

# Using Information Extraction to Improve Cross-lingual Document Retrieval

Dilek Hakkani-Tür

ICSI

Berkeley, CA, 94704, USA

dilek@ICSI.Berkeley.EDU

Heng Ji

Department of Computer Science, New York University

New York, NY, 10003, USA

{hengji, grishman}@cs.nyu.edu

Ralph Grishman

## Abstract

We present a filtering mechanism using two cross-lingual information extraction (CLIE) systems for improving document relevance of cross-lingual information retrieval (CLIR) for queries conforming to predefined templates. Experiments on retrieving Chinese documents in response to English GALE<sup>1</sup> arrest queries show that this approach can obtain a 12.7% absolute improvement in relevance (representing a 24.8% relative error reduction) for the top 25 retrieved documents. We also demonstrate that Chinese IE can provide a valuable supplement to English IE to enhance retrieval performance.

## Keywords

Cross-lingual, Information Extraction, Document Retrieval

## 1. Introduction

A shrinking fraction of the world's Web pages are written in English, and so the ability to access pages across a range of languages is becoming increasingly important for many applications. This need is being addressed in part by the research on Cross-Lingual Information Retrieval (CLIR), which, given an arbitrary query stated in one language, seeks to retrieve relevant documents in one or more foreign languages. CLIR combines two difficult problems, document retrieval and term or text translation, and these have limited the performance of CLIR systems in general.

For some applications, however, we are able to identify particular types of queries which are of primary interest to a community of users. In such circumstances we can optimize the document retrieval function for these queries, while still using general CLIR mechanisms for other queries. In this paper we study the benefits of this optimization.

Specifically, we conducted this study in the context of the "distillation evaluation" of the GALE program. As we will describe in more detail below, the evaluation task involves answering queries with respect to a multilingual collection, where the queries must conform to one of a set of query templates. Some of these templates allow for very general queries, while others involve specific types of relations or events. We examine one of the more specific query templates in detail, focusing on the

problem of retrieving relevant Chinese documents in response to the (English) query. We compare several retrieval strategies, using as a baseline keyword-based retrieval on the English (machine) translations of the documents, and then adding filters based on the output of two information extraction systems, one operating on the Chinese source, the other on the translated documents. We show in particular the benefits (and some of the limitations) of using source-language information extraction for document filtering.

The rest of this paper is structured as follows. Section 2 describes our main research task and experimental setting. Section 3 briefly describes the previous efforts made by researchers of using sophisticated linguistic analysis to enhance IR performance. Section 4 describes the motivation for our approach. Section 5 presents an overview of our system architecture and strategies for using IE to filter out the irrelevant documents. Section 6 presents the experimental results. Section 7 compares our approach with two possible alternative approaches and Section 8 then concludes the paper and sketches our future work.

## 2. Task and Terminology

Our approach has been evaluated in the framework of the U.S. Government's DARPA GALE program. One of the GALE evaluations (the *distillation* task) involves responding to queries based on a set of question templates (17 templates in GALE 2007). For the experiments presented here, we used the *arrest* class of questions (template number 15, according to the GALE program). These take the form: "*Describe arrests of persons from X and give their role in the organization*", in which X is an organization name such as "Peruvian government", "Shining Path", "WorldCom", "US Federal Bureau of Investigation", "Enron Corporation", "Jemaah Islamiyah", "ETA", "al Qaeda" and "the PLF".

We use the University of Massachusetts INDRI IR system<sup>2</sup> as a major component to return the top N ( $N \leq 50$  in this paper) relevant documents in response to a query. We then use a statistical approach to extract sentences from these documents. Our goal in this paper is to improve the precision of identifying Chinese documents relevant to these questions (we don't consider the sentence extraction phase in this paper).

---

<sup>1</sup> Global Autonomous Language Exploitation

---

<sup>2</sup> <http://www.lemurproject.org/indri/>

We make use of two cross-lingual IE systems developed around the ACE<sup>3</sup> evaluations to reach these goals. ACE defines 8 types of events, with 33 subtypes, including *Arrest-Jail* events which cover arrests, capture events, extraditions, jailing events, incarceration, etc.

### 3. Related Work

The attempt to marry natural language processing (NLP) techniques with large-scale IR is not new, but effective integration of the two remains an open research question. Researchers have experimented with syntactically derived word pairs (Strzalkowski et al., 1996; Zhai et al., 1996; Arampatzis et al., 1998), case frames (Croft and Lewis, 1987), paraphrases (Duclaye and Yvon, 2003), part-of-speech tagging (Eichmann 2003), name tagging (Eichmann 2003; Harabagiu et al., 2005; Katz and Lin, 2003), reference resolution (Harabagiu et al., 2005), parsing (Smeaton et al., 1994; Harabagiu et al., 2005) and syntactic relation patterns (Shen et al., 2005) as units of indexing. However, none of these experiments resulted in a dramatic improvement in precision or recall, and sometimes even resulted in degraded performance. Note however these efforts aim to handle arbitrary retrieval queries, whereas we can take advantage of specific types of query templates. In that regard, our work is more similar to the document filtering experiments which are sometimes used to assess information extraction (Yangarber et al., 2000).

Our spirit of using IE results as a post-processor for IR is closest to (Schiffman et al., 2007). But we have a different focus, on the problem of CLIR, whereas they emphasized the extraction of sentences from English texts only. They use IR at the first stage to return a small number of relevant documents; IE results are then used to select highly relevant words to revise the query for a second retrieval pass. But in our cross-lingual environment, the candidate documents returned by IR engine are much more noisy, therefore we use IR in high recall mode to return a large collection, then use extracted events to filter the irrelevant documents to improve precision. Our work can be considered as an application of the filtering approach in (Hakkani-Tür et al., 2007) in a cross-lingual environment.

### 4. Motivation

Despite the intuition that linguistically sophisticated techniques should be beneficial to IR, real gains in performance have yet to be demonstrated empirically in a reliable manner. We believe that the key to effective application of NLP technology is to selectively employ it in situations where we can expect to improve performance, without abandoning simple techniques. We felt that these template-based queries offer such an opportunity.

The setting of template-based CLIR has encouraged the development of CLIE systems, but it also raises the issue of the best approach for CLIE performance. We can

employ the following two cross-lingual IE (CLIE) pipelines to process Chinese documents:

**MT\_English IE:** Translate Chinese texts into English, and then run English IE on the translated texts.

**Chinese IE\_MT:** Run Chinese IE on the Chinese texts, and then use MT word alignments to translate (project) extracted information into English.

In the mono-lingual environment English IE systems generally perform better than Chinese IE, so it's natural to apply English IE. However, Chinese is linguistically very different from English and statistical MT performance for Chinese-English is still quite poor. Therefore the process of applying English IE on MT output becomes much more lossy than on English documents.

In order to quantify the information lost by MT, we count the number of "arrest" event trigger words (A trigger is the word that most clearly expresses the event occurrence) which have associated arguments, detected from 19 ACE Chinese texts, and their overlap with the true events. The results are shown in Figure 1.

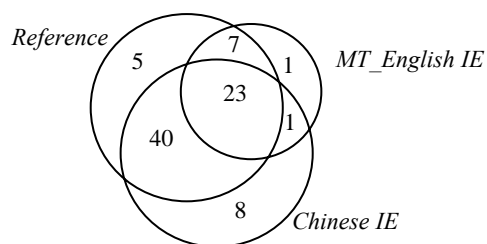


Figure 1. Number of event triggers with arguments

We can see that applying English IE on MT provides high precision (only 2 of 32 extracted events are not in reference), but only covers a small fraction (30/74) of the key events. On the other hand, by bypassing MT, Chinese IE can be a useful supplemental procedure, recovering 40 correct events missed by English IE, although it has lower accuracy (9 wrong events were generated at the same time).

In the following we use two translated example texts, to give more intuition about how these relevant events were missed by MT.

#### Example 1:

**[Query]** Describe arrests of persons from [al Qaeda] and give their role in the organization.

**[Chinese Text]** [迪拜]昨天阿拉伯新闻频道报道说盖达组织第二首脑, 欧萨玛的得力助手扎瓦里已经在伊朗落网。

**[Text Machine Translation]** DUBAI Arab news channel, reported yesterday that the terrorist organization Al-Qaida

<sup>3</sup> <http://www.itl.nist.gov/iad/894.01/tests/ace/>

second, Osama's right-hand man, Abdurrahman Wahid in Iran.

[Text Reference Translation] (DUBAI) Arab News Channel reported that Zawahiri, the second most important person of Al-Qaeda and Osama's right-hand man, was arrested in Iran.

### Example 2:

[Query] Describe arrests of persons from [the PLF] and give their role in the organization.

[Chinese Text] 据报道,“巴勒斯坦解放阵线”领导人阿巴斯在巴格达南郊的住所内被美军逮捕。

[Text Machine Translation] According to reports, PLF leader Abbas in Baghdad outside the residence of the US military.

[Text Reference Translation] According to reports, the leader of the PLF, Abbas was arrested by the US military in his residence in the outskirts of Baghdad.

Although the organization names “al Qaeda” and “PLF” in these queries are correctly translated, the event trigger words representing “arrest” are missing. In example 1, the “arrest” trigger “落网 (fall into meshwork)” is fairly rare and metaphorical; in example 2, MT met difficulty probably because of the re-ordering of phrases. In these cases, applying IE directly on source (Chinese) texts could help.

To sum up, combining both CLIE pipelines could allow us to incorporate information from a much wider knowledge base, spanning both the original and the translated documents. In the following section we will describe the algorithms to capture these intuitions.

## 5. System Architecture

### 5.1 System Overview

The overall system pipeline is presented in Figure 2. We split document retrieval into a two-stage process. The first stage simply applies cross-lingual IR, without any IE knowledge, and initially retrieves the top N ( $N \leq 50$ ) documents for each query. Then we use the events detected from CLIE as additional constraints to determine whether a document is relevant or not. In the following we shall present the system components and detailed filtering algorithm.

### 5.2 Machine Translation

Both CLIE systems need machine translation to translate Chinese documents (for English IE) or project the extraction results from Chinese IE into English. We use the RWTH Aachen Chinese-to-English machine translation system (Zens and Ney, 2004) for these purposes. It's a statistical, phrase-based machine translation system which memorizes all phrasal translations that have been observed in the training corpus. It computes the best translation using a weighted log-linear combination of various statistical models: an n-gram language model, a phrase translation model and a word-based lexicon model. The latter two models are

used in source-to-target and target-to-source directions. Additionally, it uses a word penalty and a phrase penalty.

The model scaling factors are optimized on the development corpus with respect to the BLEU score similar to (Och 2003). Almost all bilingual corpora provided by LDC were used for training, which account for about 200 million running words in each language. Language modeling used the English part of the bilingual training corpus and in addition some parts of the English GigaWord corpus. The total language model training data consists of about 600 million running words.

This MT system produces a translation for each source document, and also the *word-to-word* mapping derived from phrase-based alignment.

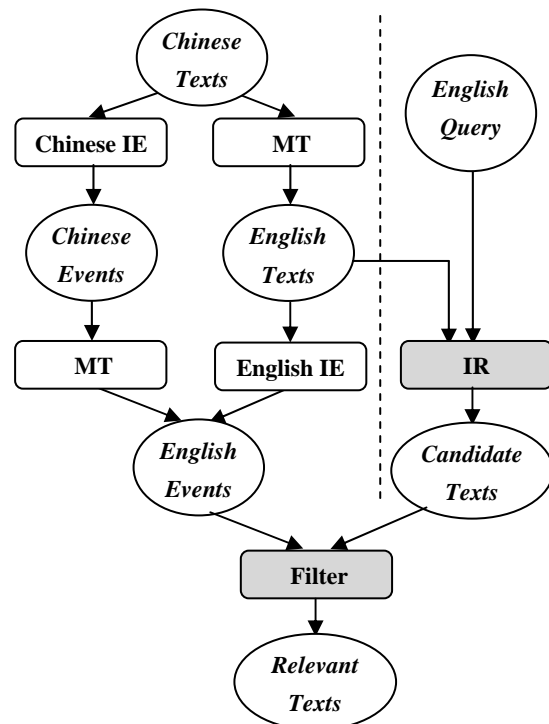


Figure 2. System Architecture

### 5.3 Baseline Cross-lingual IR

For an English query, we use the INDRI retrieval system (Strohman et al., 2005) to identify the top N translated documents, ranked by IR score. INDRI combines language modeling and inference network approaches in an architecture designed for large-scale applications. INDRI expands the query by keyword parsing, operator conversion, and date/numeric word processing. Like most other IR systems, INDRI is task independent and does not perform any deep analysis that particularly addresses the event information. However, since the query templates are known in advance, it has been possible to hand-craft queries for each template, incorporating additional keywords into the query (for example, adding “apprehend” and other related terms to the arrest query).

## 5.4 IE

We present the training and test procedures for the two IE systems as follows. Both of them are trigger-based and use pattern matching.

The English IE system combines pattern matching with statistical models (Grishman et al., 2005). For every instance of an event in the ACE training corpus, we construct two types of patterns representing the connection between the trigger word and the event arguments, and record the type and subtype of the associated event. One pattern is the sequence of constituent heads separating the trigger and arguments. The other pattern is the predicate-argument structure (Meyers et al., 2001) connecting the trigger to all the event arguments. For each argument, we record its ACE type and subtype and its head. In addition, we train a set of MaxEnt classifiers to distinguish events from non-events, to classify events by type and subtype, to distinguish arguments from non-arguments, and to classify arguments by argument role. In the test procedure, each document is scanned for instances of triggers from the training corpus. When an instance is found, we first try to match the environment of the trigger against the set of patterns associated with that trigger. This pattern-matching process, if successful, will assign some of the phrases in the sentence as arguments of a potential event. The argument classifier is applied to the remaining roles in the sentence; for any passing that classifier, we use the role classifier to assign a role to the mention. Finally, once all arguments have been assigned, we apply the event classifier to the potential event; if the result is successful, we report this as an event instance.

The Chinese IE system is based on patterns semi-automatically extracted from the ACE training corpus. For each event instance we replace the trigger word by its event type and subtype, and each argument by its entity type. Then we apply the Purdue University POS tagger (Huang et al., 2007) and chunker (Harper et al., 2005) on each event training instance. Then we edit the patterns by hand, replacing tokens by their POS tag, chunk type, or a wild card or deleting them entirely if they are not relevant to detecting the event type. Some patterns are collapsed, and some patterns which appear too specific or too general are deleted. To ensure that patterns are not over-generalized by the hand editing, the training corpus is split in two and patterns derived from one half are, after hand editing, applied to the other half to review their accuracy in event prediction. In the test procedure, each document is annotated with POS tagging and chunking, and then scanned against the patterns derived from the training corpus. Unlike the English IE system, currently we don't have statistical models following pattern matching.

## 5.5 Filtering Approach

In the GALE task we are given templates for queries in advance. When a template-based query is submitted to this IR engine, one challenge is determining the number of relevant documents that can be used for template-

based question answering from the information retrieval output. This is problematic since it is hard to know the optimal value that holds for all queries. Sometimes a query has only one relevant document in the huge document repository and sometimes thousands. If the sentence extraction system processes a larger number of returned documents, this is expected to result in a higher number of false alarms unless document level processing is available.

One solution might be getting fewer documents from IR but this may result in poor recall. Alternatively one could exploit document and argument scores returned by INDRI. Here, the argument score reflects matches between the document and the value of the slots in the queries. However the document and argument scores usually have different dynamic ranges depending on the query and it is not easy to perform thresholding that works optimally for all queries using them.

Therefore, we use an intermediate processing stage between the IR engine and the sentence extraction module, to filter out irrelevant documents. The basic idea is as follows: Since the distillation query templates are known beforehand, it is sometimes possible to associate expected document contents with one or several types of ACE event annotations. For example for:

**Query template 15:** *Describe arrests of persons from [organization] and give their role in the organization.*

the relevant document must have the ACE event of subtype *Arrest-Jail*. Since the CLIE systems provide such annotations from both source language and translated documents, the post-processing stage needs to check only whether the event mentioned in the query appears in the documents returned by IR. A more extensive version can also require that the organization name specified in the query be present in the returned document. However, in this work, we only considered filtering according to event types and subtypes.

## 6. Experimental Results

In this section we shall present the results of applying this method to improve GALE cross-lingual document retrieval.

### 6.1 Data

We evaluated our approach by using 12 example queries of GALE template 15 to retrieve documents from the English machine translation of the TDT5 Chinese corpus consisting of 56,485 newswire texts from four different news agencies. For each query the baseline IR system returns the top N ( $N \leq 50$ ) documents, in total 421 documents.

Then these documents were manually labeled as relevant or not-relevant by one of the authors to construct the reference set for the evaluation. The decisions were made against the original Chinese documents, using the

rules in the GALE annotation guidelines<sup>4</sup>. For example, a document describing “the history of the Muslim Brotherhood” (e.g. “The Muslim Brotherhood is viewed as the second strongest political force in Egypt”) is not relevant to the query for X= “the Muslim Brotherhood”. We did not judge against MT output because when the translation quality is poor that procedure tends to be too subjective and MT system-specific. In total 130 documents were judged as relevant.

## 6.2 Evaluation Metric

### 6.2.1 Precision, Relative-Recall and F-Measure

We use the traditional IR metrics, *precision* and *relative-recall*, to evaluate our approach. As we have noted, the INDRI engine was set to return a maximum of 50 documents per query, and alternative results were obtained by filtering the set returned by INDRI. Consequently, for evaluation purposes our ‘relevant document set’ **R** was the subset of these 50 (or fewer) documents per query judged to be relevant. The baseline system therefore had a maximum relative recall of 100%.

If the system (with filtering) returns document set **D**, then precision and relative-recall can be defined as follows:

$$Precision = \frac{\#(D \cap R)}{\#D}$$

$$Relative\_Recall = \frac{\#(D \cap R)}{\#R}$$

F-measure combines these two metrics:

$$F-Measure = \frac{2 * Precision * Relative\_Recall}{Precision + Relative\_Recall}$$

### 6.2.2 Search Length i

We also use the Search Length i Measure (Cooper, 1968) to measure user effort in terms of the number of non-relevant documents that a user must examine before finding i relevant documents. In our study we set i = 5.

## 6.3 Overall Performance

Figure 3 presents the overall precision and relative recall for the baseline and after adding cross-lingual IE to filter the irrelevant documents. The numbers on the curve show the F-measure (%) scores when N (the number of documents returned by INDRI) = 10, 20, 30, 40 and 50.

We can see that using English IE on MT output does indeed help achieve significant improvements in precision, but also relatively large loss in recall. Then by adding Chinese IE at the end, the system dramatically boosts recall, with a small loss in precision. As N

increases, the overall improvement in F-measure increases from 4.8% to 18.8%.

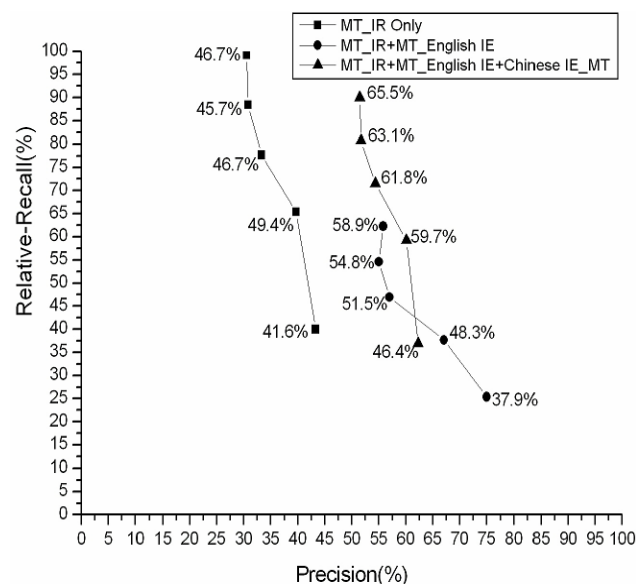


Figure 3. Overall Precision/Relative-Recall

## 6.4 Performance Breakdown

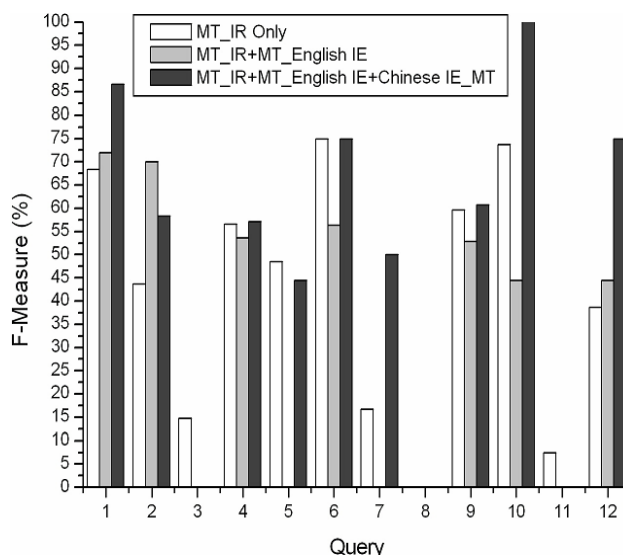


Figure 4. F-Measure for Each Query When N=25

We then performed experiments to evaluate the impact of IE on each query, with a fixed N=25<sup>5</sup>. Figure 4 shows the results. Overall the F-measure increases from 48.8% to 50.5% by using English IE on MT, and increases to 61.5% by further adding Chinese IE.

Except for the small losses for queries 3, 5 and 11, IE produces clear improvements for all the other nine queries, with 1.5%-36.3% gain in F-measure. By adding Chinese IE, the F-measures for all the queries except for

<sup>4</sup><http://projects.ldc.upenn.edu/gale/Distillation/DistillationTrainingDataGuidelinesV2.3.pdf>

<sup>5</sup> We assume N=25 reflects the number of texts a user would look at for a document retrieval task.

query 2 are significantly boosted. In the contrast, the performance by adding only English IE is much less consistent.

In order to check how robust our approach is, we conducted the Wilcoxon Matched-Pairs Signed-Ranks Test on relevance F-measures for these 12 queries, for all the retrieved documents ( $N \leq 50$ ). The results show that we can reject the hypothesis that the improvements using IE were random at a 94.6% confidence level, and reject the hypothesis that the improvements by adding Chinese IE were random at a 94.5% confidence level.

## 6.5 User Effort Measure

For the 8 queries which retrieved at least 5 relevant documents, we measure their search length with  $i=5$  as defined in section 6.2.2. In general, our approach using IE knowledge significantly outperforms the baseline, reducing the mean of search length from 1.625 to 0.375.

## 6.6 Error Analysis

Getting a reasonably complete set of patterns (linguistic expressions) for an event type is inherently difficult. Gaps in the pattern sets used by IE lead to the filtering out of relevant documents:

### Example 3:

[Query] Describe arrests of persons from [Hezbollah] and give their role in the organization.

[Chinese Text] 2000年10月以来,真主党已有4名成员被以军扣押。

[Text Reference Translation]: Since October of 2000, four Hezbollah members have been (detained and seized) by the Israeli army.

The arrest event was missed because the trigger compound word “扣押 (detain and seize)” doesn’t appear in the ACE training data.

Other errors reflect the drawbacks of not using richer CLIE outputs:

### Example 4:

[Query] Describe arrests of persons from [Peruvian government] and give their role in the organization

[Chinese text]玻利维亚警方30日宣布在该国中部缴获186.3公斤纯可卡因,并逮捕5名贩毒嫌疑人。据警员路易斯介绍,5名贩毒嫌疑人经常在玻利维亚收集毒品,并定期运往秘鲁,然后再贩卖给美国和欧洲。

[Text Reference Translation] On 30th Bolivian Police announced that 186.3 kilogram of pure cocaine were captured in the middle of the country, and five suspects of selling drugs were arrested. According to the policeman Louis, these suspects often collected drugs in Bolivia, and regularly shipped to Peru, and then sold to America and Europe..

The document is not relevant to the query because “Peru” is not an event argument of “arrest”. But our filter mistakenly confirmed it as a relevant text without using event argument information. Taking advantage of this information (which is generated by IE) could filter out some irrelevant documents, but (due to errors in IE

argument output) could also block some relevant documents.

## 7. Discussion

### 7.1 Comparison with Query Expansion

One alternative solution to reduce query/document mismatch is to add a more extensive list of “arrest” trigger words to the query. But when such trigger words are used without context, they may be highly ambiguous, leading to many false hits. For example, “pursue” can mean “to follow in an effort to overtake or capture” to indicate an arrest event: “US-Britain special troop pursued Bin Laden”; but it also can mean “to strive to gain or accomplish” such as “pursue lofty political goals”.

Furthermore, the Chinese ACE 2005 training corpus includes 43 different “arrest” trigger words, some of which appear in different forms, noun and verb, sometimes singular and plural, yielding many different trigger translations for a given event type (e.g. armies attack, bombs explode...can all indicate ‘attack’).

### 7.2 Comparison with Query Translation

The work described here complements the cross-lingual question answering (CLQA) research such as (Mitamura et al., 2006). They presented an English-to-Chinese QA system that translates the query from English to Chinese and then searches the translated query among Chinese documents. Combining this approach into our framework can magnify the gains possible with IE for cross-lingual IR.

However, as Mitamura et al. mentioned, the English to Chinese translation accuracy is low because of word sense ambiguities as well as regional language differences for Chinese. In our task the main difficulty of applying this approach is to properly translate English names into Chinese. In particular, many names in our task are written in abbreviations, such as “ETA” and “the PLF”, which will be difficult to translate/expand into Chinese. The regional language difference problem also commonly exists in English-to-Chinese name translation. For example, “al Qaeda” is translated into “基地组织” (based on meaning) in mainland China but “阿尔盖达” (based on its pronunciation) in Taiwan, and “卡伊达” in Singapore (based on part of its pronunciation). This could immediately lead to failure in document retrieval using a query-translation approach.

## 8. Conclusion and Future Work

We identified the linguistic phenomena that cross-lingual IR has difficulty with, even with predefined query templates, and demonstrated that IE enables us to successfully handle these difficulties by effectively exploiting events that can be reliably extracted from texts and matching queries with documents at the event level, thereby significantly improving precision of document retrieval with little loss in recall.

The experiments suggest that some further gains can be achieved through employing richer IE results such as

event arguments, combining this with additional shallow linguistic features such as the positions of events, titles, document structure, document topic and name concurrence. We are also interested in applying the filtering approach to the snippet (sentence and sub-sentence) retrieval system described in (Hakkani-Tür and Tur, 2007).

## 9. Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency under Contract No. HR0011-06-C-0023, and the National Science Foundation under Grant IIS-00325657. Any opinions, findings and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the U. S. Government. The authors would like to thank Gokhan Tür, Mary Harper and Satoshi Sekine for their valuable help and comments on this work.

## 10. References

- [1] A. Arampatzis, Th.P. van der Weide, C.H.A. Koster, and P. van Bommel. 1998. Phrase-based Information Retrieval. *Information Processing and Management*. 34(6):693-707, December
- [2] W. S. Cooper. 1968. Expected search length: A Single Measure of Retrieval Effectiveness based on the Weak Ordering Action of retrieval systems. *Journal of American Society of Information Science*, 19(1), 30-41.
- [3] B. Croft and D. Lewis. 1987. An Approach to Natural Language Processing for Document Retrieval. In *ACM SIGIR 1987*.
- [4] Florence Duclaye and Francois Yvon. 2003. Learning Paraphrases to Improve a Question-Answering System. In *ACE2003 workshop on natural language processing for question answering*.
- [5] David Eichmann. 2003. Issues in Extraction and Categorization for Question Answering. In *AAAI Spring Symposium on New Directions in Question Answering*, Stanford, CA, US.
- [6] Ralph Grishman, David Westbrook and Adam Meyers. 2005. NYU's English ACE 2005 System Description. In *ACE 2005 Evaluation Workshop*. Washington, US.
- [7] Sanda Harabagiu, Dan Moldovan, Christine Clark, Mitchell Bowden, Andrew Hickl and Patrick Wang. 2005. Employing Two Question Answering Systems in TREC-2005. In *TREC 2005*.
- [8] Mary Harper, Bonnie Dorr, John Hale, Brian Roark, Izhak Shafran, Matthew Lease, Yang Liu, Matthew Snover, Lisa Yung, Anna Krasnyanskaya and Robin Stewart. 2005. *Parsing and Spoken Structural Event Detection*. Technical Report, The John-Hopkins University, 2005 Summer Research Workshop.
- [9] Zhongqiang Huang, Mary Harper and Wen Wang. 2007. Mandarin Part-Of-Speech Tagging and Discriminative Reranking. In *EMNLP 2007*. Prague, Czech Republic.
- [10] Boris Katz and Jimmy Lin. 2003. Selectively Using Relations to Improve Precision in Question Answering. In *EACL 2003 Workshop on Natural Language Processing for Question Answering*.
- [11] Adam Meyers, Michiko Kosaka, Satoshi Sekine, Ralph Grishman and Shubin.Zhao. 2001. Parsing and GLARFing. In *Proceedings of RANLP-2001*.
- [12] Teruko Mitamura, Mengqiu Wang, Hideki Shima and Frank Lin. 2006. Keyword Translation Accuracy and Cross-lingual Question Answering in Chinese and Japanese. In *EACL 2006 Workshop on Multilingual Question Answering*.
- [13] F. J. Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *ACL 2003*.
- [14] Barry Schiffman, Kathleen R. McKeown, Ralph Grishman and James Allan. 2007. Question Answering using Integrated Information Retrieval and Information Extraction. In *HLT-NAACL 2007*. Rochester, US.
- [15] Dan Shen, Geert-Jan M. Kruijff, Dietrich Klakow. 2005. Exploring Syntactic Relation Patterns for Question Answering. In *IJCNLP 2005*.
- [16] A. Smeaton, R. O'Donnell, and F. Kellely. 1994. Indexing Structures Derived from Syntax in TREC-3: System description. In *TREC-3*.
- [17] T. Strohman, D.Metzler, H. Turtle and W.B. Croft. 2005. Indri: A Language-model based Search Engine for Complex Queries (extended version). Technical Report IR-407, CIIR, Umass Amherst.
- [18] T. Strzalkowski, L. Guthrie, J. Karlgren, J. Leistensnider, F. Lin, J. Perez-Carballo, T.Straszheim, J. Wang and J. Wilding. 1996. Natural Language information retrieval: TREC-5 report. In *TREC-5*.
- [19] Dilek Hakkani-Tür and Gokhan Tur. 2007. Statistical Sentence Extraction for Information Distillation. In *ICASSP-2007*.
- [20] Dilek Hakkani-Tür, Gokhan Tür and Michael Levit. 2007. Exploiting Information Extraction Annotations for Document Retrieval in Distillation Tasks. In *Interspeech 2007*.
- [21] Roman Yangarber, Ralph Grishman, Pasi Tapanainen and Silja Huttunen. 2000. Automatic Acquisition of Domain Knowledge for Information Extraction. In *COLING 2000*. Saarbruecken, Germany.
- [22] Richard Zens and Hermann Ney. 2004. Improvements in Phrase-Based Statistical Machine Translation. In *HLT/NAACL 2004*. New York City, NY, US
- [23] C. Zhai, X. Tong, N. Milic-Frayling, and D. Evans. 1996. Evaluation of Syntactic Phrase Indexing – CLARIT NLP track report. In *TREC-5*.