# The Effects of High Quality Translations of Named Entities in Cross-Language Information Exploration


**Dan WU**
**School of Information Management, Wuhan University**
**Wuhan, Hubei, China**
**woodan@whu.edu.cn**

**Daqing HE**
**School of Information Sciences, University of Pittsburgh**
**Pittsburgh, PA, USA**
**dah44@pitt.edu**

**Heng JI, Ralph GRISHMAN**
**Computer Science Department, New York University**
**New York, NY, USA**
**{hengji, grishman}@cs.nyu.edu**



## Abstract:

**Named entities (NEs) are the expressions in human languages that explicitly link notations in languages to the entities in the real world. They play important role in cross-language information retrieval (CLIR) because most users' requests have been found to have NEs, and majority of out-of-vocabulary terms are NEs. Therefore, missing their translations has a significant impact to the retrieval effectiveness. In this paper, we examined the effect of high quality translations of NEs in event driven information exploration, where the existence of NEs is even more common. With the focus on the effect of NE translations obtained by using information extraction (IE) techniques, we conducted several experiments using TDT test collections. Our results demonstrate that NEs and their translations play critical roles in improving CLIR effectiveness, and it makes positive impact in CLIR to use high quality translations of NEs obtained by IE techniques.**



## Keywords:

**Named entity; cross-language information exploration**


## 1. Introduction

Retrieving information cross different languages often means that translation is used as the technique to cross the language barriers between a query and the documents. Because of its flexibility and effectiveness, dictionary based query translation approach has been the dominant method for cross-language information retrieval (CLIR) [13]. In this approach, a dictionary, manually constructed or automatically generated, is used to provide possible translations to each query term or phrase. However, with the dictionary's limited coverage, it is possible that some query words cannot find corresponding translations in the dictionary, which results in these query terms failing to contribute to locating relevant documents. These words are called out-of-vocabulary (OOV) terms, and their effects in CLIR could be significant. Experiments show that the retrieval effectiveness can reduce up to 60% if OOV terms are common in the search topics and they are not handled properly [4].

OOV terms can be many types, some of them are newly formed words, loan words, abbreviations, or domain specific terms, but the biggest group of OOV terms, which was observed to be as many as half of the whole observed OOV terms in [4], belongs to a group called named entities (NEs). They are the expressions in human languages that explicitly link notations in languages to the entities in the real world. Common examples of NEs are proper nouns like Ming Yao, locations like Beijing, and brand names like Sony.

NEs form a very dynamic set; there already exists a large quantity of them, and at the same time people are creating new NEs every day. This is the reason that, no matter how large the coverage of a dictionary is, many NEs still do not have their translations in the dictionary. At the same time, NEs are critical information in search requests. Not only almost all search topics in cross-language evaluation forum (CLEF) 2000 – 2003 contain at least one or several NEs [15], these NEs plays important roles in locating relevant documents. When correct translations of NEs are available, it seems that "the occurrence of named entities in topics makes them 'easier' for retrieval systems" [15].

All these studies have been concentrated on generic ad hoc retrievals. Our interests are concentrated on one specific type of information explorations that are driven by a seminal event. Here, the information needs and the corresponding relevance judgments are heavily influenced by the event. This kind of information exploratory search is typical for a range of professional users, such as intelligence analysts. After an event happens, a "Request For Information" (RFI), which typically contains one overall investigation goal and a set of more specific questions that call for more information related to the event, is issued. The analyst's job is to collect relevant and useful information from various sources to answer the RFI questions. The complexity of event driven information exploration is further increased since the event is frequently evolving during the exploration. Because the information can come from different sources in multiple languages, this type of


978-1-4244-2780-2/08/$25.00 ©2008 IEEE

information exploration often are multilingual. In addition, since events are always associated with various locations, people names, organizations, and numbers, NEs should be even more critical for capturing the characteristics of the search topics and results.

The research presented in this paper aims to examine the effect of NEs in event driven cross-language information exploration. Specifically, we are interested in the effect of NEs and their high quality translations obtained by information extraction (IE) techniques. The research objectives in this study are:

- Although all these exploration topics are event driven, does the increase of the number of NEs indeed reflect in the human generated topic statements for these types of searches? To answer this question, our study is on TDT collections[1] that were built for this type of event driven information exploration. The comparison will between the search topics on TDT collections and that of cross-language evaluation forum (CLEF) collections, which are built for generic CLIR retrievals.
- Our second objective is to study the usage of information extraction (IE) techniques for obtaining NE translations. Our hypothesis is that since NEs are so important in event driven information exploration, the dedicated IE component, which gives higher quality of NE identification and translations, would further improve the retrieval effect of cross-language information exploration than simply rely on a dictionary and simple OOV handling methods.

In the remainder of this paper, we will first review the related work on NE in CLIR in section 2; then in section 3, we talk about our methods of obtaining NE translations using an IE component. Then, we will talk about the integration of NE translations in CLIR in section 4 and a group of experiments are conducted to find answers to the research objectives in section 5. Finally, in section 6, we will conclude with discussions and future works.

## 2. Related Work

Out-of-vocabulary terms have been treated as a serious problem in dictionary-based CLIR [12]. Although most widely used dictionaries are automatically generated through large parallel corpora, OOV terms are still common exist. Demner-Fushman and Oard [4] examined the effect of out-of-vocabulary terms in CLIR by artificial degradation of the dictionary coverage. They find that handling OOV terms can play significant role in improving CLIR retrieval effectiveness. Because about half of the OOV terms are named entities, the performance can decrease up to 60% when NEs are removed from the translations [4]. By reviewing the search topics and retrieval systems in several years of Cross-Language Evaluation Forum (CLEF) experiments, Mandl and his colleagues examined the NEs and their effects in CLIR [15]. They find that majority of CLEF topics contain at least one NEs, and NEs often make retrieval topics relative easier to obtain good retrieval effectiveness than those topics that do not have NEs. Of course, here their assumption is that reasonable translations can be found for these NEs.

Many methods have been developed for obtaining translations for NEs. The simplest way is to look into dictionaries for the translations of NEs. However, many NEs are OOV terms, so based on the assumption that many NEs and their translation could sound similar, transliteration methods have been introduced [11, 16, 19]. People also develop techniques and tools for identifying translations for NEs using Web mining [9], and machine translation [7].

Research and practice in CLIR have moved from the classic ad-hoc retrieval scenarios to embracing new challenges and applications. NE and other shallow NLP processing would play an important role in this new kind of CLIR applications. Pablo et al [14] reported multilingual named entity processing in cross lingual question answering and in web cross-language information retrieval. Chen et al [6] dealt with query translation issue in CLIR, proper names in particular. Their results showed that only 0.79% and 1.11% of candidates for English person names and location names, respectively have to be proposed during name translation.

## 3. Information Extraction Component for Named Entities and Their Translations

The IE component is designed to provide two functions in our CLIR process. The first one is to identify named entities in a given text, which could be queries, documents, or any parts of queries and documents. The NE identification uses the NYU English and Chinese HMM-based name taggers trained on several years of ACE (Automatic Content Extraction) corpora. Both name taggers can identify seven types of names (Person, GPE, Location, Organization, Facility, Weapon and Vehicle) and achieve about 87%-90% F-measure on newswire.

The English HMM NE tagger [5] includes six states for each of the seven name types, as well as a not-a-name state. These six states correspond to the token preceding the name; the single name token (for names with only one token); the first token of the name; an internal token of the name (neither first nor last); the last token of the name; and the token following the name. These multiple states allow the HMM to capture context information and information about the internal structure

[1] http://projects.ldc.upenn.edu/TDT/

of the name.

The Chinese name tagger consists of a HMM tagger augmented with a set of post-processing rules. The HMM tagger generally follows the Nymble model [2]. Within each of the name class states, a statistical bi-gram model is employed, with the usual one-word-per-state emission. The various probabilities involve word co-occurrence, word features, and class probabilities. Since these probabilities are estimated based on observations seen in a corpus, several levels of "back-off models" are used to reflect the strength of support for a given statistic, including a back-off from words to word features, as for the Nymble system. To take advantages of Chinese names, we extend a model to include a larger number of states, 14 in total. The expanded HMM can handle name prefixes and suffixes, and has separate states for transliterated foreign names. Finally a set of post-processing heuristic rules are applied to correct some omissions and systematic errors.

The second function of the IE component is to handle the translations of identified NEs. As stated in Section 1, a bottleneck in building high-performance cross-lingual IR system is that many NEs do not appear in the translation resources, such as the bilingual dictionary. Therefore, many NEs would not be translated or be mistranslated if without special handling of NEs. Our IE component exploits the NYU name translation system [8], which uses the following methods to locate possible translations from a variety of resources:

- Extracting cross-lingual name titles from Wikipedia pages. We run a web browser [1] to extract titles from Chinese Wikipedia pages and their corresponding linked English pages (if the link exists). Then we apply heuristic rules based on Chinese name structure to detect name pairs, for example, foreign full names must include a dot separator, Chinese person names must include a last name from a closed set of 437 family names.
- Tagging NEs in parallel corpora. Within each sentence pair in a parallel corpus, we run the NE taggers on both sides. If the types of the NE tags on both sides are identical, we extract the NE pairs from this sentence. Then at the corpus-wide level, we count the frequency for each NE pair, and only keep the NE pairs that are frequent enough. Each member of the NE pair then becomes the translation of the other member. The corpora used for this approach were all GALE MT training corpora and ACE 07 Entity Translation training corpora. We didn't use word alignment information because the state-of-the-art Chinese-English word alignment is only about 60%.
- Using patterns for Web mining. We constructed heuristic patterns such as "Chinese name (English name)" to extract NE pairs from Web pages with mixed Chinese and English.

## 4. Applying Named Entity Translations to CLIR

We viewed the application of IE component for NE translations as an extra step in query translation. As shown in Figure 1, with the help of NE tagger, we first markup all English NEs inside queries, then, we translated NEs using the NE translation resources presented in Section 3. Because the IE component does not provide weights to differentiate the multiple translations for a NE, we developed some method to calculate weights for each translation (more detail in Section 4.1). In addition, if any NE cannot find its translations in the NE resources, we used Web to mine their possible translations (more detail in Section 4.2).

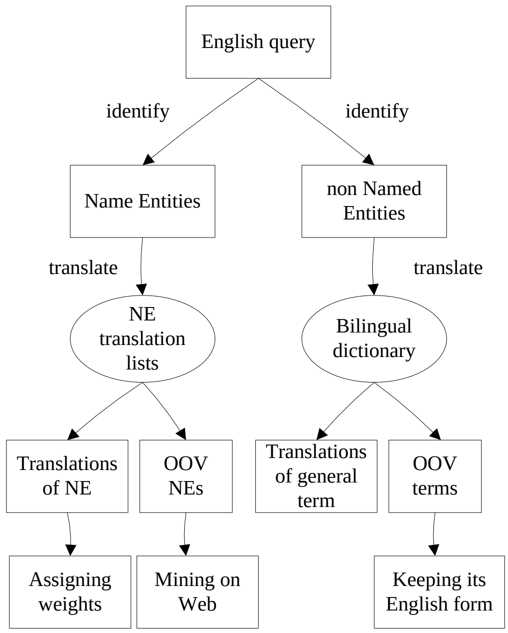


Figure 1. Process of Applying NE Translations to CL information exploration

### 4.1 Differentiating Weights for NE Translations

Due to different sources, usages, and preferences, the same NE may have different translations. Some of them are all correct but with different transliteration schemes. For example, the person name “Albright” can have Chinese translation "奥尔布赖特" or "阿尔布赖特" according to different scheme. The organization name "UNESCO" has its Chinese translation "联合国教科文组织" or just "教科文组织" according to its different abbreviation. For these types of translation alternatives, it is actually more important that possible translations are all there. It is less important to have a weight to differentiate their importance. However, the weights are essential to handle the following situation. Some translations are errors generated through the automatic IE process. If without any further indication, the retrieval system would treat them equally important to the correct translations, which would reduce the retrieval effectiveness. To solve this problem, we utilized a popular commercial Web search engine[2] to obtain the number of Web sites containing the NEs, the translation, and the NE and the translation together in the same page:

$$weight(NE_i, tran_{i,j}) = \frac{|NE_i \cap tran_{i,j}|}{|NE_i| + |tran_{i,j}|} \quad (1)$$

where weight($NE_i$,$tran_{i,j}$) is the weight for translation j of $NE_i$; |$NE_i$,$tran_{i,j}$| is the number of returned pages containing both $NE_i$ and $tran_{i,j}$; |$NE_i$| is the number of returned pages containing $NE_i$; and |$tran_{i,j}$| is the number of returned pages containing $tran_{i,j}$.

After getting all the weights for each $NE_i$, we normalized the sum of weight($NE_i$,$tran_{i,j}$) to 1:

$$\sum_{j=1}^{n} weight'(NE_i, tran_{i,j}) \quad (2)$$

where n is the total number of translations of $NE_i$.

## 4.2 Exploiting OOV NEs from Web

Same as there are OOV terms when using normal dictionary for query translations, there are OOV NEs when using our IE generated NE resources. To fix this problem, we developed two simple patterns that look for the English NEs either appear in brackets after a Chinese word (i.e., … Chinese word (English NE)…) or a Chinese word in brackets after the English NE (i.e., … English NE (Chinese Word)…). The motivation is that many Chinese Web pages would give the English or Chinese translation when a new NE is introduced. Our search was performed on the most popular Chinese search engine[3].

# 5. Experiment Design and Result Analysis

The goal of our experiments is to examine the effectiveness of applying the NE translations obtained by using IE techniques into event triggered information exploration. In particular, we identified three research questions that are corresponding to the two research objectives stated in Section 1:

(1) Are the NEs distributions inside the search topics of TDT collections different to that in CLEF collections?

(2) What is quality (measured by precision and recall) of the IE component in identifying named entities?

(3) What are the effects of applying the NEs and their translations into event-triggered information exploration?

To answer these research questions, we conducted a set of collection analysis and English-Chinese CLIR experiments.

## 5.1 Experiment Setup

Our CLIR experiments were performed on Chinese documents using English queries. For testing event-driven information exploration, we chose TDT4 and TDT5 Multilingual News Text corpora issued by Linguistic Data Consortium (LDC) as the document collections. The collections were developed for Topic Detection and Tracking (TDT) tasks which are triggered by seminal events. TDT corpora have multi-lingual data which are English, Chinese and Arabic, of which we used Chinese documents and English topics.

The Chinese news articles were collected daily in two time periods: Oct. 2000-Jan. 2001 and Apr.-Sept. 2003. In total there are 83,627 Chinese documents with the size of 328 MB, which contain 23,626,923 Chinese characters and 412,942 unique Chinese characters. All documents were converted into UTF-8 encoding. We used the Stanford Chinese word segmenter[4] to segment Chinese documents, and the Porter stemmer[5] for stemming English queries and the dictionary when necessary. Stop words were removed using a Chinese stopword list[6] and an English stopword list[7].

During query translation, when NE tagger and NE translations are used, the translation process is divided into two steps. First NEs are identified using the NE tagger, then the NEs are searched inside the NE dictionary for their translations. Found translations are added into the translated queries, whereas those NEs that can not find translations are combined with the rest of non-NEs query terms to go through the translation process of using normal dictionary. Because the translation of NEs can still be too long, we apply Chinese segmenter on them just in case the segmenter would break the same phrases in the documents.

The normal dictionary used to translate non-NEs query terms was compiled by training GIZA++[8] on multiple sources including the Foreign Broadcast Information Service (FBIS) corpus, HK News and HK Law, UN corpus, and Sinorama, et al. The dictionary contains 126,320 English entries, which are all words, not including phrases [17].

During translating those non-NE terms using the normal dictionary, to remove low probability translations which often are noises, a fixed threshold called

2 http://www.google.com/
3 http://www.baidu.com/

4 http://nlp.stanford.edu/software/segmenter.shtml/
5 http://tartarus.org/~martin/PorterStemmer/
6 http://www.unine.ch/info/clef/englishST.txt/
7 http://www.langtech.org.cn/html/bbs.html/
8 http://www.fjoch.com/GIZA++.html/

Cumulative Probability Threshold (CPT) is selected. This is done by ranking the translations in decreasing order of their normalized probabilities, then iteratively selecting translations top-down until the cumulative probability of the selected translations is first reached or exceeds the threshold. A threshold of 0 thus corresponds to using the single most probable translation (a well-studied baseline) and a threshold of 1 corresponds to use of all translation alternatives.

Either for NEs or non-NE terms, in order to reduce out-of-vocabulary problem, we adopted the back-off translation strategy [3] that is to match the surface form of an input English term to surface forms of English term in the dictionary and the NE translation lists first, if it fails, stem the input English term and match the stem of the input term to surface forms of English term in the dictionary and the NE translation lists, if this still fails, stem the dictionary and the NE translation lists and match the surface form of the input term to stems of terms in the dictionary and the NE translation lists, if all else fails, match the stem of the input term to the stem of terms in the dictionary and the NE translation lists.

Once the queries were translated, they were sent to Indri v.2.4[9] to perform the document retrieval.

The measures used for testing the results are presented in the corresponding sections, and the statistical significance test used was two tailed paired samples t-test.

### 5.2 Named Entities in the Topic Set

For evaluating another research topic [18], we developed a test collection for event-triggered information exploration. The documents are all Chinese articles from TDT4 & 5 collections. The search requests are selected 44 English TDT topics, each of which has at least more than 20 Chinese relevant documents. We also translated the English topics into Chinese with the help of Chinese native speakers who have good English skills. This helps us to obtain Chinese monolingual results which can be useful in experiments.

Inside both the original TDT topics and our modifications of TDT topics to TREC style (see Figure 2), clearly there are many NEs in various types. For example, there are people names “Laurent Gbagbo”, “Alassane Quattara”, location “Ivory Coast”, and time “October 25, 2000”. To reveal how different these TDT topics are to the generic CLIR topics, we ran a study examining the distributions of NEs.

Based on each modified TDT topic in TREC style, we generated three types of queries: short length queries that contain titles only (T query), medium length queries with title and description (TD query), and long length queries with all the three fields (TDN query). The average lengths of queries were: T queries (4 terms), TD queries (27 terms), and TDN queries (127 terms). This gives us three different query scenarios in our experiments.

<num> Number: 41012
<E-title> Trouble in the Ivory Coast

<E-desc> Description:
Presidential election; Laurent Gbagbo, Alassane Ouattara, Ivory Coast voters; Ivory Coast; October 25, 2000

<E-narr> Narrative:
On October 25, Laurent Gbagbo, head of the Ivorian Popular Front, declared himself president, as early polls showed him in the lead. Alassane Ouattara called the election unfair, but then conceded, though tens of thousands of his supporters took to the streets. A recent history of power struggle that led to the current election. Disputes concerning the election including violence by the opposition groups.

Figure 2. An Example of Modified TDT Topics

When studying the distributions of NEs search topics, we divided the TDT topics into three classes based on the query length. With the increasing information in the topic statement, there are more NEs (see Table 1).

Table 1: NE distributions in 44 TDT topics

| **Field** | **Total No. of NE** | **Average No. of NE** | **Stdev. of NE** |
|---|---|---|---|
| Title | 65 | 1.48 | 0.51 |
| Desc. | 154 | 3.5 | 1.70 |
| Narr. | 221 | 5.02 | 2.89 |

Using the data from Mandl et al’s study [15], we can compare the NE distributions inside TDT collections and that of the CLEF collections. Mandl’s paper gives data from 2000 to 2003. To save the space, we only list CLEF 2002 data, which has the highest number of NEs in each topic. As shown in Table 2, our event-driven information exploration topics have much higher numbers of NEs (unique) in each topic.

Table 2: Comparison of NEs in TDT and CLEF Topics

| **Coll** | **No. of topics** | **All No. of NE** | **Ave. No. of NE** | **Topic with no NE** | **Topic with 1 or 2 NEs** | **Topic with over 3 NEs** |
|---|---|---|---|---|---|---|
| TDT 4&5 | 44 | 238 | 5.4 | 0 | 3 | 41 |
| CLEF 2002 | 50 | 86 | 1.7 | 14 | 21 | 15 |

The large number of named entities in the TDT topic confirms our hypothesis that event-driven information exploration topics have more NEs, and

[9] http://sourceforge.net/projects/lemur/

therefore, the handling of the translation of NEs would be more important in event-driven information explorations. Whether this is true, we will test in the experiments below.

### 5.3 Evaluation of Named Entities Extraction

Because our NE translation is based on correctly recognizing NEs in queries so that they can be translated by the special NE lists, it is important to know how well the NE taggers can perform on TDT topics. Therefore, we manually annotated NEs in the topics to be used as the ground truth, and used the precision and recall as the metrics:

$$precision = \frac{NE_m \cap NE_k}{NE_m} \quad (3)$$

$$recall = \frac{NE_m \cap NE_k}{NE_k} \quad (4)$$

where $NE_m$ is the number of NEs that machine identified, and $NE_k$ is the number of NEs in the human generated ground truth.

The evaluation results are shown in Table 3, the accuracy of the NE identification is reasonable high, especially the precision is above 0.8 in all situations. This gives us the confidence of using the NE tagger.

Table 3: The Precision and Recall of Named Entity Identification

| | **Title** | **Description** | **Narrative** |
|---|---|---|---|
| Precision | 0.98 | 0.81 | 0.87 |
| Recall | 0.81 | 0.68 | 0.82 |

### 5.4 Experiments on CLIR with NE Extraction and Translation

To examine whether the application of the NE translations obtained through our IE component would generate positive effect to the event-driven information exploration, we ran a set of monolingual and cross-language retrievals:

*Monolingual Baseline (MONO-BASE)*: a run of retrieving Chinese documents using the manually translated 44 Chinese queries. This baseline is used as the top end of CLIR performance, and it is also used as the relative measure for the performance of the CLIR runs (that is, the performance of CLIR runs are converted to be the percentages of the monolingual performance.)

*CLIR Baseline (CLIR-BASE)*: a run using 44 English queries to retrieve Chinese documents without using the information from our NE component. Instead, the bilingual dictionary presented in Section 5.1 was used to translate the queries word by word. The translation probabilities associated with each translation are used to establish the cut off under different cumulative probability threshold (CPT) from 0.0 to 1.0 with an increment of 0.1 at each time for the translation.

*Higher CLIR Baseline (CLIR-QE)*: a run using 44 English queries to retrieve Chinese documents and performing pre-translation query expansion since pre-translation query expansion has been used as a method for resolving OOV problems. Here, although the translation of OOV terms cannot be solved directly. By performing query expansion on the results from an initial search on a comparable English collection, we hope that some highly content related terms would be added to the query so that the relevant criteria expressed by those OOV terms can be reasonably expressed by the expanded, highly co-occurred terms. For the pre-translation query expansion, we used TDT 4 and 5 English document collection, which in total have 306,498 English documents from the same period as the Chinese document collections. The QE module used was Indir's query expansion mechanism, which is an adaptation of Lavrenko's relevance models [10]. Top scored 20 terms were extracted from top 20 returned documents.

*CLIR run with basic NE Translation (CLIR-NE-1)*: a run using 44 English queries to retrieve Chinese documents using NE resources obtained from IE component.

*CLIR run with weighted NE Translation (CLIR-NE-2)*: a run that is the same as CLIR-NE-1 with one expansion where all NE translations have their corresponding weights obtained using the methods presented in Section 4.1.

*CLIR run with weighted NE Translation plus Web OOV resolution (CLIR-NE-3)*: a run that is the same as CLIR-NE-2 but with the extension of finding the translations for OOV NE on the Web using the methods presented in Section 4.2.

Table 4: Performance of Each Run

| **Run ID** | **T** | | **TD** | | **TDN** | |
|---|---|---|---|---|---|---|
| | **MAP** | **P10** | **MAP** | **P10** | **MAP** | **P10** |
| MONO-BASE | 0.4739 | 0.6273 | 0.5817 | 0.7386 | 0.6215 | 0.7591 |
| CLIR-BASE | 0.3336 | 0.4955 | 0.4251 | 0.5682 | 0.4701 | 0.6364 |
| CLIR-QE | 0.3714 | 0.5045 | 0.4373 | 0.5682 | 0.4477 | 0.6136 |
| CLIR-NE-1 | 0.3931 | 0.5477 | 0.4880 | 0.6682 | 0.5515 | 0.7091 |
| CLIR-NE-2 | 0.3934 | 0.5500 | 0.4887 | 0.6659 | 0.5522 | 0.7114 |
| CLIR-NE-3 | 0.3934 | 0.5500 | 0.5034 | 0.6773 | 0.5563 | 0.7182 |

Table 5: Comparison of Each Run (MAP) (P-value<0.05 is Considered Statistically Significant and * means 0.05<p<0.01, ** means p-value<0.01)

| Run ID | T | | | TD | | | TDN | | |
|---|---|---|---|---|---|---|---|---|---|
| | **% of MONO-BASE** | **Impr. over CLIR-BASE** | **Impr. over CLIR-QE** | **% of MONO-BASE** | **Impr. over CLIR-BASE** | **Impr. over CLIR-QE** | **% of MONO-BASE** | **Impr. over CLIR-BASE** | **Impr. over CLIR-QE** |
| MONO-BASE | 100% | 42.06% | 10.25% | 100% | 36.84% | 33.02% | 100% | 32.21% | 38.82% |
| CLIR-BASE | 70.39% | 0 | ---- | 73.08% | 0 | ---- | 75.2% | 0 | ---- |
| CLIR-QE | 78.37% | 11.33% | 0 | 75.18% | 2.87% | 0 | 72.04% | -4.76% | 0 |
| CLIR-NE-1 | 82.95% | 17.84%** | 5.84% | 83.89% | 14.8%** | 11.59% | 88.74% | 17.32%** | 23.19%** |
| CLIR-NE-2 | 83.01% | 17.93%** | 5.92% | 84.01% | 14.96%** | 11.75% | 88.85% | 17.46%** | 23.34%** |
| CLIR-NE-3 | 83.01% | 17.93%** | 5.92% | 86.54% | 18.42%** | 15.12%* | 89.51% | 18.34%** | 24.26%** |

To evaluate the performance, we adopted mean average precision (MAP) as the primary measure because it is designed to work on ranked lists, and it consider both precision and recall of a search. Besides that, we also look at precision at top 10 documents (P10) to examine how many relevant documents are at the top of the ranked list.

Darwish et al. demonstrated that there is basically no need to fine tune CPT for a CLIR system once you use probabilistic structured query method to take the advantage of the probabilities associated with each translation [3]. Basically CLIR-BASE performed almost indistinguishable between CPT 0.0 to 1.0 on all query length conditions (i.e., T, TD, TDN queries). Since at CPT 0.5, all three conditions have a small peak, and CPT 0.5 give us less noisy translated queries to handle, we fixed the CPT at 0.5 for all other CLIR runs.

Table 4 shows the results of all our six runs. MONO-BASE clearly is the one that performed the best, whereas CLIR-BASE performed the lowest. The performance of pre-translation QE is not as greater as that of CLIR-NE-1, CLIR-NE-2 and CLIR-NE-3.

Table 5 gives more details about the comparisons among the different NE runs against various baselines. Clearly, none of the improvement by using either QE or NE translations helps the CLIR runs to be comparable to the monolingual baseline MONO-BASE. In appearance, this seems to be lower than the state of the art of CLIR performance, since it is common for CLIR results to be close to 100% if not higher than 100% of monolingual baseline. However, as we stated, we are interested in using IE techniques to solve the OOV problems associated with NE translations. Therefore, our concerns are concentrated on whether IE techniques significantly improved over the several CLIR baselines.

Indeed, all NE runs significantly outperformed CLIR-BASE run in all three query length conditions. When comparing to the CLIR-QE run which has some capabilities of handling OOV terms, all three NEs significantly outperformed it under TDN query length condition, and CLIR-NE-3 even significantly higher than it under TD query condition. Therefore, we conclude that our IE techniques can significantly help the performance by better handling NE translations. The fact that CLIR-NE-2 consistently better than CLIR-NE-1, and CLIR-NE-3 consistently better than CLIR-NE-2 demonstrates the usefulness of finding weights for NE translations, and using the Web to resolve OOV NEs.

Because there are less NEs in the title field, the improvement of having NEs translation often is thought to be less effective. However, our results show that all runs performed relatively the same under all three query conditions. This may indicate that it is the relative ratio between NEs and other non-NE terms rather than the number of NEs is the important factor to predicate the impact of solving NEs and even maybe OOV problems.

After query translation in CLIR-NE-2, there were in total 25 OOV NEs among the 44 topics. In CLIR-NE-3, all these 25 OOV NEs were searched on the Web, 14 of them found their translations via this method. It is these 14 translations that made CLIR-NE-3 have further improvement over CLIR-NE-2, so that over all, CLIR-NE-3 is significantly better than CLIR-QE.

## 6. Conclusions

In this paper, through a set of experiments on testing the number of NEs in event-driven information exploration topics, and on examining the effectiveness of using high quality NE translations, we have demonstrated that NEs are important information in CLIR search queries. A better way of handling NEs and their translations would significantly improve the performance. Within the context of event-driven information exploration, we found that NEs are even more common. Therefore, we developed a strategy to integrate IE techniques to obtain NE translations in better quality. The effect is that it not only reduces the number of OOV terms in query translations, but also provides better translations for the NEs. To further improve over the IE techniques, we developed two methods to 1) provide different weights to NE translation alternatives, 2) mine the Web for translations when a NE cannot find its translation in our IE generated NE translation list. All these efforts on handling NEs and their translations have helped the CLIR system to obtain significant improvement not only over a simple dictionary based CLIR baseline, but also over a pre-translation query expansion method. The query translation method can help to reduce the effect of OOV terms and significantly improve over the basic CLIR approach by introducing context related terms.

Our work raises some interesting new questions, perhaps the most important of which is how much integration of IE technique should be to the CLIR process. Our results show that NEs and their translations can help, even though the integration is still loose. What if we tightly integrate IE components with CLIR system so that the identification of translations can be tailored to individual search topics? Another question is the application of our method to a full CLIR system where all techniques for improving CLIR performance have been applied except the handling of OOV. Our system will provide the handling of OOV NEs.

## References

[1] Artiles, J., Gonzalo, J. and Sekine, S. The SemEval-2007 WePS Evaluation: Establishing a benchmark for the Web People Search Task. In the 4th International Workshop on Semantic Evaluations (Semeval-2007) at ACL07, pp. 64-69, 2007.

[2] Bikel, D. M., Miller, S., Schwartz, R. and Weischedel, R. Nymble: a high-performance Learning Name-finder. In Fifth Conference on Applied Natural Language Processing, pp. 194-201, 1997.

[3] Darwish, K. and Oard, D. W. Probabilistic Structured Query Methods. In Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 338-344, 2003.

[4] Demner-Fushman, D. and Oard, D. W. The Effect of Bilingual Term List Size on Dictionary-Based Cross-Language Information Retrieval. In 36th Annual Hawaii International Conference on System Sciences (HICSS'03), pp. 108-118, 2003.

[5] Grishman, R., Westbrook, D. and Meyers, A. NYU's English ACE 2005 System Description. In ACE 2005 Evaluation Workshop. 2005.

[6] Hsin-Hsi, C., Sheng-Jie, H., Yung-Wei, D. and Shih-Chung, T. Proper name translation in cross-language information retrieval. In Proceedings of the 17th international conference on Computational linguistics, pp. 232-236, 1998.

[7] Huang, F., Vogel, S. and Waibel, A. Extracting Named Entity Translingual Equivalence with Limited Resources. ACM Transactions on Asian Language Information Processing (TALIP), 2(2): 124-129. 2003.

[8] Ji, H., Blume, M., Freitag, D., Grishman, R., Khadivi, S. and Zens, R. NYU-Fair Isaac-RWTH Chinese to English Entity Translation 07 System. In NIST ET 2007 PI/Evaluation Workshop. 2007.

[9] Jiang, L., Zhou, M., Chien, L.-F. and Niu, C. Named Entity Translation with Web Mining and Transliteration. In IJCAI'07, pp. 1629-1634, 2007.

[10] Lavrenko, V. and Croft, W. B. Relevance-based Language Models. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 120-127, 2001.

[11] Nasreen, A. and Leah, S. L. Statistical transliteration for English-Arabic cross language information retrieval. In Proceedings of the twelfth international conference on Information and knowledge management, pp. 139-146, 2003.

[12] Oard, D. W. When You Come to a Fork in the Road, Take It! In Proceedings of SIGIR2002 workshop "Cross-Language Information Retrieval: A Research Roadmap". 2002.

[13] Oard, D. W. and Diekema, A. R. Cross-Language Information Retrieval. Annual Review of Information Science and Technology. B. Cronin (Eds). American Society for Information Science. 33 pp. 223-256, 1998.

[14] Pablo-Sanchez, C. d., Martınez-Fernandez, J. L. and Martınez, P. Named Entity Processing for Cross-lingual and Multilingual IR applications. In CLEF 2005. 2005.

[15] Thomas Mandl and Womser-Hacker, C. The Effect of Named Entities on Effectiveness in Cross-Language Information Retrieval Evaluation. In ACM SAC'05, pp. 1059-1064, 2005.

[16] Virga, P. and Khudanpur, S. Transliteration of Proper Names in Cross-Language Applications. In the ACM SIGIR 2003, pp. 365-366, 2003.

[17] Wang, J. and Oard, D. W. Combining Bidirectional Translation and Synonymy for Cross-language Information Retrieval. In Proceedings of the ACM SIGIR 2006, pp. 202-209, 2006.

[18] Wu, D. and He, D. ICE-TEA: an Interactive Cross-language Search Engine with Translation Enhancement. In the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 882, 2008.

[19] Yaser, A.-O. and Kevin, K. Machine transliteration of names in Arabic text. In Proceedings of the ACL-02 workshop on Computational approaches to semitic languages, pp. 31-46, 2002.