of string frequencies less meaningful. Local Maxima based approaches are not effective in a small corpus. In comparison, our approach calculates the difference between character frequencies. In a small corpus, characters still have a relatively high value. As a result, our approach performs better than Local Maxima based approaches in small corpora. For example, local maxima based approaches were unable to extract the translation of "Nissan Motor Company" because the corpus is too small-Google only returns 73 results for the query "Nissan Motor Company".

5.1.3 SQUT Approach

Most of the translations can be extracted by the SQUT algorithm. As our approach monitors the change in R value to determine MLUs rather than using the absolute value of R, it does not have the difficulty of using predefined thresholds. In addition, the use of single character frequencies in RMSE calculation makes our approach suitable in small corpora. Therefore, we have much higher translation accuracy than the MI-based approaches and also about 10% improvement over the Local Maxima based approaches.

However, the SQUT algorithm has difficulty in extracting the translation of "Wang Dan". In analyzing the result summaries, we found that the Chinese character "王" ("Wang") is a very high-frequency character in the summaries. It is also used in other terms such as "霸王" (the Conqueror), "帝王" (regal); "國王" (king); "女王" (queen) and "王朝" (dynasty). Those terms also appear frequently in the result summaries. In our approach, where we are using the count of individual characters, the very high frequency of "王" breaks Observation 2. Thus, the translation of "Wang Dan" cannot be extracted. However, in most cases, our observations are true in small corpora as demonstrated by the high translation accuracy of our approach in

query expansion from Chinese/English Web search summaries.

5.2 Experiment Set 2

Table 4 below gives the results from the four runs defined in Section 4.3.2.

Table 4. NTCIR 5 retrieval performance

	Average precision	Percentage of MonoRun
Mono	0. 3713	-
IgnoreOOV	0.1312	35.3%
SimpleSelect	0.2482	66.8%
TQUT	0.2978	79.3%

5.2.1 IgnoreOOV

The performance of the IgnoreOOV is 0.1312 which is only 35.3% of the monolingual retrieval performance. This result shows the extent to which an OOV term can affect a query. By looking at the translated queries, we found that 62 queries out of 100 have OOV terms. By removing all 62 queries, the Mono's average precision becomes 0.3026 and the IgnoreOOV's average precision becomes 0.2581 which is about 85.3% of the Mono's precision. This is a reasonable result and indicates that our disambiguation technique works well to find the correct translations. The reason that we cannot get 100% precision is mainly due to the limited coverage of the dictionary introducing inappropriate translations. By "inappropriate translation", we mean that the translation is a valid translation in some other context but not in the current query context. In query 24: for "space station, Mir", 儲存信息暫存器 (Memory Information Register) is the only translation returned from the dictionary. However, it should be translated to 和平號太空站 here. In this case, when a dictionary only returns one translation, it is difficult to tell if it is suitable in the context. As the dictionary only gives one translation, we have no opportunity to correct this translation error using a disambiguation technique. Some translations from the dictionary are inappropriate in some given contexts because the translations are different in different regions. For example, the query "mad cow disease" is translated to 瘋牛病 in the dictionary which is used in mainland China and Hong Kong. However, in the NTCIR collection which is obtained from Taiwan, "mad cow disease" is translated to 狂牛症 or to 狂牛病. We also find the same problem in query 24 "syndrome". Its translation is 症候群 in Taiwan. The translations given in the dictionary, though, are 併發症狀 and 綜合症狀, which are used in Hong Kong and mainland China. With these inappropriate translations, the retrieval precision for these queries is very low, thus it is impossible to achieve 100% of Mono performance.

5.2.2 SimpleSelect

The performance of SimpleSelect, which achieved 0.2482 in precision, was much better than IgnoreOOV and it is 66.8% of the Mono performance. This result shows quite clearly that some of the OOV terms in English are found and translated to Chinese correctly.

Table 5. Retrieval performance on queries that contains OOV terms only

	Average precision	Percentage of Mono Run
Mono	0.4134	-
SimpleSelect	0.2149	52.0%
TQUT	0.2946	71.3%

The results of the 62 queries that have OOV terms are given in Table 5. From Table 5, we can see that the precision of Mono is 0.4134 and the precision of SimpleSelect is 0.2149 which is 52.0% of the Mono's precision. This indicates that just choosing the longest and highest frequency terms as the translation of OOV terms results in performance that is actually lower than looking them up the dictionary. The performance is quite close to the performance of looking up terms in a dictionary without translation disambiguation technology reported by other researchers. However, some of our results show that this approach is quite useful in looking up proper names. As there is no standard for name translation in Chinese, it is quite common that a person's name might be translated into different forms with similar pronunciation (akin to phonetic form). Different people may choose different translations due to their custom. As our test collection contains articles from four different news agents, if we only choose one of the translations, we may not retrieve all the relevant documents.

For example, in query 12, the precision of SimpleSelect is 0.3528 and the precision of Mono is 0.0508 which means SimpleSelect's performance is vastly superior to Mono. This is a notable performance boost. The English OOV term in query 12 is Jennifer Capriati (the name of a tennis player). The human translation is 卡普莉雅蒂. The translations from our approach are 卡普裏亞蒂, 卡普莉雅蒂, 卡普裏雅蒂 and 雅蒂. They are all correct translations. It is clear that we miss many relevant documents when we only use the translation 卡普莉雅蒂. When we take a deep look into the collection, actually three out of four news agents have sports news. Those three news agents use three different translations for Jennifer Capriati. These translations are 卡普莉雅蒂 in the mhn, 凱普莉雅蒂 in the ude and 卡普莉亞蒂 in the udn. Obviously, our translated query takes advantage of adding 雅蒂. Since we use a character-based index for our collection, the documents containing 雅蒂 will include the documents that contain both 卡普莉雅蒂 and 凱普莉雅蒂. Therefore, although we cannot find the correct translation 凱普莉雅蒂, we can still retrieve the documents that contain 凱普莉雅蒂 by using 雅蒂.

Although using part of the translation might improve the retrieval performance, it also

introduces noise information and the noise information may make it harder for the search engine to find the relevant documents. For example, America Online is translated as 美國線上 in Taiwan but 美國在線 in mainland China. If we only choose 美國 (American) as the translation, we lose the information of the term. If it is the only term in the query, obviously, we are not going to retrieve any relevant documents.

5.2.3 TQUT

Table 6. OOV translation accuracy NTCIR5&6

	Correct	No. of OOV	Accuracy (%)
TQUT	50	71	70
SimpleSelect	43	71	60

Table 6 shows that, using translation disambiguated technology in Web Translation Extraction, we can get more accurate translation than in previous approaches. We have 65% accuracy of the translation while the simulation of previous approach only achieves 51%. The IR performance of disambiguated queries achieved 79.3% of the Mono which is 0.2978. If we only look at the results of 62 queries that contain OOV terms, the precision is 0.2846 which is 71.3% of the Mono's precision. This result is much higher than the result in SimpleSelect, which is only 52% of Mono. There are 71 OOV terms over 100 queries. 50 of the OOV terms' translations can be found using our proposed approach. And 43 of the translations are equivalent to the human translation. It is about 70% in precision.

There are many reasons for not being able to get 100% precision. The first reason is the different translation customs that we described earlier. Since we cannot control from where the Web search engine gets the documents and to whom the Web search engine returns documents, we cannot guarantee the translation will be suitable for the collection. For example, we may be able to find the translation for an OOV term from the Internet, but this translation may be used only in Hong Kong and is not suitable for a collection from Taiwan. The translation of the term "Kursk" is a good example. Our Web translation extraction method only returns one translation 庫爾斯克 as the translation of "Kursk". This result shows that most of the documents over the Internet use 庫爾斯克 as the translation of "Kursk". However, the NTCIR5 collection uses 科斯克 as its translation. This kind of inappropriate translation is very hard to avoid even by human interpreters. Another good example is the translation of "National Council of Timorese Resistance". We believe 帝汶抵抗全國委員會 (from our Web translation extraction system) and 東帝汶人抗爭國家委員會 (from NTCIR human translation) are both correct. The difference of the two translations comes from the different customs of translation. However, when using the two translations as two queries, our IR system cannot retrieve any documents. This means that the documents in the NTCIR5

collection use a different translation for "National Council of Timorese Resistance". Actually the translation in the NTCIR5 collection is: 東帝汶全國反抗會議.

Another reason that we cannot get 100% precision is that our Web translation extraction system does not consider the query context. As we described before, we only put the OOV terms into a Web search engine. This may lead to a situation where we get a translation suitable for other context. For instance, in query 36, we are looking for some articles about the use of a robot for remote operation in a medical context. "Remote operation" is an OOV term in this query. Our Web translation extraction method returns the term 遠程操作服務 as its translation. Disregarding the query context, this is a correct translation. However, this translation is only correct when it is used in computer science. If we do not consider the query context, 27 of the translations are correct with about 87% precision. This result is close to the disambiguated queries of dictionary translations which is 85%.

6. Conclusion and Future Work

In this paper, first, we reviewed some existing popular OOV translation approaches. Then, we described an approach to tackling the OOV problem in English-Chinese information retrieval. As the first step of this approach, we proposed a bottom-up term extraction approach suitable for small corpora for generating candidate translations for query OOV terms. This method introduces a new measurement of a Chinese string based on frequency and RMSE, together with a Chinese MLU extraction process based on the change to a new string measurement that does not rely on any predefined thresholds. The method considers a Chinese string as a term based on the change of R's value when the size of the string increases rather than based on the absolute value of R. Our experiments show that this approach is effective for translation extraction of unknown query terms.

We also proposed a simple translation selection approach to improve translation accuracy. Our experimental results show that OOV terms can significantly affect the performance of CLIR systems. Using the translation extraction method proposed in this paper, the overall performance can be boosted by almost 174% relative to the case of not processing OOV terms. With our proposed translation selection approach, the accuracy of OOV term translation can be improved by up to 85%. The overall performance shows about 200% improvement relative to the case of not processing OOV terms. Also, it is about 120% relative to our implementation of previous approaches.

Although our proposed approach shows impressive accuracy for OOV term translation, there is still some work to be done in the future. First, our experiments were conducted using a relatively small scale test set from NTCIR5 and NTCIR6 along with CLIR task queries which only have 108 OOV terms. It is necessary to test our approach to a larger-scale test set such as a test set that has over 1000 OOV terms. Second, inappropriate translation is still a problem in

query translation. The main reasons include the limited size of the dictionary, different customs of translation, and ignoring query context. Some work should be done to minimize these problems. Our experiments provide hints for some possible approaches. If we have a large amount of resources, we may be able to find all the possible translations. For translation selection, if some of the translations hit a similar number of documents, we may keep all of them as correct translations. It may be useful to include more results from the Google search for instance or combining different translation result together. We will validate these ideas in the future.

REFERENCES

"The List of common use Chinese Characters", Ministry of Education of the People's Republic of China.

Chen, A. and F. Gey, "Experiments on Cross-language and Patent retrieval at NTCIR3 Workshop," in *Proceedings of the 3rd NTCIR Workshop*, Japan, 2003.

Chen, A., H. Jiang, and F. Gey, "Combining multiple sources for short query translation in Chinese-English cross-language information retrieval," in *Proceedings of the fifth international workshop on on Information retrieval with Asian languages*, Hong Kong, China, ACM Press, 2000.

Cheng, P.-J., J.-W. Teng, R.-C. Chen, J-H. Wang, W.-H. Lu, and L.-F. Chien, "Translating unknown queries with web corpora for cross-language information retrieval," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, Sheffield, United Kingdom, ACM Press, 2004.

Chien, L.-F., "PAT-tree-based keyword extraction for Chinese information retrieval," in *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval,* Philadelphia, Pennsylvania, United States, ACM Press, 1997.

Eijk, P. v. d., "Automating the acquisition of bilingual terminology," in Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics, Utrecht, The Netherlands, 1993.

Gao, J., J.-Y. Nie, E. Xun, J. Zhang, M. Zhou, and C. Huang, "Improving query translation for cross-language information retrieval using statistical models," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, New Orleans, Louisiana, United States, ACM Press, 2001.

Geva, S.,"Gardens Point XML IR at INEX 2006," Comparative Evaluation of XML information Retrieval Systems 5th International Workshop of the Initiative for the Evaluation of XML Retrieval, Dagstuhl Castle, Germany, Springer, 2006.

Jang, M.-G., S. H. Myaeng, and S.Y. Park, "Using mutual information to resolve query translation ambiguities and query term weighting," *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*,

College Park, Maryland, Association for Computational Linguistics, 1999.

Kupiec, J. M., "An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora," in *Proceedings of the 31st Annual Meeting of the ACL*, Columbus, Ohio, 1993.

Lu, C., Y. Xu, and S. Geva, "Translation disambiguation in web-based translation extraction for English-Chinese CLIR," in *Proceeding of The 22nd Annual ACM Symposium on Applied Computing*, 2007, pp. 819-823.

Maeda, A., F. Sadat, M. Yoshikawa, and S. Uemura, "Query term disambiguation for Web cross-language information retrieval using a search engine," in *Proceedings of the fifth international workshop on on Information retrieval with Asian languages*, Hong Kong, China, ACM Press, 2000, pp. 25-32.

Nie, J.-Y., M. Simard, P. Isabelle, and R. Durand, "Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, Berkeley, California, United States, ACM Press, 1999, pp. 74-81.

Paola, V. and K. Sanjeev, "Transliteration of proper names in cross-lingual information retrieval," *Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition - Volume 15*, Association for Computational Linguistics, 2003.

Pirkola, A., T. Hedlund, H. Keskustalo, and K. Järvelin,"Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings," *Information Retrieval*, 4(3-4), 2001, pp. 209 - 230.

Salovesh, M., "How many words in an "average" person's vocabulary?" *http://unauthorised.org/anthropology/anthro-l/august-1996/0436.html*, DOI:, 1996.

Silva, J. F. d., G. Dias, S. Guilloré, and J.G. Pereira, "Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units," *Progress in Artificial Intelligence: 9th Portuguese Conference on Artificial Intelligence*, 1999.

Silva, J. F. d. and G. P. Lopes., "A Local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units," *International Conference on Mathematics of Language*, 1999.

Smadja, F., K. R. McKeown, and V. Hatzivassiloglou, "Translating collocations for bilingual lexicons: a statistical approach," *Computational Linguistics,* 22(1), 1996, pp. 1-38.

Wu, G., "Research and Application on Statistical Language Model," *Computer science and technology*, Beijing, Tsinghua University, China, 2004.

Yan, Q., G. Gregory, and D.A. Evans., "Automatic transliteration for Japanese-to-English text retrieval," *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, Toronto, Canada, ACM Press, 2003.

Zhang, Y. and P. Vines, "Using the web for automated translation extraction in cross-language information retrieval," *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, Sheffield, United Kingdom, ACM Press, 2004.

Appendix 1

Sample translations of OOV terms from NTCIR

OOV term	SQUT	SCP	SCPCD	SE	MI
Chiutou:					
Autumn Struggle:	秋鬥大遊	從秋鬥	秋鬥	秋鬥	秋鬥
Jonnie Walker:	約翰走路	約翰走路	黑次元	高雄演唱	高雄演唱
Charity Golf Tournament:	慈善高爾夫球賽	慈善高爾夫球賽		慈善高	慈善高
Embryonic Stem Cell:	胚胎幹細胞	胚胎幹細胞	胚胎幹細胞		
Florence Griffith Joyner:	花蝴蝶	葛瑞菲絲	葛瑞菲絲	花蝴蝶	花蝴蝶
FloJo:	佛羅倫薩格里菲斯	花蝴蝶	花蝴蝶	花蝴蝶	花蝴蝶
Michael Jordan:	麥可喬丹	麥可喬丹	喬丹	喬丹	喬丹
Torrijos Carter Treaty:					
Viagra:					
Hu Jin tao:	胡錦濤	胡錦濤	胡錦濤	胡錦濤	胡錦濤
Wang Dan:		天安門	王丹	王丹	王丹
Tiananmen	天安門廣場	天安門	天安門	天安門	天安門
Akira Kurosawa:	黑澤明	黑澤明	黑澤明	黑澤明	黑澤明
Keizo Obuchi:	小淵惠三	小淵惠三	小淵惠三	小淵惠三	小淵惠三
Environmental Hormone:	環境荷爾蒙	環境荷爾蒙	環境荷爾蒙	環境荷爾蒙	
Acquired Immune Deficiency Syndrome:	後天免疫缺乏症候群	愛滋病	愛滋病	愛滋病	愛滋
Social Problem:	社會問題	社會問題	社會問題		
Kia Motors:	起亞汽車	起亞汽車	起亞汽車	起亞	起亞
Self Defense Force:	自衛隊	自衛隊	自衛隊	自衛隊	自衛隊
Animal Cloning Technique:	動物克隆技術	動物克隆技術			
Political Crisis:	政治危機	政治危機	政治危機		
Public Officer:	公職人員	公職人員	公職人員	公職人員	

Research Trend:	研究趨勢	研究趨勢	研究趨勢	研究趨勢	
Foreign Worker:	外籍勞工	外籍勞工	外籍勞工	外籍勞工	
World Cup:	世界盃	世界盃	世界盃	世界盃	世界盃
Apple Computer:	蘋果公司	蘋果電腦	蘋果電腦	蘋果電腦	蘋果電腦
Weapon of Mass Destruction:	大規模毀滅性武器	大規模毀滅性武器	性武器		
Energy Consumption:	能源消費	能源消費	能源消費		
International Space Station:	國際太空站	國際太空站	國際太空站		
President Habibie:	哈比比總統	哈比比總統	哈比比總統	哈比比	
Underground Nuclear Test:	地下核試驗	地下核試驗	地下核試		
F117:	戰鬥機	隱形戰鬥機	隱形戰	隱形戰	隱形戰
Stealth Fighter:	隱形戰機	隱形戰機	形戰鬥機	形戰鬥機	形戰鬥機
Masako:	雅子	太子妃	雅子	雅子	雅子
Copyright Protection:	版權保護	版權保護	版權保護	版權保護	版權保護
Daepodong:	大浦洞	大浦洞	大浦洞	大浦洞	大浦洞
Contactless SMART Card:	智慧卡	非接觸式智慧卡	非接觸式智慧卡	非接觸式	非接觸式
Han Dynasty:	漢朝	大漢風	漢朝	漢朝	漢朝
Promoting Academic Excellence:	學術追求卓越發展計畫	卓越計畫	卓越發展計畫	卓越發展計畫	卓越發
China Airlines:	中華航空	中華航空	中華航空	中華航空	長榮
ST1:					
El Nino	聖嬰	聖嬰現象	聖嬰現象	聖嬰	聖嬰
Mount Ali:	阿里山	阿里山	阿里山	阿里山	阿里山
Kazuhiro Sasaki:	佐佐木主浩	佐佐木主浩	佐佐木	佐佐木	佐佐木
Seattle Mariners:	西雅圖水手	西雅圖水手	西雅圖水手		
Takeshi Kitano:	北野武	北野武	北野武	北野武	北野武
European monetary union:	歐洲貨幣聯盟	歐洲貨幣聯盟	歐洲貨幣	歐洲貨幣	歐洲貨幣
capital tie up:					
Nissan Motor Company:	日產汽車公司	汽車公司	汽車公司	處經濟	處經濟
Renault:	雷諾	休旅車	雷諾	雷諾	雷諾

Pol Pot:	波布	紅高棉	紅高棉	紅高棉	紅高棉
war crime:	戰爭罪	戰爭罪	戰爭罪	戰爭罪	
Kim Dae Jung:	金大中	金大中	金大中	金大中	金大中
Clinton:	克林頓	克林頓	克林頓		
New Year Holiday:	新年假期	新年假期	新年假期		
Drunken Driving:	醉後駕車	醉後駕車	醉後駕車	醉後駕車	後駕車
Science Camp:	科學營	科學營	科學營	科學營	
Nelson Mandela:	曼德拉	曼德拉	曼德拉	曼德拉	曼德拉
Kim Il Sung:	金日成	金日成	金日成	金日成	金日成
anticancer drug:	抗癌藥物				
consumption tax:	消費稅	消費稅	消費稅	消費稅	費稅
Uruguay Round:	烏拉圭回合	烏拉圭回合	烏拉圭回合		
Kim Jong Il:	金正日	金正日	金正日	金正日	金正日
Time Warner	時代華納	時代華納	時代華納	時代華納	時代華納
American Online	美國線上	美國線上	美國線上	美國線上	美國線上
Alberto Fujimori	藤森	藤森	藤森	藤森	藤森
Taliban	塔利班	塔利班	塔利班	塔利班	塔利班
Tiger Woods	老虎伍茲	老虎伍茲	老虎伍茲	老虎伍茲	伍茲
Harry Potter	哈利波特	哈利波特	哈利波特	哈利波特	哈利波特
Greenspan	葛林斯班	葛林斯班	葛林斯班	葛林斯	
monetary policy	貨幣政策	貨幣政策	貨幣政策	貨幣政策	
abnormal weather	天氣異常	天氣異常	天氣異常	天氣異常	天氣
National Council of Timorese Resistance	帝汶抵抗全國委員會	帝汶抵抗全國委員會	帝汶抵抗全國委員會	帝汶抵抗全國委員會	帝汶抵抗全國委員會

Computational Linguistics and Chinese Language Processing
Vol. 13, No. 1, March 2008, pp. 91-120

Improving Translation of Queries with Infrequent Unknown Abbreviations and Proper Names

Wen-Hsiang Lu*, Jiun-Hung Lin+, and Yao-Sheng Chang*

Abstract

Unknown term translation is important to CLIR and MT systems, but it is still an unsolved problem. Recently, a few researchers have proposed several effective search-result-based term translation extraction methods which explore search results to discover translations of frequent unknown terms from Web search results. However, many infrequent unknown terms, such as abbreviations and proper names (or named entities), and their translations are still difficult to be obtained using these methods. Therefore, in this paper we present a new search-result-based abbreviation translation method and a new two-stage hybrid translation extraction method to solve the problem of extracting translations of infrequent unknown abbreviations and proper names from Web search results. In addition, to efficiently apply name transliteration techniques to mitigate the problems of proper name translation, we propose a mixed-syllable-mapping transliteration model and a Web-based unsupervised learning algorithm for dealing with online English-Chinese name transliteration. Our experimental results show that our proposed new methods can make great improvements compared with the previous search-result-based term translation extraction methods.

Keywords: CLIR, Transliteration, Unknown Term Translation, Web Search Result, Machine Translation.

* Dept. of Computer Sci. and Eng., National Cheng Kung University, No.1, University Road, Tainan City 701, Taiwan (R.O.C.)
Tel.: +886-6-2757575 ext: 62545 Fax: +886-6-2747076
E-mail: whlu@mial.ncku.edu.tw, ys.chang1976@gmail.com
+ Penpower Technology Ltd. 7F, NO.47, Lane 2, Sec. 2, Guanfu Rd., Hsinchu City 300, Taiwan, R.O.C.
Tel: +886-3-5722691 Fax: +886-3-5716243
E-mail: hunter@penpower.com.tw

[Received June 30, 2007; Revised October 8, 2007; Accepted January 12, 2008]

1. Introduction

Many existing cross-language information retrieval (CLIR) systems [Ballesteros and Croft 1997; Hull and Grefenstette 1996] encounter great difficulties in dealing with unknown term translation since these systems rely mostly on general-purpose bilingual dictionaries, which usually lack translations of abbreviations and proper names. Moreover, according to the report in a previous work [Cheng *et al*. 2004], even for frequent Web queries, about 64% of them are not covered in an English-Chinese lexicon with about 120K entries (provided by Linguistic Data Consortium). However, several automatic translation extraction methods based on parallel [Brown *et al*. 1993; Melamed 2000; Nie *et al*. 1999; Smadja *et al*. 1996] or comparable corpora [Rapp 1999; Fung and Yee 1998] eventually suffer from the problems of insufficient parallel texts and the shortage of translation accuracy of comparable corpora in various subject domains.

The Web has been expanded with an enormous amount of multilingual hypertext resources in diverse subjects. Recently, a number of studies in natural language processing (NLP) have concentrated on the use of Web resources to complement insufficient text corpora [Cao and Li 2002; Kilgarriff and Grefenstette 2003]. To automatically collect huge amounts of parallel corpora from the Web in various domains, some researchers have developed feasible techniques of utilizing similar file names, text length, and link structures to extract parallel text pages from bilingual Web sites [Nie *et al*. 1999; Resnik 1999; Yang and Li 2003]. On the other hand, Lu *et al.* [2002] made the first attempt of mining unknown term translations from Web anchor texts. Both Cheng *et al*. [2004] and Zhang and Vines [2004] have explored language-mixed search-result pages for extracting translations of frequent unknown queries. Although these approaches have successfully enhanced the performance of frequent unknown query translation, they still suffer from the problems of data sparseness and indirect association errors in finding translations of infrequent unknown query terms, particularly for abbreviations and proper names [Melamed 2000].

In this paper, we focus on dealing with two kinds of translation of unknown query terms, including proper names and abbreviations. According to the report in Davis and Ogden [1998], about 50% of unknown terms in queries are proper names. Most methods handling translations of proper names are based on name transliteration techniques [Knight and Graehl 1998; Lin and Chen 2002; Lin *et al*. 2003; Li *et al*. 2004]. One major drawback of these methods is that they do not consider semantic information. Lam *et al*. [2004] proposed a named entity matching model, which considers both semantic and phonetic information, and applied it in mining unknown named entity translations from online daily Web news. Huang *et al*. [2005] also presented a method to extract key phrase translations from the language-mixed search-result pages with phonetic, semantic and frequency-distance features. As for abbreviation translation, less attention has been put on this research topic in the past few years.

Different from the above works, our major goal is to solve the problems of query translation to help users access English/Chinese information in cross-lingual Web searches. In this paper, therefore, we concentrate our attention on the challenge of dealing with the translations of infrequent unknown abbreviations and transliterated names in Web search queries, *i.e*., these unknown queries that appear infrequently in Web query logs. We present two new methods to effectively extract translations of these two kinds of infrequent unknown queries. First, we propose a search-result-based abbreviation translation method for handling bidirectional translation of abbreviations in Chinese/English. Second, a new two-stage hybrid translation extraction method, which combines Cheng *et al*.'s [2004] search-result-based term translation extraction method and a new Web-based transliteration method, is proposed to extract Chinese/English translations for infrequent unknown English/Chinese proper names. In addition, to train an effective transliteration model, we also present a Web-based unsupervised learning algorithm to automatically collect large amounts of diverse English-Chinese transliteration pairs from the Web. For application, we provide a real prototype website[1] for users to translate unknown terms in practice. Our experimental results show that the proposed new methods can make great improvements in extracting infrequent unknown term translation.

The remainder of this paper is organized as follows: Section 2 describes the problems of unknown term translation and our search-result-based term translation extraction approach. Section 3 evaluates the proposed approach. Section 4 provides a simple description and comparison with the related work. Section 5 gives our conclusions.

2. Search-Result-Based Unknown Term Translation

2.1 Problems

Cheng *et al*.'s search-result-based term translation extraction method (refer to Section 2.3) is effective in extracting translations for frequent unknown query terms. However, for a lot of infrequent abbreviations and proper names, their translations are still difficult to extract. For example, while submitting an English abbreviation "AMIA" to LiveTrans[2], an incorrect Chinese translation "系列" (series) is obtained. The reason might be that some abbreviations are semantically ambiguous and co-occur relatively infrequently with the correct Chinese translations of their full names (or original forms). However, we observe that for an English abbreviation, its full name may co-occur more frequently with its corresponding Chinese translation. Thus, to effectively extract correct translation for an infrequent abbreviation, our idea is to first identify its full name in search results, and then extract correct translation of its

[1] http://ws.csie.ncku.edu.tw/~jhlin/cgi-bin/index.htm

[2] http://livetrans.iis.sinica.edu.tw/: This website is developed based on the search-result-based term translation extraction method by Web Knowledge Discovery lab of Academia Sinica, Taiwan.

full name, using the search-result-based term translation extraction method mentioned above. Generally, it should be more feasible to extract the correct translation of an abbreviation via its full name. For example, if we can extract the full name of the abbreviation "AMIA", "American Medical Informatics Association", then we can get its correct Chinese translation "美國醫學資訊協會" via LiveTrans.

On the other hand, an English proper name might have multiple Chinese transliterated names which often vary with different translators due to phonetic variation and the lack of standard transliteration rules [Gao *et al*. 2004]. In other words, there may be several Chinese transliterated names corresponding to an English name. For example, the name "Disney" has various Chinese transliterated names, including "迪士尼", "迪斯尼", "迪斯奈", "狄斯奈", and "狄士尼"; the name "Hussein" also has several different Chinese transliterated names, including "海珊", "哈珊", and "侯塞因". Obviously, it will be helpful for query translation in cross-lingual Web search if we can collect all possible transliterated names from the Web for each unknown proper name. However, it is a real challenge to find all the various transliterated names. Thus, we consider integrating name transliteration techniques into the process of translation extraction for infrequent unknown proper names. Our idea is that we first extract high-frequency terms from the search-result pages as transliteration candidates, and then filter out impossible candidates by using a name transliteration model. In fact, it is still challenging to build an effective transliteration model while lacking sufficient transliteration pairs for training. Therefore, we propose a Web-based unsupervised learning algorithm to automatically collect large amounts of English-Chinese transliteration pairs from Web search results.

2.2 Overview of the Proposed Approach

Figure 1 demonstrates the process of our search-result-based query translation method. First, an unknown term is determined by a general-purpose dictionary. Then, an unknown term is recognized as an abbreviated term using our search-result-based abbreviation translation extraction methods. If the unknown term does not belong to an abbreviated term, we have to examine whether the unknown term is a transliteration based on our two-stage hybrid translation extraction method. To deal with unknown term translation, we employ the search-result-based term translation extraction method (described in Section 2.3) to handle translation of frequent (popular) unknown query terms, and propose two new infrequent unknown translation methods, namely the search-result-based abbreviation translation extraction method (Section 2.4) and two-stage hybrid translation extraction method (Section 2.5), to solve the problems of translation of abbreviated terms (*i.e*., abbreviations) and transliterated terms (*i.e*., proper names). To recognize the abbreviated terms in queries, we

collected an abbreviation list containing about 4K entries from the Wikipedia[3] website and then generated some pre-defined abbreviation patterns like those used in Park and Byrd (2001). Besides these, we used a Web-based transliteration model to recognize a transliterated term (Section 2.5).

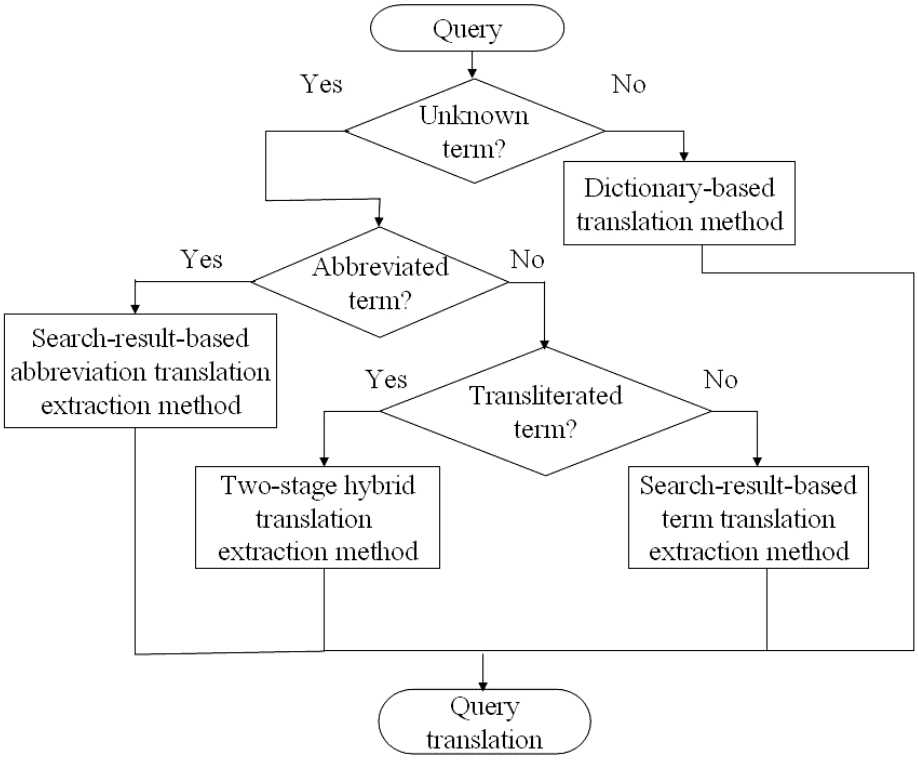

Figure 1. The process of our search-result-based query translation method

2.3 Search-Result-Based Term Translation Extraction Method

In this section, we will describe Cheng *et al.*'s [2004] search-result-based term translation extraction method, which explores search-result pages utilizing co-occurrence relation and contextual information for extraction of translations of unknown query terms.

(1) Chi-square Test Method

On the basis of co-occurrence analysis, chi-square test ($\chi 2$) is adopted to estimate semantic similarity between the source term *E* and the target translation candidate *C*. The similarity measure is defined as:

[3] http://en.wikipedia.org/wiki/List_of_acronyms_and_initialisms

$$S_{\chi^2}(E, \mathrm{C}) = \frac{N \times (a \times d - b \times c)^2}{(a+b) \times (a+c) \times (b+d) \times (c+d)}, \quad (1)$$

where a, b, c and d are the numbers of pages retrieved from search engines by submitting Boolean queries: "E and C", "E and not C", "not E and C", and "not E and not C", respectively; N is the total number of pages, i.e., $N = a + b + c + d$.

(2) Context-Vector Analysis Method

Due to the property of Chinese-English mixed texts often appearing in Chinese pages, the source term E and the target translation candidate C may share common contextual terms in the search-result pages. The similarity between E and C is computed based on their context feature vectors E_{cv} and C_{cv} in the vector-space model. The conventional tf-idf weighting scheme for each feature term t_i in E_{cv} and C_{cv}, $E_{cv} = <w_{e1}, w_{e2}, \ldots, w_{em}>$, and $C_{cv} = <w_{c1}, w_{c2}, \ldots, w_{cm}>$, is used and defined as:

$$w_{t_i} = \frac{f(t_i, p)}{\max_j f(t_j, p)} \times \log(\frac{N}{n}), \quad (2)$$

where $f(t_i, p)$ is the frequency of term t_i in the search-result page p, N is the total number of Web pages, and n is the number of the pages containing t_i. Finally, we use the cosine measure to estimate the similarity between E and C as follows:

$$S_{CV}(E, \mathrm{C}) = \frac{\sum_{i=1}^{m} w_{e_i} \times w_{c_i}}{\sqrt{\sum_{i=1}^{m} (w_{e_i})^2 \times \sum_{i=1}^{m} (w_{c_i})^2}}. \quad (3)$$

2.4 Search-Result-Based Abbreviation Translation Extraction Method

To effectively extract correct translations for infrequent abbreviated terms, we propose an integrated method in which an abbreviated term is transformed to its full name first, and then we extract the correct translation of the full name using the search-result-based term translation extraction method described above (Section 2.3). In the following, we describe two new proposed methods exploiting search results to extract full names for English and Chinese abbreviations, respectively.

2.4.1 Extracting Full Names for English Abbreviations

To deal with the full names for a given English abbreviation, we designed an efficient process of identifying full names, which consists of three major steps based on the hybrid text mining approach proposed by Park and Byrd [2001]. First, we use the contextual terms around an abbreviated term in the search results to extract possible full name candidates. Second, we use occurrence frequency and Part-of-Speech (POS) information of full name candidates to filter out some impossible candidates. Finally, we propose a simple adaptive co-occurrence model

which utilizes several different augmenting and decaying factors in selecting the best full name candidate. More details are described in the following.

(1) Identifying Full Name Candidates

To solve the problem of identifying full names without sufficient texts [Park and Byrd 2001], we take advantage of Web search results as a corpus. Our idea is to take the given abbreviated term as a search term to fetch the top 200 search result snippets from Google. To extract possible full name candidates by exploring the search result snippets, we utilize contextual information of the abbreviated term in the snippets. These full name candidates must appear in the same snippets with the abbreviated term, and should have a minimum word length between $|A|\times 2$ and $|A|+5$, where $|A|$ is the length of characters of the abbreviated term. In addition, to select more reliable full name candidates, we put a constraint on the identification process in which the first character of the first word of each full name candidate should match the first character of the abbreviated term.

(2) Filtering Impossible Full Name Candidates

To reduce computation time while extracting many full name candidates, we first select the top 20 frequent full name candidates and then filter out some impossible candidates whose first word or last word are prepositions, be-verbs, modal verbs, conjunctions, or pronouns [Park and Byrd 2001].

(3) Selecting Best Full Name Candidate

To select the best full name candidates, we propose an adaptive co-occurrence model by employing mutual information as well as four augmenting or decaying factors to compute the similarity between an abbreviated term A and its full name candidates F_C.

(A) Mutual Information: In this step, mutual information is used to compute the similarity between an abbreviated term A and its full name candidate F_C. Mutual information is defined as follows:

$$MI(A, F_C) = P(A, F_C) \times \log\left(\frac{P(A, F_C)}{P(A) \times P(F_C)}\right). \quad (4)$$

Here $P(A, F_C)$ is the probability of co-occurrence of A and F_C. $P(A)$ and $P(F_C)$ are the probabilities of occurrence of A and F_C in the Web, respectively. We can get the occurrence frequencies from search engines by submitting queries: "A", "F_C", and "A and F_C", respectively.

(B) Syntactic Cues: To augment the identification of full names, we utilize the information of orthographic and syntactic structure. N_{SC} indicates the number of abbreviation-full-name pairs appearing in the same snippets. Several frequent patterns of abbreviation-full-name pair are used as the syntactic cues [Park and Byrd 2001], including:

- abbreviation (full name)
- full name (abbreviation)
- abbreviation, **or** full name
- full name, **or** abbreviation
- full name, abbreviation **for short**
- abbreviation … **stands/short/acronym** …full name

(C) Similarity of Character: To further determine correct full names, we add another augmenting factor to estimate the similarity between an abbreviated term and its full name candidates by adopting a fast and simple character matching method. We use two kinds of character matching: (1) first-letter matching is used to compute the total number N_F of matching the first letter of each word in the full name candidate F_C with each character in the abbreviated terms, and (2) non-first-letter matching is used to computer the total number N_{NF} of matching the non-first letters of each word in the F_C with each character in A. The score of character matching of A and F_C is defined as:

$$Overlap\left(A, F_C\right) = \alpha * N_F + (1-\alpha) * N_{NF} . \tag{5}$$

Here, the weighting parameter α is empirically set to 0.8. Basically, the first-letter matching should be reasonably assigned higher weight for each matching pair. The character similarity is defined as follows:

$$CharSim(A, F_C) = \frac{Overlap(A, F_C)}{|A|}, \tag{6}$$

where $|A|$ is the number of characters of the abbreviated term A.

(D) Difference of Length: The number N_{LD} to represent the difference between character length $|A|$ of the abbreviated term A and word length $|F_C|$ of the corresponding full name candidate F_C as a decaying factor. N_{LD} is defined as follows:

$$N_{LD} = \left\| A \right| - \left| F_C \right\| . \tag{7}$$

(E) Number of Stop Words: The number N_{SW} of stop words in the full name candidate F_C is also used as a decaying factor.

(F) Adaptive Co-occurrence Model: We adaptively integrate the above two augmenting and two decaying factors into the basic co-occurrence model to compute the similarity between A and F_C. Our adaptive co-occurrence model is defined as follows:

$$S_{AC}(A, F_C) = \frac{MI(A, F_C) \times F_{Augument}}{F_{Decay}}, \tag{8}$$

where the augmenting factor $F_{Augument}$ is integrated as

$$F_{Augument} = CharSim\left(A, F_C\right) \times \left(\beta_1 + N_{SC}\right); \tag{9}$$

and the decaying factor F_{Decay} is integrated as

$$F_{Decay} = \left(\beta_2 + N_{LD} + N_{SW}\right). \tag{10}$$

To avoid the product being zero, here, β_1 and β_2 are the adaptable parameters and set to 1 heuristically.

2.4.2 Extracting Full Names for Chinese Abbreviations

Due to language differences between Chinese and English, such as no space delimitation between Chinese words, it is more difficult to identify the full name for a given Chinese abbreviated term. Therefore, we designed a method slightly different from the method of extracting English full names described above. Our Chinese full name extraction method consists of three major steps. First, the possible full name candidates are extracted by using the PAT-tree-based keyword extraction method proposed by Chien [1997]. Second, we use the character similarity between an abbreviated term and its full name candidates to filter out some impossible candidates. Finally, to select the correct Chinese full name, we use the adaptive co-occurrence model (Equation (8)) but slightly modify the decaying factors. The following description will explain the different points in more details.

(1) Identifying Full Name Candidates

To identify the possible full name candidates for a given Chinese abbreviated term A, we adopt a PAT-tree-based keyword extraction method [Chien 1997] to extract Chinese phrases in the search results related to the abbreviated term A as full name candidates. In addition, to select more reliable full name candidates, we put a length constraint on the candidates. These candidates should have more than ($|A|$ +2) characters, where $|A|$ is the number of characters of A.

(2) Filtering Impossible Full Name Candidates

According to our observations, the Chinese full name candidates extracted by the PAT-tree-based keyword extraction method generally have higher reliability. Thus, we just use Equations (5) and (6) with a threshold of character similarity to filter out some impossible candidates.

(3) Selecting Best Full Name Candidate

Like the above method of selecting the best English full name candidates, we still use the proposed adaptive co-occurrence model (Equation (8)) to select the best Chinese full name candidates. Please note, though, that the processing of augmenting/decaying factors is a little different. For example, we remove the decaying factor of stopword number since most stopwords seldom appear in Chinese full names. Some different points will be described below.

(A) Syntactic Cues: We also manually choose several syntactic patterns of Chinese abbreviation- full name pairs as the augmenting factor:

- abbreviation (full name)
- full name (abbreviation)
- abbreviation, 或 full name
- full name, 或 abbreviation
- abbreviation … 代表/簡稱/縮寫 …full name

Here the Chinese cues "或", "代表", "簡稱", "縮寫" correspond to the English words "or", "present", "short", and "acronym", respectively.

(B) Similarity of Character: First, we use the Chinese POS tagger to segment full name candidates. Then, we take character similarity (Equation (5) and (6)) as an augmenting factor.

(C) Difference of Length: Due to the fact that there is no space delimitation between Chinese words, we adopt a Chinese POS tagger[4] to do word segmentation for full name candidates. Then, we use the number N_{LD} to represent the difference between character length $|A|$ of the abbreviated term A and word length $|F_C|$ of the corresponding full name candidate F_C; this is considered a decaying factor (Equation (7)).

(D) Adaptive Co-occurrence Model: We adopt the same adaptive co-occurrence model (Equation (8)) with two augmenting factors and one decaying factor to compute the similarity between A and F_C. The augmenting factors are the same as Equation (9), but the decaying factor in Equation (10) is modified adaptively by removing the stopword number as:

$$F_{Decay} = (\beta_3 + N_{LD}). \tag{11}$$

To avoid the product being zero, here β_3 is an adaptable parameter and set to 1, heuristically.

2.5 Search-Result-Based Transliteration Name Extraction Method

To improve the performance of unknown term translation extraction for infrequent proper names, we consider integrating name transliteration techniques into the process of translation extraction in order to filter out impossible transliterated name candidates. Our idea is to first extract terms from the search-result snippets as translation candidates (see Section 2.3), and then filter out impossible transliterated name candidates based on the name transliteration model (described in Section 2.5.2). Therefore, in this section we propose a two-stage hybrid translation extraction method, a Web-based transliteration model to deal with transliteration mapping between an English proper name and its corresponding Chinese, and a Web-based

[4] http://ckipsvr.iis.sinica.edu.tw/demo.htm, which is a Chinese POS tagger developed by Chinese Knowledge and Information Processing group of Academia Sinica.

unsupervised learning algorithm to automatically collect diverse English-Chinese transliteration name pairs from Web search results for transliteration model training (Section 2.5.3).

2.5.1 Two-Stage Hybrid Translation Extraction

Our proposed two-stage hybrid translation extraction method is composed of two major steps. First, we use the search-result-based translation extraction method (Section 2.3) to extract k (k = 20) terms with higher similarity scores as transliteration candidates. Second, some impossible candidates included in general-purpose bilingual dictionaries are filtered out, and then the rest of the candidates are ranked according to transliteration similarity with the source proper name, which is computed based on the proposed Web-based transliteration model below (Equation (15)).

2.5.2 Filtering Impossible Candidates Using Web-Based Transliteration Model

(A) English Syllable Segmentation: Wan and Verspoor [1998] have developed a fully rule-based algorithm to transliterate English proper names into Chinese names. We simplify their syllabification techniques to generate a few simple heuristic rules of segmenting an English name into a sequence of syllables. Each English syllable is regarded as an English transliteration unit (ETU) in this work and has at most one corresponding character of the Chinese transliterated name. Initially, we used only five rules for English syllable segmentation listed below:

- a, e, i, o, u are vowels, and y is also regarded as a vowel if it appears behind a consonant. All other letters are consonants.
- Separate two consecutive vowels except the following cases: ai, au, ee, ea, ie, oa, oo, ou, etc.
- Separate two consecutive consonants except the following cases: bh, ch, gh, ph, th, wh, ck, cz, zh, zk, ng, sc,ll, tt, etc.
- l, m, n, r are combined with the prior vowel only if they are not followed by a vowel.
- A consonant and a following vowel are regarded as an ETU.

For example, “Nokia” (諾基亞) is segmented into three ETUs “no”, “ki”, and “a”, and “Epson” (愛普生) is segmented into three ETUs “e”, “p”, and “son”. Currently, although some English names may be segmented incorrectly, it is easy to manually update new rules to improve English syllable segmentation.

(B) Web-based Transliteration Model: To avoid double errors of converting English phonetic representation to Chinese Pinyin and from Pinyin to Chinese characters, in this work, we adopted direct orthographic mapping for name transliteration. We use the probability $P(e_i, c_i)$ to estimate the possibility of the mapping between an ETU e_i and a Chinese character c_i. Additionally, to build an efficient online name transliteration model, we propose a more simple transliteration

model. Our Web-based transliteration model is called forward-syllable-mapping transliteration model:

$$S_{FSM}(E,C)=\frac{P_{FSM}(E,C)}{D(E,C)}, \tag{12}$$

where $P_{FSM}(E, C)$ is the co-occurrence probability of E and C and defined as

$$P_{FSM}(E,C)\approx\prod_{i=1}^{\min(m,n)}[(1-\gamma_1)P(e_i,c_i)+\gamma_1], \tag{13}$$

and γ_1 is the smoothing weight. The decaying factor $D(E, C)$ indicates the number of syllable difference between an English name E and a Chinese transliterated name C and is defined as:

$$D(E,C)=\varepsilon+|m-n|. \tag{14}$$

Here ε is a decaying parameter, m is the total number of ETUs, and n is the total number of Chinese characters.

To improve incorrect transliteration mapping between ETUs and Chinese characters while an English-Chinese transliterated name pair with different numbers of transliteration unit, we propose the reverse-syllable-mapping transliteration model to assist in learning more correct mapping, which is defined below:

$$S_{RSM}(E,C)=\frac{P_{RSM}(E,C)}{D(E,C)}, \tag{15}$$

where

$$P_{RSM}(E,C)\approx\begin{cases}\prod_{i=m-n+1}^{m}[(1-\gamma_2)P(c_{i-(m-n)},e_i)+\gamma_2], & m\geq n;\\ \prod_{i=n-m+1}^{n}[(1-\gamma_2)P(c_i,e_{i-(n-m)})+\gamma_2], & m<n.\end{cases} \tag{16}$$

Here γ_2 is the smoothing weight and $D(E, C)$ is the same as Equation (14).

Our alternative transliteration model will combine the forward-syllable-mapping and reverse-syllable-mapping transliteration model, which is called **mixed-syllable-mapping transliteration model**, and defined as:

$$S_{MSM}(E,C)=\sqrt{S_{FSM}(E,C)\times S_{RSM}(E,C)}. \tag{17}$$

2.5.3 Web-Based Unsupervised Learning Algorithm

To deal with the problems of the diversity of Chinese transliterated names to English proper names, we intend to take advantage of abundant language-mixed texts on the Web to collect various English-Chinese transliterated name pairs from the Web and build a an effective online transliteration model. Thus, we designed an unsupervised learning process for English-Chinese transliterated name mapping. The process is composed of three main stages:

extraction of Chinese transliterated names, extraction of English original names, and learning of transliterated name mapping. More details are described below and the unsupervised learning algorithm is illustrated as well in Figure 2.

Web-based Unsupervised Learning Algorithm for Collecting English-Chinese Transliteration Pairs and Training a Transliteration Model

Input: initial transliterated name pair set V_{ec} and a general-purpose bilingual dictionary D.

Output: updating V_{ec} and a transliteration model T.

1 Extraction of Chinese transliterated names: select a transliterated name pair (E, C) from V_{ec}, two characters from the Chinese name C as seed characters V_c, and two corresponding English syllables from the English name E as seed ETUs V_e.

1.1 Search-result crawling: send the two selected Chinese seed characters V_c to a search engine and get search-result pages.

1.2 Chinese transliterated name identification: use a Chinese POS tagger to find unknown terms in the search-result pages, and then take the unknown terms containing the two seed characters V_c as potential Chinese transliterated names C_p.

2 Extraction of English original names: for each potential Chinese transliterated name C_p in V_c,, perform the following sub-steps:

2.1 Two-Stage hybrid translation extraction

2.1.1 English name candidate extraction: use search-result-based term translation extraction method to find English name candidates (see Section 2.3).

2.1.2 English name candidate filtering: first filter out impossible English name candidates included in D; second, compute transliteration mapping scores based on the English syllable segmentation rules and the name transliteration model T; third, choose the candidates with the highest scores as the possible English original names. Update V_{ec} by adding the new transliterated name pairs extracted.

2.2 Learning of transliterated name mapping: update T by computing the scores of transliterated name mapping of the new extracted transliterated name pairs (Equation (17)).

3 Repeat from step1 until the desired number of transliteration pairs is reached.

Figure 2. Web-based unsupervised learning algorithm for collecting English-Chinese transliterated name pairs and building a transliteration model

(1) Extraction of Chinese Transliterated Names: Xiao *et al*. [2002] have proposed a bootstrapping algorithm that uses only five frequent Chinese transliterated characters as initial seed character set: {阿, 爾, 巴, 斯, 基} to automatically collect over 100,000 Chinese transliterated names by utilizing search-result pages. Inspired by this work, we further propose a bootstrapping algorithm to automatically find English-Chinese transliterated name pairs from search-result pages. Initially, we need at least one English-Chinese transliterated name pair containing two frequent Chinese transliterated characters as seed transliteration pair set V_{ec}, *e.g*., V_{ec} = {(Bush, 布希)}. We select two Chinese characters from the Chinese name of the seed pair, and then send them to search engines for getting search-results pages. To efficiently extract more Chinese transliterated names from search-result pages, we use the CKIP tagger (Section 2.4.2), which is a representative Chinese POS tagger and performs well in segmenting Chinese texts into meaningful words and extracting unknown words.

(2) Extraction of English Original Names: We use the proposed two-stage hybrid translation extraction method described above (Section 2.5.1) to find possible English original names.

(3) Learning of Transliterated Name Mapping: On the basis of the rules of English syllable segmentation, we will gradually train an English-Chinese name transliteration model by computing the scores of the transliterated name mapping of the new extracted transliterated name pairs (Equation (17)).

3. Experimental Results

We conducted the following experiments to evaluate the performance of our proposed search-result-based abbreviation translation extraction method and two-stage hybrid translation extraction method.

Evaluation Metric: For the following experiments on full name identification of abbreviations and translation of abbreviations, the average top-*n* inclusion rate is adopted as a metric. For a set of abbreviated terms to be expanded/translated, its top-*n* inclusion rate was defined as the percentage of the abbreviated terms whose correct full names/translations could be found in the first *n* extracted full name candidates/translation candidates [Cheng *et al*. 2004].

Correct Translation / Transliteration: The correct translation / transliteration or correct definition is judged by us according to more popular sense in general cases.

3.1 Evaluation for the Search-Result-Based Abbreviation Translation Extraction Method

In this experiment, we intend to compare the performance of our proposed search-result-based abbreviation translation method with that of the search-result-based term translation extraction method proposed by Cheng *et al*. [2004].

3.1.1 Translation Extraction Results for English Abbreviations

Test data: Four test sets of English abbreviated terms are prepared in the following.

- FA-Dreamer-E: 28 frequent English abbreviated terms which have correct Chinese translations were manually selected from about 20K frequent queries with occurrence frequency over 10 in the Dreamer query log[5] which contains 228,566 unique queries. (The partial test data is listed in Appendix).
- IA-Dreamer-E: 27 infrequent English abbreviated terms (frequency < 3 in Dreamer query log) which have correct Chinese translations were manually selected from infrequent English queries in the Dreamer query log (about 40K entries). (The partial test data is listed in Appendix).
- FA-Wiki-E: 62 popular English abbreviated terms which have correct Chinese translations were manually selected from Wikipedia abbreviation list containing about 4k entries (Section 2.2). (The partial test data is listed in Appendix).
- RA-Wiki-E: 25 English abbreviated terms which have correct Chinese translations were randomly selected from Wikipedia abbreviation list due to the list without frequency information. (The partial test data is listed in Appendix).

(1) Results for English Full Name Extraction

Table 1 shows that our full name extraction method is effective for the test abbreviated terms with various subjects. Our method can achieve the top-1 inclusion rate of over 85% and the top-5 inclusion rate of over 92% for all test sets. Different from existing methods, our full name extraction method is very promising even for infrequent abbreviated terms by utilizing search results from Web search engines. However, some errors still result from the problem of data sparseness. For example, given the abbreviated term "MPEG", its correct full name "Motion Picture Experts Group" might appear quite rarely in the top 200 search results snippets. Therefore, the correct full name is filtered out by the filtering step and this causes trouble in extracting incorrect full names.

Table 1. Inclusion rates on full name extraction for different test sets of English abbreviated queries

Test Set	Inclusion Rates		
	Top-1	Top-3	Top-5
FA-Dreamer-E	93%	96%	96%
IA-Dreamer-E	85%	96%	96%
FA-Wiki-E	90%	94%	94%
RA-Wiki-E	88%	88%	92%

[5] http://www.dreamer.com.tw, which was a popular Chinese search engine and is closed now.

(2) Search-Result-based Abbreviation Translation Extraction Method vs. Search-Result-based Term Translation Extraction Method

Tables 2 to 5 show that the proposed search-result-based abbreviation translation extraction method actually performs better than the previous search-result-based translation extraction method proposed by Cheng *et al*. For example, for the infrequent English abbreviated queries from the Dreamer query log, the search-result-based abbreviation translation extraction method achieve the top-1 inclusion rate of 48% (see Table 3) but the search-result-based translation extraction method achieve the top-1 inclusion rate of 0%. Given the example query "ISS", the search-result-based term translation extraction method cannot obtain the correct Chinese translation "國際太空站" among the top five extracted candidates. However, our search-result-based abbreviation translation extraction method can extract the correct full name "International Space Station", and then extract correct Chinese translation "國際太空站" via the full name "International Space Station". As mentioned in Section 2.1, the reason might be that the abbreviated terms are semantically more ambiguous and co-occur relatively infrequently with the correct translations of their full names.

(3) Linear Combination Results

To further improve the performance of our search-result-based abbreviation translation extraction method, we intuitively intend to combine our method and Cheng *et al*.'s method. We expect that such a combination would make both methods mutually complementary by extracting translations from abbreviations and their full names simultaneously. Tables 2 to 5 show that the linear combination method is effective in improving the top-5 inclusion rate. For example, for the abbreviated query "AOL", its correct full name "America Online" is correctly extracted via our abbreviation expansion method. It fails to find the correct translation among the top five extracted candidates using our search-result-based abbreviation translation method, but the correct translation "美國線上" can be ranked at third place using the linear combination method.

Table 2. Inclusion rates on translation of frequent English abbreviations from Dreamer query log

Translation Extraction Method	Inclusion Rates		
	Top-1	Top-3	Top-5
Search-result-based Translation Extraction Method	43%	54%	57%
Search-result-based Abbreviation Translation Extraction Method	75%	82%	86%
Linear Combination	71%	82%	93%

Table 3. Inclusion rates on translation of infrequent English abbreviations from Dreamer query log

Translation Extraction Method	Inclusion Rates		
	Top-1	Top-3	Top-5
Search-result-based Translation Extraction Method	0%	19%	19%
Search-result-based Abbreviation Translation Extraction Method	48%	59%	63%
Linear Combination	44%	63%	67%

Table 4. Inclusion rates on translation of frequent English abbreviations from Wikipedia abbreviation list

Translation Extraction Method	Inclusion Rates		
	Top-1	Top-3	Top-5
Search-result-based Translation Extraction Method	24%	40%	44%
Search-result-based Abbreviation Translation Extraction Method	65%	79%	79%
Linear Combination	65%	77%	81%

Table 5. Inclusion rates on translation of randomly selected English abbreviations from Wikipedia abbreviation list

Translation Extraction Method	Inclusion Rates		
	Top-1	Top-3	Top-5
Search-result-based Translation Extraction Method	24%	36%	36%
Search-result-based Abbreviation Translation Extraction Method	64%	76%	76%
Linear Combination	64%	72%	80%

3.1.2 Translation Extraction Results for Chinese Abbreviations

Test data: Two test sets of Chinese abbreviated terms are prepared in the following.

- FA-Dreamer-C: 35 frequent Chinese abbreviated terms with correct English translations were manually selected from about 20K frequent queries with occurrence frequency over 10 in the Dreamer query log. (The partial test data is listed in Appendix).
- IA-Dreamer-C: 28 infrequent Chinese abbreviated terms (frequency < 3 in Dreamer query log) with correct English translations were manually selected from infrequent Chinese queries in the Dreamer query log (about 115K entries). (The partial test data is listed in Appendix).

(1) Results for Chinese Full Name Extraction

Table 6 shows that our Chinese full name extraction method is effective and can achieve top-1 inclusion rate of over 86% for the two test sets. We observed that some errors resulted from

incorrect matching between the abbreviated query terms and their highly related full name candidates in the search results. For example, given the abbreviated term "中影" (Central Motion Picture Corporation), our method extracted the incorrect full name "中國電影" (Chinese Movie) at first place. Since the correct full name "中央電影公司" co-occurs infrequently with the abbreviated query term "中影" in the search results, it can't be extracted by the PAT-tree-based keyword extraction method. As a result, our method extracted the incorrect full name "中國電影" because the abbreviated term "中影" and the incorrect full name candidate "中國電影" have stronger correlation in the search results and higher character similarity.

Table 6. Inclusion rates on full name extraction for two test sets of Chinese abbreviated queries

Test Set	Inclusion Rates		
	Top-1	Top-3	Top-5
FA-Dreamer-C	94%	100%	100%
IA-Dreamer-C	86%	89%	89%

(2) Performance Comparison between Search-Result-based Abbreviation Translation Extraction Method and Term Translation Extraction Method

Tables 7 and 8 show that, for the extraction of Chinese abbreviation translation, the proposed search-result-based abbreviation translation extraction method still performs better than the previous search-result-based translation extraction method proposed by Cheng *et al*. For example, for the infrequent Chinese abbreviated queries from the Dreamer query log, Cheng *et al*.'s method performs very poorly with a top-5 inclusion rate of 4%, but our method achieves great improvement with the top-5 inclusion rate of 29%. For example, given the Chinese abbreviated query "國安局", Cheng *et al*.'s method cannot obtain the correct English translation "National Security Bureau" among the top five extracted candidates. However, our method can extract the correct Chinese full name "國家安全局", and then extract the correct English translation "National Security Bureau", which is ranked at second place.

In addition, Table 8 shows that the linear combination method just achieves the same performance as our method, and is unable to further improve the top-*n* inclusion rates. In fact, we need larger amounts of test data to determine the effectiveness using the linear combination method in the future.

Table 7. Inclusion rates on translation of frequent Chinese abbreviations from Dreamer query log

Translation Extraction Method	Inclusion Rates		
	Top-1	Top-3	Top-5
Search-result-based Translation Extraction Method	17%	46%	54%
Search-result-based Abbreviation Translation Extraction Method	40%	66%	71%
Linear Combination	49%	63%	71%

Table 8. Inclusion rates on translation of infrequent Chinese abbreviations from Dreamer query log

Translation Extraction Method	Inclusion Rates		
	Top-1	Top-3	Top-5
Search-result-based Translation Extraction Method	4%	4%	4%
Search-result-based Abbreviation Translation Extraction Method	11%	21%	29%
Linear Combination	11%	21%	29%

3.2 Evaluation for the Two-Stage Hybrid Translation Extraction Method

The following two experiments are focused on the evaluation of the performance of extracting translations for infrequent unknown English and Chinese proper names, respectively, using the proposed mixed-syllable-mapping transliteration model and the two-stage hybrid translation extraction method.

3.2.1 Translation Extraction Results for English Proper Names

Test data: Two test sets of unknown English proper names are prepared, including:

- FP-Dreamer-E: 28 frequent unknown English proper names are manually selected from the 169 unknown terms out of the 430 frequent English queries in the Dreamer query log. (The partial test data is listed in Appendix).
- IP-Dreamer-E: 41 infrequent unknown English proper names (frequency < 3 in the query log) are manually selected from the Dreamer query log. (The partial test data is listed in Appendix).

(1) Two-Stage Hybrid Translation Extraction Method vs. Search-Result-based Term Translation Extraction Method

According to the results shown in Tables 9 and 10, we can obtain the following findings. For the two test sets, the proposed two-stage hybrid translation extraction method made great improvements compared with the search-result-based translation extraction method and the general name transliteration method [Wan and Verspoor 1998; Knight and Graehl 1998; Lin

and Chen 2002; Virga and Khudanpur 2003; Gao *et al*. 2004; Li *et al*. 2004]. In this work, we just use our proposed transliteration model as a "Name Transliteration" method for performance comparison. For example, the two-stage hybrid translation extraction method can achieve the top-1 inclusion rate of 41% (Table 10) for infrequent unknown English proper names, but the search-result-based translation extraction method only achieved 17%. The main reason is that most of the incorrect translation candidates extracted via the search-result-based translation extraction method can be filtered out based on our mixed-syllable-mapping transliteration model. For example, given the English proper name "Pamela", the correct Chinese transliterated name "潘蜜拉" can be extracted and ranked at second place (see Table 11).

(2) Linear Combination Results

Tables 9 and 10 also demonstrate that the simple linear combination method obtained slight improvement on transliterated name performance since the general name transliteration method is still limited in generating correct transliteration candidates. However, note that for many English-Chinese transliteration pairs with different numbers of transliteration units, the mixed-syllable-mapping transliteration model is still effective to learn correct transliteration mapping between English syllables and Chinese characters.

Table 9. Inclusion rates on translation of frequent unknown English proper names from Dreamer query log

Translation Extraction Method	Inclusion Rates		
	Top-1	Top-3	Top-5
Search-result-based Translation Extraction Method	32%	71%	82%
Name Transliteration	11%	18%	21%
Linear Combination	32%	50%	86%
Two-Stage Hybrid Translation Extraction Method	61%	64%	68%

Table 10. Inclusion rates on translation of infrequent unknown English proper names from Dreamer query log

Translation Extraction Method	Inclusion Rates		
	Top-1	Top-3	Top-5
Search-result-based Translation Extraction Method	17%	32%	37%
Name Transliteration	15%	15%	17%
Linear Combination	17%	37%	44%
Two-Stage Hybrid Translation Extraction Method	41%	46%	46%

Table 11. Effective results of translation extraction using the two-stage hybrid translation extraction method (underlined terms indicate correct translation)

Test Query	Translation Extraction Method	Top 5 Translation Candidates
Pamela	Search-result-based Translation Extraction Method	最後發表, 發表文章, 派米拉路, 發表, 討論區
	Name Transliteration	帕麥拉, 帕亞拉, 帕雲拉, 帕麥斯, 柏麥拉
	Linear Combination	最後發表, 發表文章, 帕麥拉, 派米拉路, 發表
	Two-Stage Hybrid Translation Extraction Method	彭美拉, 潘蜜拉, 派米拉路, 安德森, 尤德夫人

3.2.2 Translation Extraction Results for Chinese Proper Names

Test data: Two test sets of unknown Chinese proper names are prepared, including:

- FP-Dreamer-C: 28 frequent unknown Chinese proper names are obtained from the transliterated terms of the frequent unknown English proper name set FP-Dreamer-E (described in Section 3.2.1). (The partial test data is listed in the Appendix).
- IP-Dreamer-C: 41 infrequent unknown Chinese proper names are obtained from the transliterated terms of the infrequent unknown English proper name set IP-Dreamer-E (described in Section 3.2.1). (The partial test data is listed in the Appendix).

(1) Two-Stage Hybrid Translation Extraction Method vs. Search-Result-based Term Translation Extraction Method

Table 12 shows that our two-stage hybrid translation extraction method obtains the top-1 inclusion rate of 64%. Surprisingly, it performs worse than the search-result-based translation extraction method at 70%. This means that our candidate filtering method based on our trained Web-based transliteration model is unable to improve the performance of extracting translations for frequent unknown Chinese proper names in Web queries. We will investigate the possible reasons in the following discussion. However, for the test set of infrequent unknown Chinese proper names, the two-stage hybrid translation extraction method made effective improvements compared with the search-result-based translation extraction method (Table 13). For example, the two-stage hybrid translation extraction method can achieve the top-1 inclusion rate of 46% for infrequent unknown Chinese proper names, whereas the search-result-based translation extraction method only achieved 27%. It shows that most of the incorrect translation candidates extracted via the search-result-based translation extraction method can be filtered out using our mixed-syllable-mapping transliteration model. For

example, given the Chinese transliterated name "艾立克", its correct English original name "Eric" can be extracted and ranked at first place (Table 14).

Table 12. Inclusion rates on translation of frequent unknown Chinese proper names from Dreamer query log

Translation Extraction Method	Inclusion rates		
	Top-1	Top-3	Top-5
Search-result-based Translation Extraction Method	71%	89%	93%
Name Transliteration	14%	21%	25%
Linear Combination	71%	82%	86%
Two-Stage Hybrid Translation Extraction Method	64%	71%	75%

Table 13. Inclusion rates on translation of infrequent unknown Chinese proper names from Dreamer query log

Translation Extraction Method	Inclusion rates		
	Top-1	Top-3	Top-5
Search-result-based Translation Extraction Method	27%	44%	51%
Name Transliteration	12%	22%	22%
Linear Combination	27%	47%	57%
Two-Stage Hybrid Translation Extraction Method	46%	51%	51%

Table 14. Effective results of translation extraction using the two-stage hybrid translation extraction method (underlined terms indicate correct translation)

Test Query	Translation Extraction Method	Top 5 Translation Candidates
艾立克	Search-result-based Translation Extraction Method	Blog, Doll Edward, card, ebay, Eric Benet
	Name Transliteration	Elic, Eddoc, Alic, Addoc, <u>Eric</u>
	Linear Combination	Blog, Doll Edward, Elic, card, ebay
	Two-Stage Hybrid Translation Extraction Method	<u>Eric</u>, Alex, Eric idle, Clapton Eric, Eric Clapton Tears, KKBox Eric

(2) Discussion

According to our further analyses of the results shown in Tables 12 and 13, we obtain the following interesting findings.

- Our test set FP-Dreamer-C (frequent unknown Chinese transliterated terms) contains a number of company names, *e.g.*, "銳跑" (Reebok) and "新浪" (Sina). In fact, these Chinese characters like "銳", "跑", and "浪" are rarely used as transliterated characters in general cases. Thus, these characters are certainly difficult to be matched with those possibly correct ETUs since they have never appeared in the training data of our collected English-Chinese transliterated name pairs from search-result pages.
- The probabilities of some correct transliteration mapping between Chinese characters and English ETUs are lower than those of incorrect transliteration mapping trained from incorrect or partial matching transliteration pairs. However, our training data of about 10k potential transliterated name pairs extracted via our Web-based unsupervised learning algorithm should contain a number of incorrect transliteration mapping pairs and still be insufficient to build a good-quality transliteration model.
- The search-result-based term translation extraction method perform well for the test set of frequent unknown Chinese proper names while our two-stage hybrid translation extraction method is effective in improving the translation performance for infrequent unknown Chinese proper names. Therefore, we consider adding the information of term occurrence frequency in the query log into the process of unknown term translation. For a query with frequent Chinese proper names in the query log, we can use the previous search-result-based term translation extraction method to translate it. On the other hand, for queries with infrequent Chinese transliterated terms, we can use the proposed two-stage hybrid translation extraction method to translate them.

However, utilizing Web search results to translate unknown terms would lead to only partial representative candidates, which are the most popular ones. Therefore, we should continuously collect much more English-Chinese transliterated name pairs for training a better transliteration model in the future, and at the same time improve the techniques of extracting and filtering English name candidates to further collect larger amounts of correct transliterated name pairs for building a high quality transliteration model. In addition, there are still a number of cases which are difficult to be dealt with by using the simple mixed-syllable-mapping transliteration model and need to be further improved in the future.

4. Related Work

In previous works on identifying full names of abbreviations, AFP (Acronym Finding Program) [Taghva and Gilbreth 1995] used free texts to find English abbreviations and their full names. Park and Byrd [2001] used contextual information around abbreviations to extract potential full name candidates based on their pre-defined rules. However, these methods might suffer from the problem of insufficient texts. Our proposed method exploiting search results can extract English full names for abbreviations in various domains, and then effectively

extract correct Chinese translations via their full names.

Also, Leah *et al*. [2000] tried to find full name candidates from a small number of Web pages, and they used lots of syntax rules to select full name candidates of English acronyms. Instead of using many syntax rules, we propose an adaptive co-occurrence model to select the best full name candidates based on the co-occurrence relation and the integration of several augmenting and decaying factors.

For name transliteration between Latin-alphabet languages and some Asian languages with different writing forms, such as English and Chinese, researchers have proposed phoneme-based mapping techniques [Knight and Graehl 1998; Lin and Chen 2002; Meng *et al*. 2001]. Lin *et al*. [2003] proposed a statistical transliteration model and apply the model to extract English proper names and their Chinese transliterated names in a parallel corpus with high average precision and recall rates. However, Li *et al*. [2004] pointed out that the transliteration precision of the phoneme-based approaches could be limited by two main constraints. First, Latin-alphabet foreign names from different origins have different phonic rules, such as French and English. Second, transforming English words to Chinese characters will need two steps: transforming from phonemic representation to Chinese Pinyin and from Pinyin to Chinese characters. Two cascaded transforming steps may cause double errors. To avoid this problem, we propose a Web-based mixed-syllable-mapping transliteration model for dealing with online English-Chinese name transliteration based on the concept of direct orthographic mapping.

Both Cheng *et al*. [2004] and Zhang and Vines [2004] have exploited language-mixed search-result pages for extracting translations of frequent unknown queries. Moreover, Huang *et al.* [2005] takes advantage of cross-language query expansion to retrieve more relevant search-result pages and then extract translations by combining with phonetic, semantic and frequency-distance features. However, these methods haven't solved the problems of translation extraction for infrequent unknown abbreviations and proper names. Currently, our search-result-based methods presented in this paper can effectively mitigate such kinds of translation problems.

5. Conclusions

In this paper we presented two new search-result-based methods to extract unknown term translation based on the previous method proposed by Cheng *et al.*, including the search-result-based abbreviation translation extraction method and the two-stage hybrid translation extraction method. Our experimental results demonstrate the effectiveness of improving translation extraction for infrequent unknown abbreviations and proper names. Additionally, our proposed adaptive co-occurrence model is effective in aiding the process of selecting the correct full name candidates for the best abbreviated terms. However, currently,

the search-result-based abbreviation translation extraction method can perform well in the first stage of extracting the full names of those test abbreviated terms but can hardly extract correct translations via the extracted full names in the second stage. In the future, we are investigating to integrate the cross-language query expansion techniques proposed by Huang *et al*. into our search-result-based abbreviation translation extraction method.

As for the two-stage hybrid translation extraction method, we will continuously collect larger amounts of English-Chinese transliterated name pairs via our proposed Web-based unsupervised learning algorithm to build a more reliable transliteration model. In the future, referring to the methods proposed by both Lam *et al*. [2004] and Huang *et al*. [2005], we will extend our method by involving both semantic and phonetic information and expect that it can be more robust in extracting translations of unknown proper names.

References

Ballesteros, L. and W. B. Croft, "Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval," In *Proc. of 20th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1997, Philadelphia, USA, pp. 84-91.

Brown, P. F., S. A. D. Pietra, V. D. J. Pietra and R. L. Mercer, "The Mathematics of Machine Translation," *Computational Linguistics*, 19(2), 1993, pp. 263-312.

Cao, Y.-B. and H. Li, "Base noun phrase translation using Web data and the EM algorithm," In *Proc. of COLING*, 2002, Taipei, Taiwan, pp. 127-133.

Cheng, P.-J., J.-W. Teng, R.-C. Chen, J.-H. Wang, W.-H. Lu, L.-F. Chien, „Translating unknown queries with web corpora for cross-language information retrieval," In *Proc. of 27th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004, The University of Sheffield, UK, pp. 146-153.

Davis, M. W. and W. C. Ogden, "Free Resources and Advanced Alignment for Cross-Language Text Retrieval," In *Proc. of the Sixth Text Retrieval Conference* (*TREC 6*), 1998, Gaithersburg, Maryland, pp. 385-394.

Fung, P. and L.-Y. Yee, "An IR approach for translating new words from nonparallel, comparable texts," In *Proc. of 36th Annual Meeting of the Association for Computational Linguistics*, 1998, Montreal, Quebec, Canada, pp. 414-420.

Hull, D. A. and G. Grefenstette, "Querying across Languages: A Dictionary-based Approach to Multilingual Information Retrieval," In *Proc. of 19th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996, Zurich, Switzerland, pp. 49-57.

Gao, W., K.-F. Wong and W. Lam, "Phoneme-based Transliteration of Foreign Name for OOV Problem" In *Proc. of the first International Joint Conference on Natural Language Processing* (*IJCNLP*), 2004, Hainan Island, China, pp. 274-381.

Huang, F., Y. Zhang and S. Vogel, "Mining Key Phrase Translations from Web Corpora," In *Proc. of HLT-EMNLP*, 2005, Vancouver, B.C., Canada, pp. 483-490.

Kilgarriff, A. and G. Grefenstette, "Introduction to the special issue on the web as corpus," *Computational Linguistics*, 29(3), 2003, pp. 333-348.

Knight, K. and J. Graehl, "Machine Transliteration," *Computational Linguistics* 24(4), 1998, pp. 599-612.

Lam, W., R. Huang, P.-S. Cheung, "Learning phonetic similarity for matching named entity translations and mining new translations," In *Proc. of 27th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004, The University of Sheffield, UK, pp. 281-288.

Leah, L., P. Ogilvie, A. Price, and B. Tamilio, "Acrophile: An Automated Acronym Extractor and Server," In *Proc. of the 5th ACM Digital Libraries Conference*, 2000, San Antonio, TX, pp. 205-214.

Li, H., M. Zhang and J. Su, "A Joint Source-Channel Model for Machine Transliteration," In *Proc. of 42th Annual Meeting of the Association for Computational Linguistics*, 2004, Forum Convention Centre, Barcelona, pp. 160-167.

Lin, T., C.-C. Wu, J.-S. Chang, "Word-Transliteration Alignment," In *Proc. of ROCLING XV*, 2003, Hsinchu, Taiwan, pp. 1-16.

Lin, W.-H. and H.-H. Chen, "Backward machine transliteration by learning phonetic similarity," In *Proc. of CONLL*, 2002, Taipei, Taiwan, pp. 139-145.

Lu, W.-H., L.-F., Chien, H.-J. Lee, "Translation of Web Queries using Anchor Text Mining," *ACM Transactions on Asian Language Information Processing*, 1(2), 2002, pp. 159-172.

Ma, W.-Y. and K.-J. Chen, "A Bottom-up Merging Algorithm for Chinese Unknown Word Extraction," In *Proc. of ACL workshop on Chinese Language Processing*, 2003, pp. 31-38.

Melamed, I. D., "Models of translational equivalence among words," *Computational Linguistics*, 26(2), 2000, pp. 221-249.

Meng, H., W.-K. Lo, B. Chen and K. Tang, "Generate Phonetic Cognates to Handle Name Entities in English-Chinese Cross-Language Spoken Document Retrieval," In *Proc. of Automatic Speech Recognition and Understanding Workshop* (ASRU), 2001, Italy, pp. 311-314.

Nie, J.-Y., P. Isabelle, M. Simard, and R. Durand, "Cross-language Information Retrieval Based on Parallel Texts and Automatic Mining of Parallel Texts from the Web," In *Proc. of 22th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, University of California, Berkeley, pp. 74-81..

Park, Y. and R. J. Byrd, "Hybrid text mining for finding abbreviations and their definitions," In *Proc. of Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), 2001, pp. 126-133.

Rapp, R., "Automatic identification of word translations from unrelated English and German corpora," In *Proc. of 37th Annual Meeting of the Association for Computational Linguistics*, 1999, College Park, Maryland, USA, pp. 519-526.

Resnik, P., "Mining the Web for Bilingual Text," In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistic*,.1999, College Park, Maryland, USA, pp. 527-534.

Smadja, F., K. McKeown, and V. Hatzivassiloglou, "Translating collocations for bilingual lexicons: a statistical approach," *Computational Linguistics*, 22(1), 1996, pp. 1-38.

Taghva, K. and J. Gilbreth, "Recognizing Acronyms and their Definitions. Technical Report 95-03," *ISRI* (*Information Science Research Institute*), *UNLV*, June, 1995.

Virga, P. and S. Khudanpur, "Transliteration of Proper Names in Cross-Lingual Information Retrieval," *ACL 2003 workshop MLNER*.

Wan, S. and C. M. Verspoor, "Automatic English-Chinese name transliteration for development of multilingual resources," In *Proc. of 36th Annual Meeting of the Association for Computational Linguistics*, 1998, Montreal, Quebec, Canada, pp. 1352-1357.

Xiao, J., J. Liu and T.-S. Chua, " Extracting pronunciation-translated names from Chinese texts using bootstrapping approach," In Proc. of *the 1st SIGHAN workshop on Chinese Language Processing*, 2002, Taipei, Taiwan, pp. 1-6.

Yang, C. C. and K. W. Li, "Automatic Construction of English/Chinese Parallel Corpora, " *Journal of the American society for Information Science and Technology*, 54(8), 2003, pp. 730-742.

Zhang, Y. and P. Vines, "Using the web for automated translation extraction in cross-language information retrieval," In *Proc. of 27th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004, The University of Sheffield, UK, pp. 162-169.

Appendix

Partial English abbreviation test data

FA-Dreamer-E	IA-Dreamer-E	FA-Wiki-E	RA-Wiki-E
EDI	ADSM	ACM	NFL
ERP	AMIA	AMD	ABS
FMEA	ALSA	AOL	ACARS
TSMC	ATN	BBS	AGP
VLSI	BFI	CAD	ALTE
OTC	BGP	CDMA	BBS
VSAT	CGS	CEO	CICS
AIT	BSI	CMMI	DOM
CPR	CGMH	CS	DSP

Partial Chinese abbreviation test data

FA-Dreamer-C	IA-Dreamer-C
台銀 (Bank of Taiwan)	中影 (Central Motion Pictures Company)
日亞航 (Japan Asia Airways)	中選會 (Central Election Commission)
中信銀 (Chinatrust Commercial Bank)	智財權 (Intellectual Property Right)
勞保 (Labor Insurance)	台啤 (Taiwan Beer)
證交稅 (Securities Exchange Transaction Tax)	國衛院 (National Health Research Institutes)
竹科 (Hsinchu Science Park)	央銀 (Central Bank)
華航 (China Airlines)	兒福 (Child Welfare)
中研院 (Academia Sinica)	國台辦 (Taiwan Affairs Office of the State Council)
台大 (National Taiwan University)	客服 (Customer Service)

Partial English and Chinese transliteration test data

FP-Dreamer-E	IP-Dreamer-E	FP-Dreamer-C	IP-Dreamer-C
Alex	Athena	法拉利 (Ferrari)	雅典娜 (Athena)
Benz	Austin	古奇 (Gucci)	奧斯汀 (Austen)
Betty	Kournikova	辛吉斯 (Hingis)	庫妮可娃 (Kournikova)
Bosch	Bond	義大利 (Italy)	龐德 (Bond)
Calvin Klein	Brandy	肯尼 (Kenny)	布蘭蒂 (Brandy)
Ferrari	Charles	托福 (Tofel)	查爾斯 (Charles)
Gucci	David Robinson	泰迪 (Teddy)	大衛羅賓森 (David Robinson)
Hingis	Damon	茱蒂 (Judy)	達蒙 (Damon)
Italy	Duncan	迪士尼 (Disney)	鄧肯 (Duncan)

Computational Linguistics and Chinese Language Processing
Vol. 13, No. 1, March 2008, pp. 121-140

Analyzing Information Retrieval Results With a Focus on Named Entities

Thomas Mandl* and Christa Womser-Hacker*

Abstract

Experiments carried out within evaluation initiatives for information retrieval have been building a substantial resource for further detailed research. In this study, we present a comprehensive analysis of the data of the Cross Language Evaluation Forum (CLEF) from the years 2000 to 2004. Features of the topics are related to the detailed results of more than 100 runs. The analysis considers the performance of the systems for each individual topic. Named entities in topics revealed to be a major influencing factor on retrieval performance. They lead to a significant improvement of the retrieval quality in general and also for most systems and tasks. This knowledge, gained by data mining on the evaluation results, can be exploited for the improvement of retrieval systems as well as for the design of topics for future CLEF campaigns.

Keywords: Cross-Lingual Information Retrieval, Evaluation Issues, Named Entities (NEs)

1. Introduction

The Cross Language Evaluation Forum (CLEF) provides a forum for researchers in information retrieval and manages a testbed for mono- and cross-lingual information (CLIR) retrieval systems. CLEF allows the identification of successful approaches, algorithms, and tools in CLIR. Within CLEF, various strategies are employed in order to improve retrieval systems [Braschler and Peters 2004; di Nunzio *et al*. 2007].

We believe that the effort dedicated to large scale evaluation studies can be exploited beyond the optimization of individual systems. The amount of data created by organizers and participants remains a valuable source of knowledge awaiting exploration. Many lessons can still be learned from past data of evaluation initiatives such as CLEF, TREC [Voorhees and

* Information Sci., University of Hildesheim, Marienburger Platz 22, 31141 Hildesheim, Germany.
Tel.: +49-5121-883 ext: 837
The author for correspondence is Thomas Mandl.
E-mail: mandl@uni-hildesheim.de

[Received June 30, 2007; Revised October 8, 2007; Accepted January 12, 2008]

Buckland 2002], INEX [Fuhr 2003], NTCIR [Oyama *et al*. 2003], or IMIRSEL [Downie 2003].

Ultimately, further criteria and metrics for the evaluation of search and retrieval methods may be found. This could lead to improved algorithms, quality criteria, resources, and tools in cross language information retrieval [Harman 2004; Schneider *et al*. 2004]. This general research approach is illustrated in Figure 1.

Topics are considered an essential component of experiments for information retrieval evaluation [Sparck Jones 1995]. In most evaluations, the variation between topics is larger than the variation between systems. The topic creation for a multilingual test environment requires special care in order to avoid cultural or linguistic bias influencing the semantics of topic formulations [Kluck and Womser-Hacker 2002]. It must be assured that each topic provides equal conditions as starting points for the systems. The question remains whether linguistic aspects randomly appearing within the topics have any influence on the retrieval performance. This is especially important, as we observed in some cases, as leaving out one topic from the CLEF campaign changes the ranking of the retrieval systems despite the fact that 50 topics are considered to be sufficiently reliable [Voorhees and Buckley 2002; Zobel 1998].

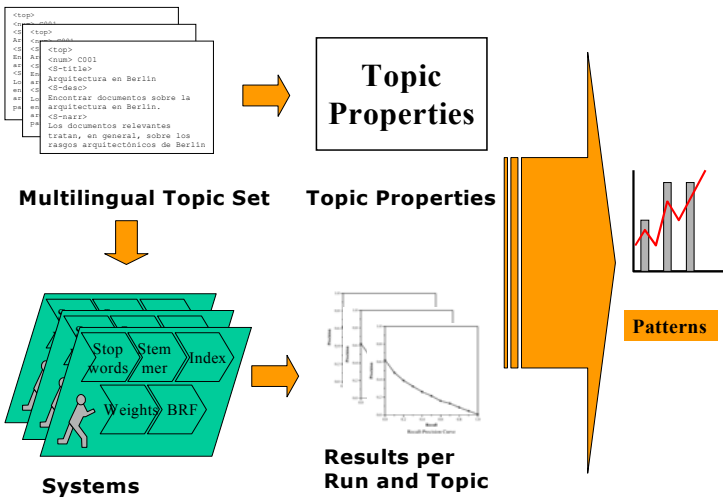

Figure 1. General overview of the research approach.

Most analysis of the data generated in CLEF is based on the average performance of the systems. This study concentrates on the retrieval quality of systems for individual topics. By identifying reasons for the failure of certain systems for some topics, these systems can be optimized. Our analysis identified a feature of the topics which can be exploited for future system improvement. In this study, we focused on the impact of named entities in topics and found a significant correlation with the average precision. Consequently, the goal of this study

is twofold:

(a) to measure the effect of named entities on retrieval performance in CLEF

(b) to optimize retrieval systems based on these results.

Named entities pose a potential challenge to cross language retrieval systems, because these systems often rely on machine translation of the query. The following problems may occur when trying to translate a named entity:

- The named entity may be out of vocabulary for translation
- Copying a named entity into the target language often does not help, as the name may be spelled differently (*e.g.* German: "Gorbatschow" vs. English: "Gorbachev")
- A named entity can actually be translated (*e.g.* "Smith" could be interpreted as a name or a profession and as the latter, translated)

Named entities are a feature which can be easily identified within queries. We consider the systems at CLEF as black boxes and have so far not undertaken any effort to analyze how these systems treat named entities and why that treatment may result in the effects we have observed. The data necessary for such an analysis is not provided by CLEF. The systems use very different approaches, tools and linguistic resources. Each system may treat the same named entity quite differently and successful retrieval may be due to a large number of factors like appropriate treatment as n-gram, proper translation by a translation service, or due to an entry in a linguistic resource. An analysis of the treatment of the named entities would lead merely to case studies. As a consequence, we find a statistical analysis of the overall effect as the appropriate research approach.

The remainder of this paper is organized as follows. The next chapter provides a brief overview of the research on evaluation results and their validity. Chapter three describes the data for CLEF used in our study. In chapter four, the influence of named entities on the overall retrieval results are analyzed. Chapter five explores the relationship between named entities and the performance of individual systems. In chapter six, we show how the performance variation of systems due to named entities could be exploited for system optimization.

2. Analysis of Information Retrieval Evaluation Results

The validity of large-scale information retrieval experiments has been the subject of a considerable amount of research. Zobel concluded that the TREC (Text REtrieval Conference) experiments are reliable as far as the ranking of the systems is concerned [Zobel 1998]. Voorhees and Buckley have analyzed the reliability of experiments as a function of the size of the topic set [Voorhees and Buckley 2002]. They concluded that the typical size of the topic set of some 50 topics in TREC is sufficient for a satisfactory level of reliability.

Human judgments are necessary to evaluate the relevance of the documents. Relevance assessment is a very subjective task. Consequently, assessments by different jurors result in different sets of relevant documents. However, these different sets of relevant documents do not lead to different system rankings according to an empirical analysis [Voorhees 2000]. Thus, the subjectivity of the jurors does not call into question the validity of the evaluation results.

Further research is dedicated toward the question of whether expensive human relevance judgments are necessary or whether the constructed document pool of the most highly ranked documents from all runs may serve as a valid approximation of the human judgments. According to a study by Cahan *et al*., the ranking of the systems in TREC correlates positively to a ranking based on the document pool without further human judgment [Cahan *et al*. 2001]. However, there are considerable differences in the ranking which are especially significant for the highest ranks.

Another important aspect in evaluation studies is pooling. Not all submitted runs can be judged manually by jurors and relevant documents may remain undiscovered. Therefore, a pool of documents is built to which the systems contribute differently. In order to measure the potential effect of pooling, a study was conducted which calculated the final rankings of the systems by leaving out one run at a time [Braschler 2003]. It shows that the effect is negligible and that the rankings remain stable.

However, our analysis shows that leaving out one topic during the result calculation changes the system ranking in most cases. It has also been noted that the differences between topics are larger than the differences between systems. This effect has been observed in TREC [Harman and Voorhees 1997] and also in CLEF [Gey 2001].

For example, when looking at run EIT01M3N in the CLEF 2001 campaign, we see that it has a fairly good average precision of 0.341. However, for one topic (nr. 44), which had an average difficulty, this run performs far below (0.07) the average for that topic (0.27). An intellectual analysis of the topics revealed that two of the most difficult topics contained no proper names and that both topics were from the sports domain (Topic 51 and 54). This effect has been noted in many evaluations and also in CLEF [Hollink *et al*. 2004]. As a consequence, topics are an important part of the design in an evaluation initiative and need to be created very carefully.

Named entities seem to play an important role especially in multilingual information retrieval [Gey 2001]. This assumption is backed by experimental results. The influence of named entities on the retrieval performance is considerable. In an experiment, the removal of named entities from the topic decreased the quality considerably, whereas the use of named entities only in the query led to a much smaller decrease [Demner-Fushman and Oard 2003].

A study for the CLEF campaign 2001 revealed no strong correlation between any single linguistic phenomenon and the system difficulty of a topic. Not even the length of a topic showed any substantial effect, except for named entities. However, the sum of all phenomena was correlated to the performance. The more linguistic phenomena available, the better the systems solved a topic on average [Mandl and Womser-Hacker 2003]. The availability of more variations of a word seems to provide stemming algorithms with more evidence for extraction of the stem, for example.

3. Named Entities in the Multi-lingual Topic Set

The data for this study stems from the Cross Language Evaluation Forum (CLEF) [Peters *et al*. 2003; Peters *et al*. 2004]. CLEF is a large evaluation initiative which is dedicated to cross-language retrieval for European languages. The setup is similar to the Text Retrieval Conference (TREC) [Harman and Voorhees 1997; Voorhees and Buckland 2002]. The main tasks for multilingual, ad-hoc retrieval are:

- The core and most important track is the **multilingual** task. The participants choose one topic language and need to retrieve documents in all main languages. The final result set needs to integrate documents from all languages ordered according to relevance regardless of their language.
- The **bilingual** task requires the retrieval of documents different from the chosen topic language.
- The **Monolingual** task represents the traditional ad-hoc task in information retrieval and is allowed for some languages.

All runs analyzed in this study are test runs based on topics for which no previous relevance judgments were known. For training runs, older topics can be used each year. Techniques and algorithms for cross-lingual and multilingual retrieval are described in the CLEF proceedings and are not the focus of this paper.

The topic language of a run is the language which the system developers use to start the search and to construct their queries. The topic language needs to be stated by the participants and can be found in the appendix of the CLEF proceedings. The retrieval performance of the runs for the topics can also be extracted from the appendix of the CLEF proceedings [Peters *et al.* 2003; Peters *et al*. 2004]. Most important, the average precision of each run for each topic can be retrieved.

3.1 Topic Creation Process

The topic creation for CLEF needs to assure that each topic is translated into all languages without modifying the content while providing equal chances for systems which start with

different topic languages. Therefore, a thorough translation check of all translated topics in CLEF was performed to check if the translations to all languages resulted in the same meaning. Nevertheless, the topic generation process follows a natural method and avoids artificial constructions [Womser-Hacker 2002].

Figure 2 shows an exemplary topic from CLEF containing a named entity. The topic's structure is built up by a short title, a description with a few words and a so-called narrative with one or more sentences. Participants of CLEF have to declare which parts are used for retrieval.

```
<top lang="ES"> <num>C083</num>
<ES-title> Subasta de objetos de Lennon. </ES-title>
<ES-desc> Encontrar subastas públicas de objetos de John Lennon.</ES-desc>
<ES-narr> Los documentos relevantes hablan de subastas que incluyen objetos que
pertenecieron a John Lennon, o que se atribuyen a John Lennon.</ES-narr>
</top>   <top> <num>C083</num>
<FR-title> Vente aux enchères de souvenirs de John Lennon </FR-title>
<FR-desc> Trouvez les ventes aux enchères publiques des souvenirs de John Lennon.
</FR-desc>
<FR-narr> Des documents pertinents décriront les ventes aux enchères qui incluent les objets
qui ont appartenu à John Lennon ou qui ont été attribués à John Lennon. </FR-narr> </top>
```

Figure 2. Example of a CLEF topic with a named entity

3.2 Data

An intellectual analysis of the results and the properties of the topics had identified named entities as a potential indicator of good retrieval performance. For that reason, named entities in the CLEF topic set were analyzed in more detail.

Named entities were intellectually assessed according a published schema [Sekine *et al*. 2002]. The analysis included all topics from the campaigns in the years 2000 through 2004. The number of named entities in each topic was assessed intellectually. We focused on English, Spanish, and German as topic languages and considered monolingual, bilingual, and multilingual tasks.

Table 1 shows the overall number of named entities found in the topic sets. The extraction was done intellectually by graduate students. We also assessed in which parts of the topic the name occurred, whether found in the title, the description, or the narrative. This detailed analysis was not exploited further because very few runs use a source other than title plus description. In very few cases, the topic narrative includes additional named entities not already present in the title and the description. For our analysis, the sum of named entities in all three parts was used. We analyzed the topic set in three languages, and in some cases, differences between the number of named entities between two versions of a topic occur.

These differences were considered. In 18 cases, a different number of named entities was assessed between German and English versions of topics 1 through 200, and in 49 cases, a difference was encountered between German and Spanish for topics 41 though 200. For example, topic 91 contains one named entity more for German because German has two potential abbreviations for United Nations (UN and UNO) and both are used.

The numbers given in Table 1 are based on the English versions of the topics and consider the number of types rather than tokens of named entities in title, description, and narrative together.

Table 1. Number of named entities in the CLEF topics

CLEF year	Number of topics	Total number of named entities	Average number of named entities in topics	Standard deviation of named entities in topics
2000	40	52	1.14	1.12
2001	50	60	1.20	1.06
2002	50	86	1.72	1.54
2003	60	97	1.62	1.18
2004	50	72	1.44	1.30

Table 2. Overview of named entities in CLEF tasks

CLEF year	Task	Topic language	Nr. runs	Topics without named entities	Topics with one or two named entities	Topics with more than three named entities
2001	Bi	German	9	16	24	7
2001	Multi	German	5	16	24	7
2001	Bi	English	3	16	24	7
2001	Multi	English	17	17	26	7
2002	Mono	German	21	12	21	17
2002	Mono	Spanish	28	11	18	21
2002	Bi	German	4	12	21	17
2002	Multi	German	4	12	21	17
2002	Bi	English	51	14	21	15
2002	Multi	English	32	14	21	15
2003	Mono	Spanish	38	6	33	21
2003	Multi	Spanish	10	6	33	21
2003	Mono	German	30	9	40	10
2003	Bi	German	24	9	40	10
2003	Bi	English	8	9	41	10
2003	Multi	English	74	9	41	10
2004	Multi	English	34	16	23	11

The large number of named entities in the topic set shows their importance. Table 2 shows the number of runs within each task. For the analysis presented in chapter five, we divided the topics into three classes: (a) no named entities, (b) one or two named entities, and (c) three or more named entities. The distribution of topics over these three classes is also shown in Table 2. It can be seen that the three classes are best balanced in CLEF 2002, whereas topics in the second class dominate in CLEF 2003.

Only topics for which no zero results were returned were considered for each sub-task. Since these topics differ between sub-tasks, there are slight differences between the numbers for each class even for one year. For further analysis, only tasks with more than eight runs were considered.

4. Named Entities and General Retrieval Performance

Our first goal was to measure whether named entities had any influence on the overall quality of the retrieval results. In order to measure this effect, we first calculated the correlation between the overall retrieval quality achieved for a topic and the number of named entities encountered in this topic. In the second section, this analysis is refined to single tasks and specific topic languages.

4.1 Correlation Between Average Precision and Number of Named Entities

Table 3. Method a: Best run for each topic in relation to the number of named entities in the topic

Number of named entities	0	1	2	3	4	5
Number of Topics	42	43	40	20	9	4
Average of Best System per Topic	0.62	0.67	0.76	0.83	0.79	0.73
Minimum of Best System per Topic	0.09	0.12	0.04	0.28	0.48	0.40
Standard Deviation of Best System per Topic	0.24	0.24	0.24	0.18	0.19	0.29

Table 4. Method b: Average precision of runs in relation to the number of named entities in the topic

Number of named entities	0	1	2	3	4	5
Number of Topics	42	43	40	20	9	4
Minimum of Average Performance per Topic	0.02	0.04	0.01	0.10	0.17	0.20
Average of Average Performance per Topic	0.20	0.25	0.36	0.40	0.31	0.40
Maximum of Average Performance per Topic	0.54	0.61	0.78	0.76	0.58	0.60
Standard Deviation of Average Performance	0.14	0.15	0.18	0.17	0.14	0.19

First, we determined the overall performance in relation to the number of named entities in a topic. The 200 analyzed topics contain between zero and six named entities. For each number *n* of named entities, we determine the overall performance by two methods: (a) take the best run for each topic and (b) take the average of all runs for a topic. For both methods, we obtain a set of values for *n* named entities. Within each set, we can determine the maximum, the average, and the minimum. For example, we determine for method (a) the following values: best topic for *n* named entities, average of all topics for *n* named entities, and worst topic among all topics with *n* named entities. The last value gives the performance for the most difficult topic within the set of topics containing *n* named entities. The maximum of the best runs is in most cases 1.0 and is, therefore, omitted. The following Tables 3 and 4 show these values for CLEF overall. Figures 3 and 4 show detailed analysis for specific tasks.

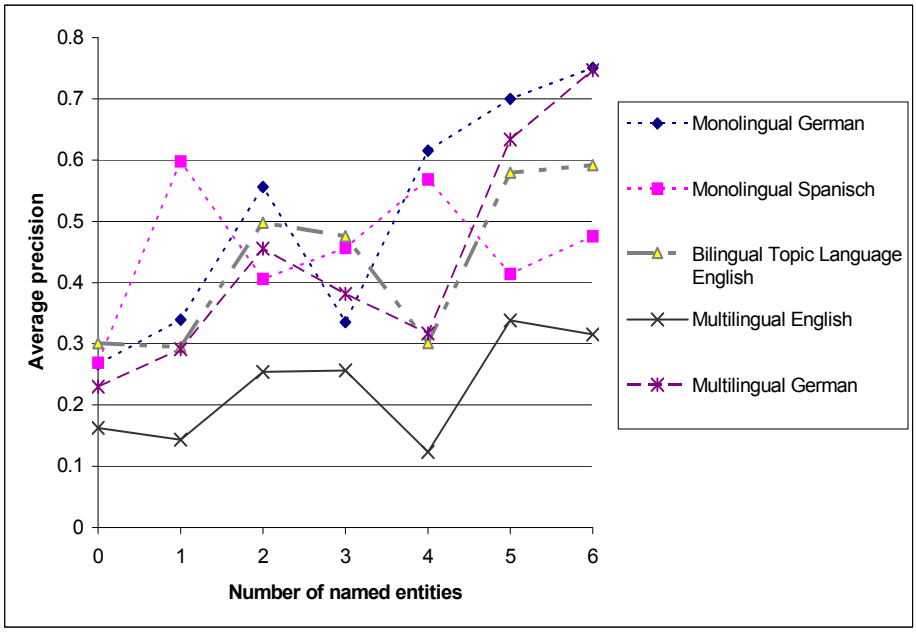

Figure 3. Method a: Average precision for topics with n named entities for CLEF 2002

The CLEF campaign contains relatively few topics with four or more named entities. The results for these values are, consequently, not significant.

It can be seen that topics with more named entities are generally solved better by the systems. This observation can be confirmed by statistical analysis. The average performance correlates to the number of named entities with a value of 0.43 and the best performance with a value of 0.26. Both correlation values are statistically significant at a level of 95%. With one exception, the worst performing category is always the one without any named entities.

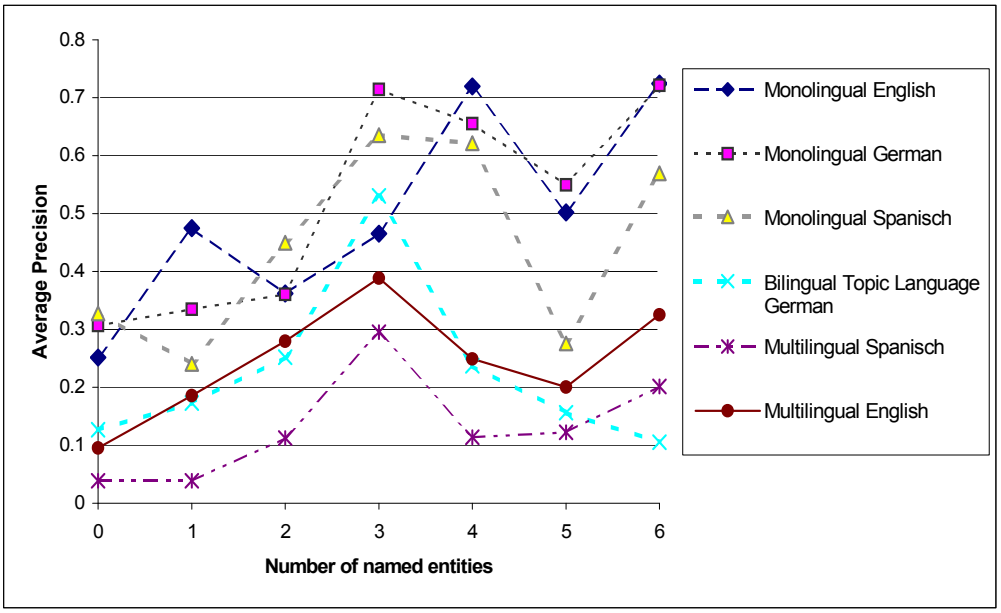

Figure 4. Method b: Relation between system performance and the number of named entities in CLEF 2002

4.2 Correlation for Individual Tasks and Topic Languages

The correlation analysis was also carried out for the individual retrieval tasks or tracks. This can be done by (a) calculating the average precision for each topic achieved within a task, by (b) taking the maximum performance for each topic (taking the maximum average precision that one run achieved for that topic), and by (c) calculating the correlation between named entities and average precision for each run individually and taking the average for all runs within a task. Both measures a and b are presented in Table 5. Except for one task (multilingual with topic language English in 2001), all observed correlations are positive. Thus, the overall effect occurs within most tasks and even within most single runs.

There is no difference in the average strength of the correlation for German (0.27) and English (0.28) as topic language. The average for each language in the last column shows a more significant difference. The correlation is stronger for German (0.19) than for English (0.15) as topic language. Furthermore, there is a considerable difference between the average correlation for the bilingual (0.35) and multilingual run types (0.22). This could be a hint that the observed positive effect of named entities on retrieval quality is smaller for multilingual retrieval.

Table 5. Correlation of system performance and number of named entities for different tasks

CLEF year	Run type	Topic language	Num-ber of runs	(a) Correlation of average precision per topic to number of NEs	Level of statistical significance (t-distribution) for prev. column	(b) Correlation of max. precision per topic to nr. of NEs
2001	Bilingual	German	9	0.44	-	0.32
2001	Multilingual	German	5	0.19	-	0.24
2001	Bilingual	English	3	0.20	-	0.13
2001	Multilingual	English	17	-0.34	-	-0.36
2002	Bilingual	German	4	0.33	-	0.25
2002	Multilingual	German	4	0.43	-	0.41
2002	Bilingual	English	51	0.40	99%	0.36
2002	Multilingual	English	32	0.29	-	0.37
2002	Monolingual	German	21	0.45	95%	0.34
2002	Monolingual	Spanish	28	0.21	-	0.27
2003	Bilingual	German	24	0.21	-	0.10
2003	Bilingual	English	8	0.41	-	0.47
2003	Multilingual	English	74	0.31	99%	0.27
2003	Monolingual	German	30	0.37	95%	0.28
2003	Monolingual	Spanish	38	0.39	99%	0.33
2003	Monolingual	English	11	0.16	-	0.24
2003	Multilingual	Spanish	10	0.21	-	0.31
2004	Multilingual	English	34	0.33	95%	0.34

It needs to be stressed, though, that the effect does not only occur for systems with overall poor performance. Rather, it can be observed in the top ranked runs as well. Figure 5 shows the strength of the correlation for all runs in one task. The runs are ordered according to their average precision. The correlation between the systems MAP for a topic and the number of named entities present in that topic is also shown in Figure 5.

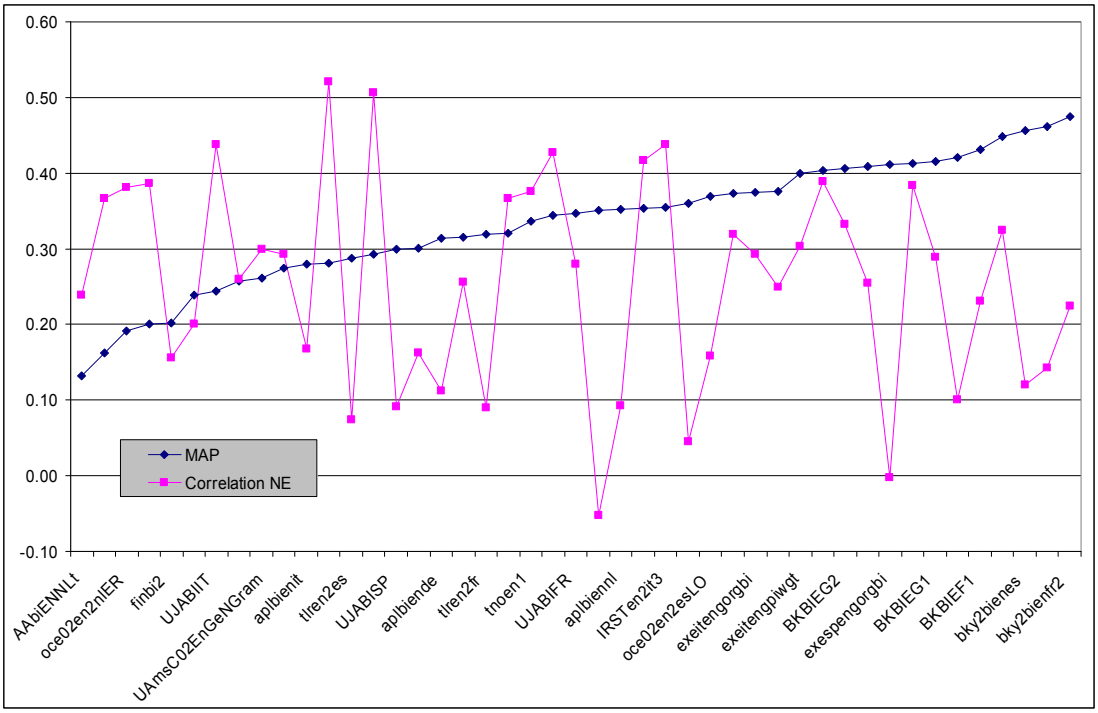

Figure 5. Correlation between named entities and performance for runs in CLEF 2002 (task bilingual, topic language English)

5. Conclusion Performance Variation of Systems for Named Entities

In this chapter, we show that the systems tested at CLEF perform differently for topics with different numbers of named entities. Although proper names make topics easier in general, and for almost all runs, the performance of systems varies within the three classes of topics based on the number of named entities. As already mentioned, we distinguished three classes of topics: (a) the first class without proper names (called "none"), (b) the second class with one or two named entities (called "few"), and (c) a third class with three or more named entities (called "lots"). This approach is suitable for implementation and allows the categorization before the experiments and the relevance assessment. It requires no intellectual intervention but, solely, a named entity recognition system.

5.1 Variation of System Performance

As we can see in Table 2, the three categories are well balanced for the CLEF campaign in 2002. For 2003, there are only few topics in the first and second categories. Therefore, the average ranking is extremely similar to the ranking for the second class "few".

Figure 5 shows that the correlation between average precision and the number of named entities is quite different for all runs for one exemplary task. The runs in Figure 6 are ordered

according to the original ranking in the task. We observe a slightly decreasing sensitivity for named entities with higher system performance. However, the correlation is still substantial and sometimes still high for top runs.

A look at the individual runs shows large differences between the three categories. We show the values for three tasks in Figure 6. The curve for many named entities lies mostly above the average curve, whereas the average precision for the class none without named entities in most cases remains below the overall average. Sometimes, even the best runs perform quite differently for the three categories. Other runs perform similarly for all three categories.

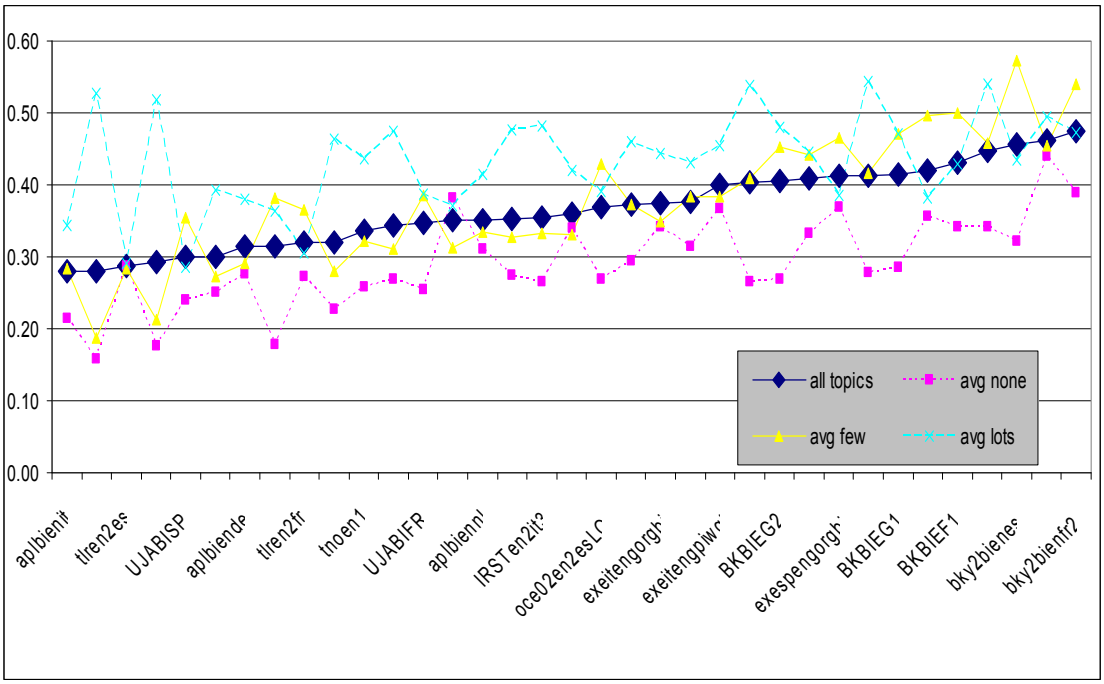

Figure 6. Performance variation of runs in CLEF 2002 (task bilingual, topic language English) depending on number of named entities in topic

5.2 Correlation of System Rankings

The performance variation within the classes leads to different system rankings for the classes. An evaluation campaign including, for example, only topics without named entities may lead to different rankings. To analyze this effect, we determined the rankings for all runs within each named entity class, *none*, *few,* and *lots*. Table 6 shows that the system rankings can be quite different for the three classes. The difference is measured with the Pearson rank correlation coefficient.

For most tracks, the original average system ranking is most similar to the ranking based only on the topics with one or two named entities. For the first and second categories, the rankings are more dissimilar. The ranking for the top ten systems in the classes usually differs more from the original ranking. This is due to minor performance differences between top runs.

Table 6. Correlation of full system ranking to ranking based on topic sub-set

Sub-Task				Topic sub-set		
CLEF year	Run type	Topic language	Number of runs	No NEs	few NEs	lots NEs
2001	Bilingual	German	9	0.92	0.93	0.92
2001	Multilingual	English	17	0.98	0.93	0.75
2002	Bilingual	English	51	0.88	0.93	0.74
2002	Multilingual	English	32	0.94	0.99	0.98
2003	Bilingual	German	24	0.81	0.99	0.91
2002	Multilingual	English	74	0.86	1.00	0.93

These findings are not always statistically significant because each category contains only few topics. As stated by Buckley and Voorhees, some 50 topics are necessary to create a reliable ranking [Buckley and Voorhees 2002].

6. Optimization by Fusion Based on Named Entities

The patterns of the systems are strikingly different for the three classes. As a consequence, there seems to be potential for the combination or fusion of systems.

We propose the following simple fusion rule. For each topic, the number of named entities is determined. Subsequently, this topic is channeled into the system with the best performance for this named entity class. The best system is a combination of at most three runs. Each category of topics is answered by the optimal system for that number of named entities. By simply choosing the best performing system for each topic, we can also determine a practical upper level for the performance of the retrieval systems. This upper level can give a hint about how much of the potential for improvement is exploited by an approach. Table 6 shows the optimal performance and the improvement by the fusion based on the optimal selection of a system for each category of topics.

The highest levels of improvement are achieved for the topic language English. For the year 2002, we observe the highest improvement of 10% for the bilingual runs. For this task, there is also the highest figure for potential, 53%. Figure 7 shows the results of the optimization.

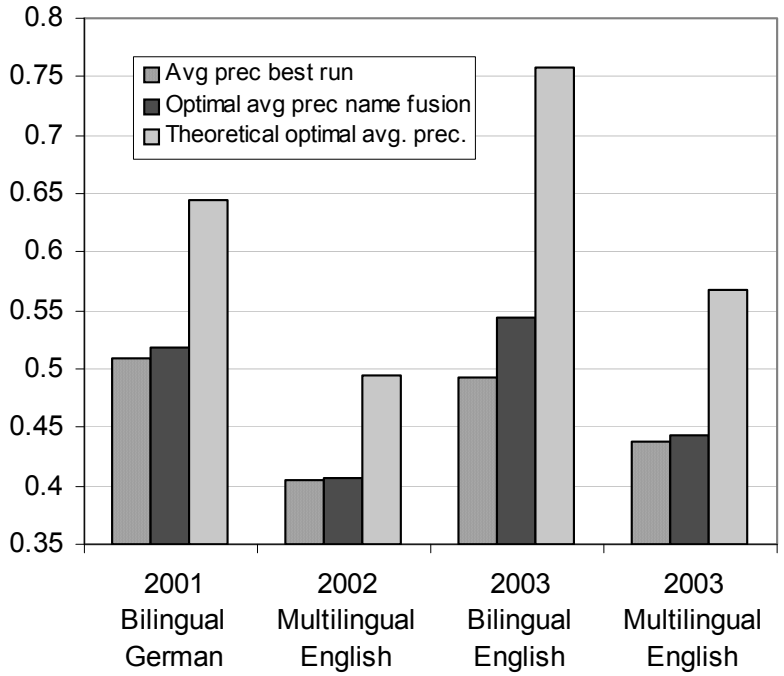

Figure 7. Optimization potential of named entity based fusion

Table 7. Improvement by fusion based on named entities for several tasks

CLEF year	Run type	Topic language	Average precision best run	Optimal average precision name fusion	Improve-ment over best run	Practical optimal average precision.	Improve-ment over best run
2001	Bilingual	German	0.509	**0.518**	**2%**	0.645	27%
2001	Multilingual	English	0.405	**0.406**	**0%**	0.495	22%
2002	Bilingual	English	0.4935	**0.543**	**10%**	0.758	53%
2002	Multilingual	English	0.378	**0.403**	**6.5%**	0.456	21%
2003	Bilingual	German	0.460	**0.460**	**0%**	0.622	35%
2003	Bilingual	English	0.348	**0.369**	**6.1%**	0.447	28%
2003	Multilingual	English	0.438	**0.443**	**1.2%**	0.568	30%

The previous analysis showed that our fusion approach has the potential to boost even top runs. Consequently, this technique may also be beneficial for lower-ranked runs. We applied the optimization through fusion for all runs. In the ordering of all runs according to the average precision (original CLEF ranking), we chose a window of three and five neighboring runs. From these three to five runs, we chose the best results for each of the three classes of number of proper names (none, few, or lots). Again, the best run for each class is chosen and

contributes to the fusion result. Table 6 shows the average improvement for this fusion technique. This analysis shows that the performance of retrieval systems can be optimized by channeling topics to the systems best appropriated for topics with none, one or two and three and more proper names. Certainly, the application of this fusion on the past results approach is artificial and, in our study, the number of named entities was determined intellectually. However, this mechanism can be easily implemented by using an automatic named entity recognizer.

7. Named Entities in Topics and Retrieval Performance for Target Languages

So far, our studies have been focused on the language of the initial topic which participants used for their retrieval efforts. Additionally, we have analyzed the effect of the target or document language. In this case, we cannot consider the multilingual tasks where there are several target languages. However, the monolingual tasks have already been analyzed and are also considered here. The additional analysis is targeted at bilingual retrieval tasks. We grouped all bilingual runs with English, German, and Spanish as document languages. The correlation between the number of named entities in the topics and the average precision of all systems for that topic was calculated. The average precision may be interpreted as the difficulty of the topic. Table 8 shows the results of this analysis.

Table 8. Correlation for target languages for CLEF 3 and 4

CLEF year	Task type	Target language	Number of runs	Correlation between number of named entities and average precision
2003	Mono	English	11	0.158
2002	Bi	English	16	0.577
2003	Bi	English	15	0.187
2002	Mono	German	21	0.372
2003	Mono	German	30	0.449
2002	Bi	German	13	0.443
2003	Bi	German	3	0.379
2002	Mono	Spanish	28	0.385
2003	Mono	Spanish	38	0.207
2002	Bi	Spanish	16	0.166
2003	Bi	Spanish	25	0.427

First, we can see a positive correlation for all tasks considered. Named entities support the retrieval also from the perspective of the document language. These results for the year 2002 may be a hint that retrieval in English or German document collections profits more

from named entities in the topic than Spanish. However, in 2003, the opposite is the case and English and Spanish switch. For German, there are only 3 runs in 2003. As a consequence, we cannot yet detect any language dependency for the effect of named entities on retrieval performance.

8. Resume

Research on failure and success stories for individual topics is a promising strategy for the analysis of information retrieval results. Several current research initiatives are focusing on this strategy and are looking at retrieval results beyond average precision [Harman 2004; SIGIR 2005 query difficulty workshop]. We identified named entities in topics as one transparent predictor in multi- and mono-lingual retrieval. Further analysis on named entities should also take the frequency and distribution of the named entities in the corpora into account.

References

Allan, J., and H. Raghavan, "Using part-of-speech Patterns to Reduce Query Ambiguity," In *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR ´02),* Tampere, Finland, Aug. 11-15, 2002, pp. 307-314.

Braschler, M., "CLEF 2002 - Overview of Results," *Evaluation of Cross-Language Information Retrieval Systems. Third Workshop of the Cross Language Evaluation Forum CLEF 2003),* Trondheim. Springer (Lecture Notes in Computer Science).

Braschler, M., and C. Peters, "Cross-Language Evaluation Forum: Objectives, Results, Achievements," *Information Retrieval,* 2004, 7, pp. 7-31.

Cahan, P., C. Nicholas, and I. Soboroff, "Ranking Retrieval Systems without Relevance Judgments," In *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '01),* New Orleans, USA, Sep. 9-13, 2001, pp. 66-73.

Cronen-Townsend, S., Y. Zhou, and B. Croft, "Predicting Query Ambiguity," In *Proceedings of the Annual Intl. ACM Conference on Research and Development in Information Retrieval (SIGIR '02),* Tampere, Finland, 2002, pp. 299-306.

Demner-Fushman, D., and D. Oard, "The effect of bilingual term list size on dictionary-based crosslanguage information retrieval," In *Thirty-Sixth Hawaii International Conference on System Sciences,* (Hawaii, Jan 6-9, 2003).

Di Nunzio, G., N. Ferro, T. Mandl, and C. Peters, "CLEF 2006: Ad Hoc Track Overview," In *Evaluation of Multilingual and Multi-modal Information Retrieval. 7th Workshop of the Cross-Language Evaluation Forum, (CLEF 2006),* Alicante, Spain, Revised Selected Papers. Berlin et al.: Springer (Lecture Notes in Computer Science 4730) 2007, pp. 21-34.

Diaz, F., and R. Jones, "Using Temporal Profiles of Queries for Precision Prediction," In *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '04),* Sheffield, UK, 2004, pp. 18-24.

Downie, S., "Toward the Scientific Evaluation of Music Information Retrieval Systems," In *International Symposium on Music Information Retrieval (ISMIR),* Washington, D.C., and Baltimore, USA 2003, http://ismir2003.ismir.net/papers/Downie.PDF.

Evans, D., J. Shanahan, and V. Sheftel, "Topic Structure Modeling," In *Proceedings of the Annual Intl. ACM Conference on Research and Development in Information Retrieval (SIGIR '02),* Tampere, Finland, 2002, pp. 417-418.

Fuhr, N., *Initiative for the Evaluation of XML Retrieval (INEX): INEX 2003 Workshop Proceedings*, Dagstuhl, Germany, December 15-17, 2003. http://purl.oclc.org/NET/duett-07012004-093151.

Gey, F., "Research to improve Cross-Language Retrieval. Position Paper for CLEF," In *Cross-Language Information Retrieval and Evaluation. Workshop of Cross-Language Evaluation Forum (CLEF 2000),* Lisbon, Portugal, September 21-22, 2000). Berlin et al.: Springer [LNCS 2069] 2001, pp. 83-88.

Hackl, R., R. Kölle, T. Mandl, A. Ploedt, J.-H. Scheufen, and C. Womser-Hacker, "Multilingual Retrieval Experiments with MIMOR at the University of Hildesheim," In *Evaluation of Cross-Language Information Retrieval Systems. Proceedings CLEF 2003 Workshop*, Trondheim, Norway, Revised Selected Papers. Berlin et al.: Springer [LNCS 3237] 2004, pp. 166-173.

Harman, D., "SIGIR 2004 Workshop. RIA and Where can we go from here?," *ACM SIGIR Forum,* 38(2), pp. 45-49.

Harman, D., and E. Voorhees, "Overview of the Sixth Text REtrieval Conference," In *The Sixth Text REtrieval Conference (TREC-6),* National Institute of Standards and Technology, Gaithersburg, Maryland, 1997, http://trec.nist.gov/pubs/.

Hollink, V., J. Kamps, C. Monz, and M. de Rijke, "Monolingual Document Retrieval for European Languages," *Information Retrieval,* 7(1-2), pp. 33-52.

Kluck, M., and C. Womser-Hacker, "Inside the Evaluation Process of the Cross-Language Evaluation Forum (CLEF): Issues of Multilingual Topic Creation and Multilingual Relevance Assessment," In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002),* Las Palmas de Gran Canaria, Spain, May 29-31, 2002, ELRA, Paris, 2002, pp. 573-576.

Lempel, R., and S. Moran, "Predictive Caching and Prefetching of Query Results in Search Engines," in *Proceedings of the Twelfth International World Wide Web Conference (WWW 2003),* Budapest, Hungary, May 20-24, 2003. pp. 19-28.

Mandl, T., and C. Womser-Hacker, "Linguistic and Statistical Analysis of the CLEF Topics," In *Advances in Cross- Language Information Retrieval: Third Workshop of the Cross-Language Evaluation Forum (CLEF 2002),* Rome, Italy, September 19-20, 2002 Springer, LNCS 2785. 2003 pp. 505-511.

Mandl, T., and C. Womser-Hacker, "A Framework for long-term Learning of Topical User Preferences in Information Retrieval," *New Library World,* 105(5/6), pp. 184-195.

Oyama, K., E. Ishida, and N. Kando, *NTCIR Workshop 3: Proceedings of the Third NTCIR Workshop on research in Information Retrieval, Automatic Text Summarization and Question Answering,* Tokio, 2003.
http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/index.html

Peters, C., M. Braschler, J. Gonzalo, and M. Kluck, "Evaluation of Cross-Language Information Retrieval Systems," In *Third Workshop of the Cross Language Evaluation Forum (CLEF 2002),* Rome. Berlin et al.: Springer (Lecture Notes in Computer Science 2785) 2003.

Peters, C., J. Gonzalo, M. Braschler, and M. Kluck, "Comparative Evaluation of Multilingual Information Access Systems," In *4th Workshop of the Cross-Language Evaluation Forum (CLEF 2003),* Trondheim, Norway, August 21-22, 2003, Revised Selected Papers. Springer Lecture Notes in Computer Science 3237, 2004.

Schneider, R., T. Mandl, and C. Womser-Hacker, "Workshop LECLIQ: Lessons Learned from Evaluation: Towards Integration and Transparency in Cross-Lingual Information Retrieval with a special Focus on Quality Gates," In 4th *Intl Conference on Language Resources and Evaluation (LREC),* Lisbon, Portugal, May 24-30, 2004. Workshop Lessons Learned from Evaluation: Towards Transparency and Integration in Cross-Lingual Information Retrieval (LECLIQ), pp. 1-4.

Sekine, S., K. Sudo, and C. Nobata, "Extended Named Entity Hierarchy," In: *Proceedings of Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands, Spain.

Soboroff, I., C. Nicholas, and P. Cahan, "Ranking retrieval systems without relevance judgements," In *Proceedings of the 24th Annual International ACM SIGIR Conference of Research and Development in Information Retrieval*, New Orleans, pp. 66-73.

Sparck, J.K., "Reflections on TREC," *Information Processing and Management*, 31(3), pp. 291-314.

Voorhees, E., and C. Buckley, "The Effect of Topic Set Size on Retrieval Experiment Error," In *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '02),* Tampere, Finland, Aug. 11-15, 2002, ACM Press, pp. 316-323.

Voorhees, E., "Variations in Relevance Judgements and the Measurement of Retrieval Effectiveness," *Information Processing and Management*, 36(5), 2000, pp. 679-716.

Voorhees, E., and L. Buckland, *The Eleventh Text Retrieval Conference (TREC 2002),* National Institute of Standards and Technology, Gaithersburg, Maryland. Nov. 2002.
http://trec.nist.gov/pubs/trec11/t11_proceedings.html.

Womser-Hacker, C., "Multilingual Topic Generation within the CLEF 2001 Experiments," In *Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum (CLEF 2001),* Peters C.; Braschler M.; Gonzalo, J.

and Kluck, Michael (Eds.). 2002, Darmstadt, Germany, September 3-4, 2001. Springer, LNCS 2406, pp. 389-393.

Zobel, J., "How Reliable are the Results of Large-Scale Information Retrieval Experiments?" In *Proceedings of the 21st Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '98),* Melbourne, Australia, 1998, pp. 307-314.