A Study of Using an Out-Of-Box Commercial MT System for Query Translation in CLIR

Dan Wu

School of Information Management
Wuhan University, Hubei, China

woodan@whu.edu.cn

Daqing He

School of Information Sciences
University of Pittsburgh, PA, USA

dah44@pitt.edu

Heng Ji, Ralph Grishman

Computer Science Department
New York University, NY, USA

{hengji, grishman}@cs.nyu.ed

ABSTRACT

Recent availability of commercial online machine translation (MT) systems makes it possible for layman Web users to utilize the MT capability for cross-language information retrieval (CLIR). To study the effectiveness of using MT for query translation, we conducted a set of experiments using Google Translate, an online MT system provided by Google, for translating queries in CLIR. The experiments show that MT is an excellent tool for the query translation task, and with the help of relevance feedback, it can achieve significant improvement over the monolingual baseline. The MT based query translation not only works for long queries, but is also effective for the short Web queries.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Query formulation

General Terms

Experimentation, Languages, Performance

Keywords

Cross-Language Information Retrieval, Query Translation, Machine Translation, Query Expansion, Relevance Feedback

1. MT FOR QUERY TRANSLATION IN CLIR

Recent rapid development of new Internet technologies has empowered more people to use the Web as their social and collaboration platform. As every Web user potentially becomes not only a Web information consumer but also an information creator, cross-language information retrieval (CLIR) becomes critical for the communication on the Web. Considering most Web users are non-native English speakers, and most Web information is in English, CLIR is a critical Web search service for non-English searchers.

CLIR often relies on translation methods to cross the language barriers between a query and the documents. Depends on whether it is the query, the documents, or both that are translated, we have query translation based CLIR, document translation based CLIR and interlingual CLIR. Many resources have been exploited for the translation task, among which the most commonly used are machine-readable dictionary (MRD), parallel or comparable corpora, and machine translation (MT) systems. MRD is probably the most commonly used translation resources in experiment setting, especially for translating queries [3, 6]. However, if consider the easy accessibility for some popular languages, MT is probably the most commonly available translation resource among the above three on the Web. Several companies such as Google and Alta Vista are actively prompting their multilingual services on the Web. Therefore, it is reasonable to assume that a layman Web user can potentially achieve a cross-language information retrieval by using one of the commercial online MT services to translate his/her queries.

Although MT in document translation based CLIR can generate competitive search results to monolingual searches, in the context of query translation based CLIR, the effectiveness of using MT to translate queries is much more uncertain in previous studies. On the one hand, researchers criticized that the quality of MT translation was poor [7], and dictionary-based CLIR techniques are often cited to outperform those using some popular commercial MT for translating queries [4]. On the other hand, Oard find that early rule-based MT systems when were used for translating queries can generate comparable or slightly better CLIR results than dictionary-based methods [10]. Other researchers also showed that MT-based query translation can achieve reasonable effectiveness, and was better than the approach that using all the translations in the dictionary [1].

However, we have seen two important developments in the fields of CLIR and MT in recent years. First, both MT and CLIR have experienced rapid development of integrating statistical based language models and resources into their handling of translations. Statistical MT has becomes the state of the art for MT [11], and even some commercial MT systems such as Google Translate are statistical MT systems. Translation probabilities are widely used in CLIR for handling translation ambiguities or are even built as a part of the statistical language modeling for CLIR [6, 9, 14]. One important insight gained in CLIR from the usage of translation probabilities is that choosing multiple possible translations with their probabilities is a superior method than choosing only the top best translation. This insight argues against of the usage of MT output, which is the list of one best translation for each query term, as the translation of the query.

Second, both MT and CLIR have to face out-of-vocabulary terms (e.g., those terms whose translations are not available in the translation resources such as MRD). In CLIR, recent studies show that it is beneficial to use dedicated data mining and information extraction methods for obtaining high quality translations for named entities [8], which are the most important and most common type of OOV [12]. This again can potentially help MRD method to outperform MT based query translation.

Therefore, with all these recent developments in both MT and CLIR, we are interested in examining again the effect of using MT in query translation based CLIR. Our work has a great benefit of answering the question that, by using current MT systems as their query translation resource, can laymen searchers obtain CLIR performance that is comparable to that of using MRD. We will concentrate on using commercial MT systems such as Google Translate since they are the ones that most people can access and integrate into their CLIR process. To cover multiple CLIR scenarios, our research questions cover three important factors:

- Do pseudo relevance feedback (PRF) and the different CLIR query expansion (QE) methods make difference to the usage of MT and MRD in query translation? In CLIR, there are pre-translation QE that performed before translating the query with the help of an extra document collection at the query language side, post-translation QE that performed after the query is translated, and the combination of pre and post-translation QE, which is called combined QE. The motivation here is that PRF has demonstrated to gain significant improvement on retrieval effectiveness in both monolingual IR and CLIR, and different QE methods seem work differently.

- Do different query lengths make difference? Web search queries are often short, but queries in TREC like evaluation frameworks are often much longer. We, therefore, utilize the title, description, and narrative fields in the search topics to simulate short Web queries and long TREC queries, and examine the differences between query translation based on MT and that based on MRD.

- Do the handling of out-of-vocabulary (OOV) terms by dedicated NE translation component change the results obtained in the two previous research questions?

2. EXPERIMENT SETTINGS

Our experiments were performed between English queries and Chinese documents. The reasons for this selection include that CLIR between English and Chinese is still a very active research topic, and all the authors have been working on Chinese related issues.

2.1 CLIR Systems in the Experiments

At MT side, we selected Google Translate[1] as the out-of-box commercial statistical MT for translating queries. Google Translate is a Web multilingual service provided by Google Inc. The underneath MT system is a statistical based system, which was developed by Franz-Josef Och, and the system won the DARPA contest for speed machine translation in 2003[2]. We did not and could not modify any parameter of Google Translate. Although this prevents us from studying Google Translate in depth, this setting is actually close to what would happen when layman Web users incorporate Google Translate into their CLIR search. When using MT for translation, rather than translate the query terms one by one, we inputted the whole query into the MT system at once.

The bilingual MRD used for our dictionary-based CLIR was an English-Chinese lexicon generated from a parallel bilingual corpus automatically [13]. It was compiled by training GIZA++[3] 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 with translation probabilities for each Chinese translation alternative. The translation probabilities are obtained based on the normalized frequency of an English word and a Chinese word being linked together by word alignment.

During the translation of the queries with the MRD, to remove low probability translations which often are noises, a fixed threshold called Cumulative Probability Threshold (CPT) is selected. This is done by ranking the translations in decreasing order of their probabilities, then iteratively selecting translations from the top of the ranking until the cumulative probability of the selected translations firstly reaches or exceeds the threshold. A threshold of 0 thus corresponds to the using the single most probable translation (a well-studied baseline) and a threshold of 1 corresponds to the use of all translation alternatives in the dictionary.

In order to improve the coverage of the dictionary as much as possible, we adopted the back-off translation strategy [5] during the translation of the query terms. The back-off strategy starts with trying to match the surface form of an input English term to the surface forms of English term in the dictionary 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, if this still fails, stem the dictionary and match the surface form of the input term to stems of terms in the dictionary, if all else fails, match the stem of the input term to the stem of terms in the dictionary.

Once the queries were translated by either MT or MRD, we used Indri v.2.4[4] as the search engine to perform document retrieval on the Chinese collection.

For all the PRF tasks in the experiments, we used the Indri's build-in PRF module which is based on Lavrenko's relevance model [9] . Formula (1)[5] summarizes this implementation, where I is the original query, r is a term to be considered for query expansion, and D is a document.

$$P(r \mid I) = \frac{\sum_{D} P(r \mid D) P(I \mid D) P(D)}{P(I)} \tag{1}$$

The parameters used in our PRF tasks were selecting top 20 terms from top 20 returned documents. The relative weight between original queries terms and expanded queries is 0.5. This was based on our previous ad hoc exploration of the parameters in Indri.

2.2 Information Extraction-based Named Entity Translation Module

In MRD based CLIR, we adopted an NE translation component based on information extraction (IE) techniques. The IE component is designed to provide two functions in the MRD based query translation. 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 function was provided by the NYU English and Chinese HMM-based name taggers trained on several years of ACE (Automatic Content Extraction[6]) corpora. Both name taggers can identify seven types of names (Person, Geo-Political Entity (GPE), Location, Organization, Facility, Weapon and Vehicle) and achieve about 87%-90% F-measure on newswire [7].

The English HMM NE tagger [7] includes six states for each main name type, 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 the information about the internal structure 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 [5]. Within each of the name class states, a statistical bigram 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. Our IE component exploits the NYU name translation system [8], which usages 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 [2] 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 close 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 at 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. The NE pair then becomes the translation of each other. The corpora used for this approach were all GALE[7] MT training corpora and ACE 07 Entity Translation training corpora to conduct this process. 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 mixed of Chinese and English.

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 schemes. 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 abbreviations. For these types of translation alternatives, it 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. This error would reduce the retrieval effectiveness. To solve this problem, we utilized a popular commercial Web search engine[8] to obtain three numbers, which correspond to the numbers of web sites contain the NEs, the translation, and the NE and the translation together in the same page:

$$weight(NE_i, tran_{i,j}) = \alpha \frac{|NE_i \cap tran_{i,j}|}{|NE_i| + |tran_{i,j}|} \qquad (2)$$

where $weight(NE_i, tran_{ij})$ is the weight for translation j of NE i, $|NE_i, tran_{ij}|$ is the number of returned pages containing both NE_i and $tran_{ij}$; $|NE_i|$ is the number of returned pages containing NE_i and $|tran_{ij}|$ is the number of returned pages containing $tran_{ij}$; α is a constant to adjust the score in case the weight is too small.

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

$$\sum_{j=1}^{n} weight'(NE_i, tran_{ij}) = 1 \qquad (3)$$

where n is the total number translations of NE_i.

Just like that 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 baidu.com[9] which is the most popular Chinese search engine.

2.3 Search Collections and Topics

The test collection used in our study combined TDT4 and TDT5 Multilingual News Text corpora. The combined collection contains 83,627 Chinese documents of 328M and 306,498 English documents of 1.13G. The collection was used for NIST multilingual Topic Detection and Tracking evaluations [reference].

We selected 44 TDT English topics and manually translated them into Chinese for monolingual Chinese search. We also rewrote them into TREC topic style with title, description and narrative fields (see Figure 1). Queries were automatically extracted from the topics with short queries containing titles only (T query), medium queries with title and description fields (TD query), and long queries with all the three fields (TDN query). The average length of the queries were: T query (4 terms), TD query (27 terms), and TDN query (127 terms).

<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 1: An example of the modified TDT topic

All Chinese texts and queries were segmented using the Stanford Chinese word segmenter[10]. The Porter stemmer[11] was used to stem English texts, queries and the dictionary when necessary. Stop words were removed using a Chinese stopword list[12] and an English stopword list[13]. After all these preprocessing steps, the queries are ready for translation by either MT or MRD,

To answer our research questions, we conducted the following experiment runs:

- **Monolingual Baseline:** a monolingual run that retrieves Chinese documents using the manually translated Chinese topics.

- **MT CLIR Baseline:** a CLIR run that translates English topics with MT.

- **MT QE-PreTrans:** a CLIR run that uses MT to translate queries and uses the pre-translation QE method for PRF.

- **MT QE-PostTrans:** a CLIR run that uses MT to translate query and uses the post-translation QE method for PRF.

- **MT QE-Combine:** a CLIR run that uses MT to translate query and uses the combination of pre-translation and post-translation QE for PRF.

- **MRD CLIR Baseline:** a CLIR run that translates English topics using the MRD.

- **MRD NE Baseline:** a CLIR run that translates queries using the MRD and translates NEs using the NE module.

- **MRD QE-PreTrans:** a CLIR run that uses the MRD to translate the queries and uses the pre-translation QE method for PRF.

- **MRD QE-PostTrans:** a CLIR run that uses the MRD to translate the queries and uses the post-translation QE method for PRF.

- **MRD QE-Combine:** a CLIR run that uses the MRD to translate the queries and uses the combined QE method for PRF.

Figure 2 illustrates the experiment procedure involving MT or MRD based query translation with and without PRF.
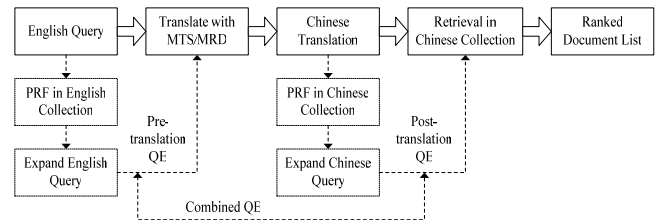

Figure 2: The whole experiment procedure

When MT is used for query translation, it only outputs the one best translation for each query term. However, for MRD runs, we tried different CPTs from 0.0 to 1.0 with an increment of 0.1 at each time. The results presented here are the best run results of the 11 CPTs (CPT=0.5 for all three query lengths).

We use mean average precision (MAP) as the basis of evaluation for all experiments and the two tailed paired samples t-test at P-value < 0.05 for statistical significance test.

---

[9] http://www.baidu.com

[10] http://nlp.stanford.edu/software/segmenter.shtml
[11] http://tartarus.org/~martin/PorterStemmer/
[12] http://www.unine.ch/info/clef/englishST.txt
[13] http://bbs.ir-lab.org/cgi-bin/topic.cgi?forum=3&topic=127

3. RESULTS AND DISCUSSION

3.1 MT-based Query Translation in CLIR without PRF

Table 1 shows the MAP results of all the experiment runs. For our discussion, we first examine the runs without PRF, which include MT CLIR Baseline, MRD CLIR Baseline, MRD NE Baseline and Monolingual IR.

Without any query expansion improvement from feedback, most CLIR runs could not generate comparable results to Monolingual IR, particular MRD based query translation method, no matter using or not using NE translation module, and whether the queries are T, TD or TDN.

Table 1: The Mean Average Precision (MAP) results of MT-based and MRD-based CLIR runs

Run ID	T	TD	TDN
Monolingual IR	0.4739	0.5817	0.6215
MT CLIR Baseline	0.4446	0.5536	0.6170
MT QE-PreTrans	0.4922	0.5443	0.5580
MT QE-PostTrans	0.5284	0.6031	0.6292
MT QE-Combine	0.5604	0.5833	0.6001
MRD CLIR Baseline	0.3336	0.4251	0.4701
MRD NE Baseline	0.3934	0.5034	0.5563
MRD QE-PreTrans	0.3714	0.4377	0.4477
MRD QE-PostTrans	0.4118	0.5080	0.5182
MRD QE-Combine	0.4415	0.5007	0.5170

However, MT based query translation method under TDN queries did generate comparable results to Monolingual IR (99% of Monolingual IR). The performance of MT based query translation was much closer to Monolingual IR under T and TD queries too (at least 94% of Monolingual IR). This indicate that it is clear that MT based query translation is a better query translation method than MRD based method when there is no query expansion.

Table 2: Significant T-test on Comparison (* indicates that the improvement is statistically significant)

	Impr. over MRD CLIR Baseline	Impr. over MRD NE Baseline
MT CLIR Baseline (T)	+33.27%*	+13.01%
MT CLIR Baseline (TD)	+30.23%*	+9.97%
MT CLIR Baseline (TDN)	+14.69%*	+10.91%*

Table 2 shows the significant testing between the runs of MT against that of MRD. MT-based query translation significantly outperformed MRD CLIR Baseline in all three lengths of queries. However, it is interesting to see that the superiority of MT method is shrinking along with the increasing of query length. This is counter intuitive since we would think that MT performs better with more context (longer queries). We will conduct further analysis on this issue.

As shown in Tables 1 and 2, introducing NE translation module does improve the performance. Although MRD NE Baseline is still inferior to that of MT CLIR Baseline, the difference is not significant in two of the three query lengths (T and TD).

3.2 MT-based Query Translation in CLIR with PRF

When PRF was added into MT-based query translation CLIR, all runs were at least 90% of monolingual performance, with many of them over 100% and the highest is 120% (MT QE-Combine at T queries). Therefore, all these MT-based CLIR run have achieved the state of the art CLIR performance which is either comparable to or higher than the performance of corresponding monolingual run. It is interesting to notice that MT works even with short queries (Query T). Lacking of context was a worry about using MT for query translation. All these results confirm again that MT is a valid tool for translating queries in CLIR.

Table 3: Comparison of QE methods for MT-based CLIR (* indicates the improvement is statistically significant)

		Perc. Of Mono IR	Impr. over MT-based CLIR
T	MT QE-PreTrans	103.86%	10.71%*
	MT QE-PostTrans	111.50%*	18.85%*
	MT QE-Combine	118.25%*	26.05%*
TD	MT QE-PreTrans	93.57%	-1.68%
	MT QE-PostTrans	103.68%	8.94%*
	MT QE-Combine	100.28%	5.36%
TDN	MT QE-PreTrans	89.78%*	-9.56%*
	MT QE-PostTrans	101.24%	1.98%
	MT QE-Combine	96.56%	-2.74%

As shown in Table 3, although in general query expansion (QE) based on PRF helps MT based query translation CLIR, its effect is affected by the query lengths. QE was most helpful when queries are short. Under T queries, all three QE methods significantly outperformed not only the MT CLIR Baseline, but also the Monolingual Baseline. The post-translation QE and the combined QE runs even significantly outperformed the monolingual baseline. However, when queries are longer, like in the cases of medium TD queries and long TDN queries, the effectiveness of QE is shrinking. Not all QE methods generate significant improvement over even the CLIR baseline. In fact, several runs (MT QE-PreTrans under TD queries, MT QE-PreTrans and MT QE-Combine under TDN queries) generated negative impact on retrieval effectiveness. In the case of MT QE-PreTrans under TDN queries, the negative effect was statistically significant. Therefore, it should be careful to use QE for MT based query translation, particularly when the queries are not short.

When comparing between the corresponding MT-based and MRD-based CLIR runs with or without PRF (see Table 4), MT-based CLIR all improved between 16%-33%. Statistical testing shows that all these improvements are statistically significant with two tailed paired-samples t-tests.

Table 4: Improvement of MT-based CLIR over MRD-based CLIR (* indicates the improvement is statistically significant)

	Baseline	PreTrans	PostTranss	Combine
T	33.27%*	32.53%*	28.31%*	26.93%*
TD	30.23%*	24.35%*	18.72%*	16.50%*
TDN	31.25%*	24.64%*	21.42%*	16.07%*

3.3 Error Analysis

When examining individual topics, we noticed that the quality of MT results were quite accurate, even for short T queries. There were many named entities in the topic statements, and Google Translate system handled them well. There were still a few named entities that Google Translate system cannot handle. But they were also out-of-vocabulary terms for the MRD. So they did not make obvious difference between MT runs and MRD runs.

However, there were 10 to 12 topics that MRD runs obtained better results than the MT runs, even though the MT translation of the queries looks better than the corresponding MRD translated queries. This probably indicates that MT generated one-best translation, although in reasonable quality, still can miss certain relevant documents which may be captured by the n-best translation approach used in MRD based CLIR. Therefore, maybe a better strategy is to combine these two approaches?

4. CONCLUSION

With further development of modern commercial statistical MT systems, it is possible to use an out-of-box commercial MT system for translating queries in CLIR task. Because many such MT systems are available on the Web, these systems can easily be used by layman Web users to build up CLIR capabilities. In this paper, therefore, we tried to examine the performance of using MT for query translation. The experiment results show that MT is a better tool for translating queries in CLIR than using MRD. Its performance is significantly better than that of using a MRD, not matter whether relevance feedback (RF) is applied or not.

When there is no help from RF, MT based query translation works the best with long queries. Its performance can reach to 99% of that of monolingual search. This is the state of the art of CLIR performance. This result may help to remove the uneasy about the worry that MT does not work under the situation that there is not much context available for translation.

However, when RF is integrated into CLIR, MT method works the best with short queries. Not only it outperforms the monolingual search under short queries, but also it achieved significant improvement over the monolingual run with both post translation based QE and combined QE methods.

Similar to that in MRD-based CLIR, when RF is helpful in MT based query translation in CLIR, it is often the post-translation QE and/or the combined QE that are the best method among the three CLIR-QE methods. In fact, post-translation QE seems to be helpful whatever the query length is, whereas pre-translation QE only works for short queries and the combined QE does not work for long queries.

Therefore, MT is not just a good query translation method in CLIR, and it actually is a better tool for query translation than MRD, not matter whether the query length are short or long, and whether RF is used or not.

Besides the main findings about using MT for query translation, our results also confirms that a specially handing of NE translation is important in CLIR, especially MRD based query translation method.

Our future work include repeat the experiments in other test collections to examine whether our findings are repeatable, and study more complex strategy of applying MT based query translation methods.

5. REFERENCES

[1] Aljlayl, M. and Frieder, O. Effective Arabic-English cross-language information retrieval via machine-readable dictionaries and machine translation. In *Proceedings of the tenth international conference on Information and knowledge management.* 2001.

[2] 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).* 2007.

[3] Ballesteros, L. and Croft, B. Dictionary Methods for Cross-Lingual Information Retrieval. *Proceedings of the 7th International DEXA Conference on Database and Expert Systems.* pages 791-801.1996

[4] Ballesteros, L. and Croft, W. B. Resolving Ambiguity for Cross-language Retrieval. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* pages 64-71.1998

[5] 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.* 1997.

[6] 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.* pages 338-344. 2003.

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

[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] Lavrenko, V., Choquette, M. and Croft, W. B. Cross-Lingual Relevance Models. *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* page 175-182.2002

[10] Oard, D. W. A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval. In *the Third Conference of the Association for Machine Translation in the Americas (AMTA).* 1998.

[11] Och, F. J. Minimum Error Rate Training for Statistical Machine Translation. In *the Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL2003).* 2003.

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

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

[14] Xu, J. and Weischedel, R. A Probabilistic Approach to Term Translation for Cross-Lingual Retrieval;. *Language Modeling for Information Retrieval.* W. B. Croft and J. Lafferty (Eds). 2003.