Mining Translations of Web Queries from Web Click-through Data*

Rong Hu[1], Weizhu Chen[2], Jian Hu[2], Yansheng Lu[1], Zheng Chen[2], Qiang Yang[3]

[1]Department of Computer Science, Huazhong University of Science and Technology, Wuhan 430074, China
{ronghu, lys}@mail.hust.edu.cn
[2]Microsoft Research Asia, 5F, Sigma Center, 49 Zhichun Road, Beijing 100080, China
{wzchen, jianh, zhengc}@microsoft.com
[3]Department of Computer Science, Hong Kong University of Science and Technology, Clearwater Bay, Hong Kong
qyang@cse.ust.hk

Abstract

Query translation for Cross-Lingual Information Retrieval (CLIR) has gained increasing attention in the research area. Previous work mainly used machine translation systems, bilingual dictionaries, or web corpora to perform query translation. However, most of these approaches require either expensive language resources or complex language models, and cannot achieve timely translation for new queries. In this paper, we propose a novel solution to automatically acquire query translation pairs from the knowledge hidden in the click-through data, that are represented by the URL a user clicks after submitting a query to a search engine. Our proposed solution consists of two stages: identifying bilingual URL pair patterns in the click-through data and matching query translation pairs based on user click behavior. Experimental results on a real dataset show that our method not only generates existing query translation pairs with high precision, but also generates many timely query translation pairs that could not be obtained by previous methods. A comparative study between our system and two commercial online translation systems shows the advantage of our proposed method.

Introduction

Cross-Lingual Information Retrieval (CLIR) has become increasingly important in Information Retrieval (IR). The main difference between CLIR and traditional IR is that the query language differs from the page language (Pirkola *et al*. 2001), so the research effort in the area of CLIR has mainly focused on how to translate a source query into a target query in the page language. In addition to CLIR, query translation is also valuable for many other applications, such as machine translation (MT), question answering, and reading/writing assistant. Therefore, it is desirable to provide automatic query translation techniques for the above applications.

In this paper, we propose a novel solution to translate web queries based on the analysis of user behavior in the click-through data. Compared to previous methods, our method generates large-scale and timely query translation pairs guided by a small set of seed word pairs from a dictionary, without relying on additional knowledge or complex models. We currently use this method to mine English-Chinese query translation pairs throughout the paper. However, the method described here does not consider any linguistic constraints, and hence could be applied to any language pairs.

Intuitively, millions of users across the world issue queries to search engines in various languages daily, which forms large-scale and live cross-lingual click-through data. The data source has covered rich user language knowledge about queries as well as their relationships to the clicked pages. Moreover, many URLs in the click-through data contain some language information. We leverage the language knowledge encoded in the URLs to construct extensive bilingual URL pair patterns. The objective of this paper is to uncover user behavior in the click-through data to mine URL pair patterns, and eventually generate query translation pairs.

We make two basic assumptions based on our observation on the click-through data collected from a famous commercial web search engine.

The first assumption is that there exists some naming convention in the URLs which specifies the language information of the corresponding pages, such as *en-us, zh-cn, .cn, .us*. In order to verify this assumption, we first extracted English words and their corresponding Chinese translation words from an English-Chinese dictionary, and obtained 22,000 word translation pairs. Then we collected the URLs clicked by these words from the click-through data. Finally, among 99,900 distinct URLs clicked, 51,000 URLs (over 50%) were found to be encoded with language information by human judges. As shown in Figure 1, the English URL is http://www.fedex.com/us/, and the Chinese corresponding URL is http://www.fedex.com/cn/. They share the common substring http://www.fedex.com/, and the only difference is the substring indicating the language type, i.e., *us* is used to indicate the English version, while

*The work was done when the first author was doing internship at Microsoft Research Asia.

cn is used to indicate the Chinese version. We denote these two URLs as a bilingual URL pair.

Figure 1: An example of a bilingual URL pair and the corresponding queries

The second assumption is that the clicked URLs are relevant to the query. It provides the connection between URLs and queries, which can be exploited to extract candidate query pairs. In the above example, http://www.fedex.com/us/ can be considered to be relevant to English queries "fedex" and "express", whereas http://www.fedex.com/cn/ can be considered to be relevant to Chinese queries "联邦快递" and "国际快递服务", since users click the URLs after submitting the queries to the search engine. By employing several filtering strategies, we can extract query translation pair "fedex" and "联邦快递" from four candidate queries.

Based on two assumptions, our solution can be divided into two steps. The first step is to use a dictionary for finding seed queries to mine URL pair patterns from the click-through data. The second step is to use the mined URL pair patterns to generate the candidate query pairs. The result query pairs are generated based on co-occurrence analysis of the click-through data.

We evaluate our proposed method over a real web click-through dataset. We compare our translated result with a bilingual dictionary CEDICT[1]. For the queries covered by CEDICT, we judged the precision of our translation method by the dictionary. For the out of vocabulary (OOV) queries, we first compared our method to two of the best MT systems available in the Internet: Systran[2] and Google translator[3]. Then we asked human experts to assess the precision of these queries. Experimental results show that our method outperforms Google translator and Systran by 3.3% and 126.8%, respectively.

To the best of our knowledge, this is the first paper which proposes to leverage click-through data and bilingual URL pairs to mine query translation pairs and demonstrates its effectiveness.

The rest of this paper is organized as follows. We first introduce related work and then the outline of our method. After that we present the detailed method and experimental results. Finally, we conclude the paper.

[1] http://www.mandarintools.com/cedict.html

[2] http://babelfish.altavista.com/

[3] http://www.google.com/translate_t/

Related Work

Lavrenko et al. (2002) introduced a relevance model in the target language to rank the documents for CLIR. However, most approaches to CLIR perform query translation followed by monolingual IR. A natural idea is to combine IR with existing MT systems (Fujii & Ishikawa 2001). However, MT systems require a complex process to produce a grammatical translation, which is not fully exploited by current IR models due to that queries are often very short and without enough contextual information.

Dictionary-based translation approaches (Pirkola *et al*. 2001) are proposed to simplify the translation process. However, besides the problem of coverage, they face the problem of ambiguity, i.e. selection of the most appropriate translation in the given context.

Corpus-based translation approaches are proposed to remedy the two major issues: translation disambiguation and OOV words. They used a statistical model based on a corpus, including parallel corpus (Nie *et al*. 1999; Resnik & Smith 2003), anchor texts (Lu, Chien, & Lee 2001), comparable corpus (Fung & Yee 1998), and mixed-language pages (Huang, Zhang, &Vogel 2005). However, these approaches require crawling web pages. And parallel corpora and anchor texts are quite limited. Using comparable corpora and mixed-language pages cannot achieve high translation quality due to lack of parallel correlation. In addition, (Can & Li 2002) searched the web for base noun phrase candidates and used an EM algorithm to select translations.

In contrast to previous studies that rely exclusively on MT systems, bilingual dictionaries, or corpora, we explore the role of the readily available click-through data. One related work (Ambati & Rohini 2006) exploited the query logs of the target language to learn translation models that could be used to perform CLIR. Their work is different from ours, as they used a bilingual dictionary to translate queries in the monolingual click-through data, and disregarded the proper nouns and domain specific terms. More recently, in (Gao *et al.* 2007), the authors also used query logs of different languages. Their method is different from ours on two aspects. First, it used query co-occurrence information in the query log, while ours uses URL pattern as a bridge to mine click-through log. Second, our method does not depend on the resources that theirs used to generate candidates, such as parallel corpora and web corpora.

Although click-through data have received much attention in the task of information retrieval and have been successfully used for improving information retrieval, they have only recently been explored as a resource of automatic acquisition of hidden relations in information extraction. In (Pasca *et al.* 2006), the authors introduced a method to extract attributes for various classes of objects based on query logs rather than page collections. In (Shen *et al.* 2007), the authors proposed to mine query hierarchies from click-through data, since the user's understanding of the query as well as its relation to the web pages is encoded in the click-through data.

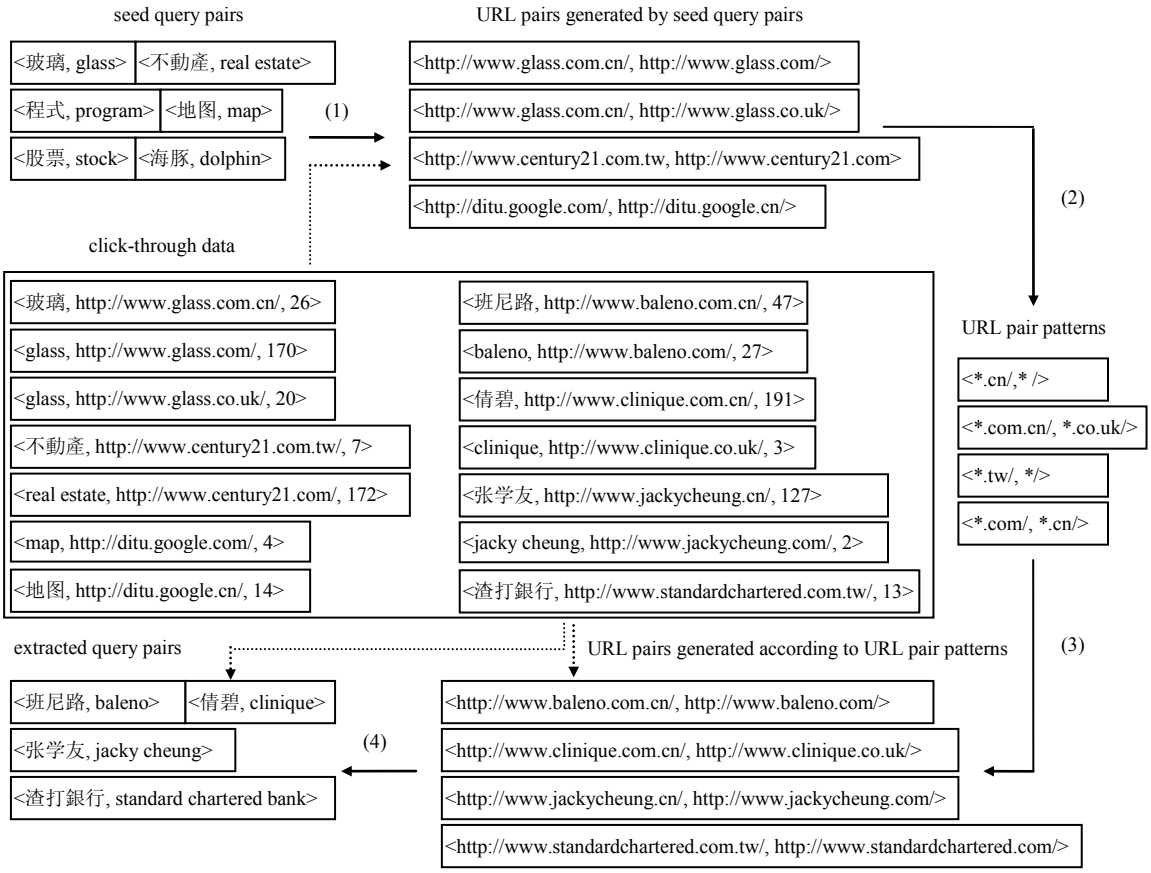

Figure 3: A fragment of data flow using our proposed method

Outline of Our Proposed Method

Consider a typical user search session: a user u submits a query q in Market m, a search engine returns a ranked list of web pages. Then the user clicks on the page p. Hence, a session is defined as <u, q, p, m>. Click-through data are the accumulation of a large number of sessions. In our case, the Market symbol denotes either English or Chinese. Figure 2 illustrates the outline of our click-through data-based query translation method.

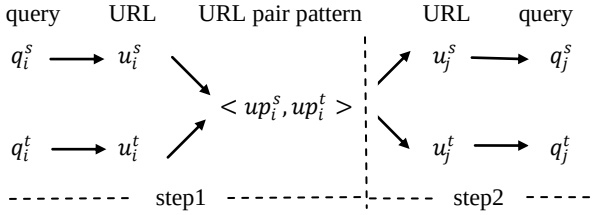

Figure 2: Outline of click-through data-based query translation method

Our method consists of two main steps. It starts from a handful of seed query translation pairs (e.g. <q_i^s, q_i^t>), identifies their corresponding URL pairs (e.g. <u_i^s, u_i^t>), and generates URL pair patterns (e.g. <up_i^s, up_i^t>). Then it pairs up the candidate queries regarding to the URL pairs (e.g. <u_j^s, u_j^t>) that are extracted by URL pair patterns as query translation pairs (e.g. <q_j^s, q_j^t>). Figure 3 shows a fragment of data flow using our proposed method.

Web Query Translation Extraction

Identifying Bilingual URL Pair Patterns

The identification of URL pair patterns is guided by seed query pairs from a bilingual dictionary. Two groups of URLs relevant to these query pairs are found from the click-through data (see (1) in Figure 3). Then the similarity between the URLs in two groups is computed. We used edit distance, which is a measure based on the minimum number of edit operations (insertion, deletion, and substitution etc.) needed to transform one URL into the other. The advantage of the measure is that it takes into account both word order and the words. It is therefore very suitable to our situation where one URL in a pair is often the substitution of the other on the same position. The similarity is inversely proportional to the ratio between edit distance and the length of the shorter URL.

$$sim(u_i^s, u_i^t) = 1 - \frac{EditDistance(u_i^s, u_i^t)}{\min(|u_i^s|, |u_i^t|)}$$

The URL pairs with similarity score above a certain threshold are chosen to extract patterns (see (2) in Figure 3). Let us explain how to obtain URL pair patterns from URL pairs with two URLs illustrated in Figure 1. Since the only difference between their names is *us* and *cn*, we can follow the first assumption to extract the pattern <*.*us*, *.*cn*>. We adopted the largest common string algorithm to collect candidate patterns and counted frequency. We also used several simple but effective filtering techniques. The URL patterns with frequency larger than two are selected.

Matching Query Translation Pairs

The generated URL patterns are used to guide URL extraction from the click-through data. That is, the bilingual URL pairs that correspond to the URL pair patterns are selected (see (3) in Figure 3), which can limit the URL set by only considering the reliable URL pairs.

Once URL pairs are generated, we can find out all queries that have been used to retrieve these URL pairs based on the second assumption. Due to the noise in the click-through data, there may have some query pairs loosely related to each other. To generate query translation pairs, we need to filter out candidate queries.

Consider a pair of candidate queries: q_i^s in the source language, and q_i^t in the target language. Intuitively, if both q_i^s and q_i^t occur frequently in the query pairs associated with the same URL pairs, then they may have a high probability to be the translation of each other. However, co-occurrence alone does not accurately reflect the association between two queries, so we must consider their individual occurrence. If either query, for example q_i^s, may pair up with other queries, that is, query q_i^s is also associated with other URL pairs together with the queries rather than q_i^t, then query q_i^t may have less chance to be the translation of query q_i^s.

Therefore, we introduce the following confidence score for each translation candidate to measure the extent to which the two queries are relevant.

$$score(< q_i^s, q_i^t >) = \frac{f(q_i^s, u_i^s) \times f(q_i^t, u_i^t)}{f(u_i^s) \times f(u_i^t) \times f(q_i^s) \times f(q_i^t)}$$

Here, $f(q_i^s, u_i^s)(f(q_i^t, u_i^t))$ stands for the number of sessions where the query is clicked in case the URL appears in the URL pairs. $f(q_i^s)\,(f(q_i^t))$ is the number of sessions that contain $q_j^s\,(q_j^t)$. And $f(u_i^s)\,(f(u_i^t))$ is the number of sessions that contain $u_i^s(u_i^t)$.

From the above formula, it is obvious that if two queries closely associate with the URL pair, and they often co-occur compared to their individual occurrence in the candidate pairs; then they are most probable to be the translations of each other, which accords with our intuition. According to the confidence score, only those query pairs with large scores are output (see (4) in Figure 3).

Extensions

Our method can be extended to generate more query pairs. The basic idea is that additional URL pair patterns can be derived from the query pairs extracted in the previous iteration to increase coverage. This iterative acquisition method is similar to the idea of bootstrapping (Riloff & Jones 1999), which learned extraction patterns from queries and then exploited the learned extraction patterns to identify more queries belong to the same category.

Experiments

Dataset

The input of our experiment involves two parts: click-through data, from which query translation pairs were mined, and seed queries acquired by the dictionary, from which our mining method started up.

Click-through data: The click-through data were collected from a commercial web search engine for eight months spanning from October 2006 until June 2007. And then the sessions containing English and Chinese queries were extracted. Next, the click-through data were preprocessed so that the queries clicked by the same URL were aggregated. Consequently, about 100M unique query-URL pairs, together with the click counts constitute the click-through data. The click-through data used are unbalanced since among 100M unique query-URL pairs, there are about 1 million unique Chinese queries, 30 million unique English queries, and 10 million unique English queries with occurrence frequency more than 5.

Seed queries: We randomly sampled 5K queries from a dictionary as seed queries.

Results

To evaluate the effectiveness of our method, two evaluation measures were used: recall, defined as the percentage of queries covered given the set of queries, and precision, denoted as the percentage of queries correctly translated given the set of queries. And three state-of-the-art translation systems were used: a Chinese-English dictionary CEDICT having 44,782 entries totally (denoted as DT), a MT system named Systran (denoted as MT1); and a MT system named Google Translator (denoted as MT2). The correctness of translation was first judged by DT, then human translators. A translation was considered correct when it can be matched exactly with one of the reference translations.

The recall values are shown in the fourth column of Table 1. Since it is difficult to directly examine the coverage of our method, we evaluate the recall of other methods compared to our method. Specifically, given the number of queries (see the second column of Table 1), we first checked how many of them were covered (see the third column of Table 1) by DT and MT, and then computed the recall values based on it.

As shown in Table 1, among the queries mined from the real click-through data, 11.1% of them are included in DT. Moreover, among all the OOV queries, only 16.1% and 34.8% of them are included by MT1 and MT2, respectively. These results indicate that by effectively leveraging click-through data, our mining method could discover many query translation pairs unavailable from the state-of-the-art methods.

Table 1: Recall in comparison with our method

methods	#queries	#queries covered	recall
DT	718	80	11.1%
MT1	638	103	16.1%
MT2	638	222	34.8%

The precision values are shown in the fourth column of Table 2. We took into account only the queries which have been not covered in DT. More specifically, given the number of OOV queries (see the second column of Table 2), we first checked how many of them were translated correctly (see the third column of Table 2), and then computed the precision values based on it.

As shown in Table 2, among 638 OOV queries, 88 queries are translated correctly by MT1, 193 queries are translated correctly by MT2, and 200 queries are translated correctly by our method. Thus, our method achieves 126.8% and 3.3% relative improvements on precision in comparison with MT1 and MT2, respectively. The result indicates that our mining method could accurately indentify many query translation pairs that the state-of-the-art methods fail to translate.

To further test coverage and precision, we conducted experiments on the queries selected based on frequency and meaningfulness of the queries. We find that among the 100 English & 100 Chinese queries, 62 & 95 queries have their translations in the click-through log, and 45 & 78 queries can be translated in our method. Our method outperforms MT1 and MT2 in precision by 82.3% and 44.9%, respectively.

When comparing with the bilingual dictionary-based method used in (Gao *et al.* 2007), we selected 500 Chinese queries from the log data. We find that the precision scores are 0.084 and 0.622 in theirs and ours, respectively. The recall scores are 0.353 and 0.665 in both methods, respectively. This validates the claim that our method outperforms theirs.

Table 2: Precision comparisons on OOV queries

methods	#queries	#queries translated correctly	precision
MT1	638	88	13.8%
MT2	638	193	30.3%
our method	638	200	**31.3%**

Some example data obtained in our experiment are listed in the following. Table 3 presents the confidence scores of two groups of candidate query pairs. It is obvious that the query pairs with the largest scores are the correct pairs.

Table 3: Scores of sample candidate query pairs

q_i^s	q_i^t	score	q_i^s	q_i^t	score
卡西欧 (casio)	casio	1.35E-06	克莱斯勒 (chrysler)	chrysler	2.64E-05
	casio camera	2.93E-07		chrysler australia	2.87E-06
	casio digital camera	5.40E-07		chrysler jeep	7.66E-06

Samples of mined query translation pairs not covered by DT, but covered by MT and our method are listed in Table 4. After examining these queries, we can find that DT is not as effective as MT and our method when translating technical terminologies and named entities.

Samples of mined query translation pairs not covered by DT and MT are provided in Table 5. We can find that our method is especially useful for translating technical terminologies, and named entities, such as person names, web site names, product names, and so on.

When comparing Table 4 with Table 5, we can find that although MT can translate some of OOV queries, it cannot achieve comparable coverage as our method. The strength of our method lies in translating technical terminologies and named entities. Since they are among the most information-bearing entities, translating them correctly will improve the performance of CLIR, MT and query answering. However, for long-tail queries, our current work cannot handle it well, and we will investigate them in the future.

Table 4: Samples of mined English-Chinese translation pairs that are not covered by DT, but covered by MT

Technical terminologies		Named entities	
dependencies	依赖项	fujitsu	富士通
emulator	模拟器	taobao	淘宝
casting	强制转换	harry potter	哈利波特
buffering	缓冲	ferrari	法拉利
concurrency	并发控制	alexander	亚历山大
constructor	构造函数	sapporo	札幌市
packaging	封装	osaka	大阪市
disassembly	反汇编	winnie	小熊維尼

Table 5: Samples of mined English-Chinese translation pairs that are not covered by DT and MT

Technical terminologies		Names of person	
quickstarts	入门	ultraman	奥特曼
live onecare	安全扫描程序	conan	柯南
mozilla firefox	火狐浏览器	Names of product or brand	
walkthrough	演练	icefish	冰鱼
currency rate	汇率	healthbanks	健银
garden show	庭院设计	yahoo	奇摩
others			
labor department	劳工處	national ballet of china	中央芭蕾舞团

Discussions

In this section, we investigate the causes of incorrect translations and try to propose some solutions. By inspecting these translations, we find that many of them have a relation to the original meaning.

First, the wrong translation may be the partially correct translation. For instance, "uninstall" was translated into "卸载工具" (uninstall tool). "print control" was translated into "打印" (print). One possible solution is to consider the queries having common phrases. For example, we find that the English queries "learn macros", "creating macros", and "learning macros" are all highly relevant to the Chinese "宏" (macro). Since they share the word "macros" in common and are different in the remaining part, it is less likely to translate them into "宏".

Second, the wrong translation may be similar or related to the correct translation. Many extracted query pairs in our query translation method belong to the case. For instance, "flight" and "机票" (ticket), "logarithm" and "数学" (math), "algorithms" and "数据挖掘" (data mining), "integer" and "数据类型" (data type). This shows our method is sometimes unable to distinguish true translations from statistically associated words well. One possible solution is to combine our method with a dictionary, since we can check the translations by comparing them with the dictionary, when queries exist there.

However, even though some mined translations are not the equivalent translations of the source queries, they are semantically relevant ones, thus they may be helpful to improve CLIR. We leave this as our future work.

Conclusions

In this paper, we have proposed a novel solution to leverage user behavior in the click-through data to extract query translation pairs. Our solution is a two-stage method: identifying bilingual URL pair patterns and matching query translation pairs. Different from previous methods that depend on strong language resources, our solution is a weak-resource learning method with large-scale automatic knowledge acquisition from click-through data. To demonstrate the effectiveness of our solution, we evaluate its performance over a real web click-through dataset. Experimental results show that the proposed method yields a number of translations not covered by the state-of-the-art methods, as well as many timely query translation pairs that cannot be obtained by these methods.

Acknowledgement

Qiang Yang would like to thank the support of Hong Kong CERG Project 621307. We thank Peng Bai and the anonymous reviewers for their useful comments.

References

Ambati, V.; and Rohini, U. 2006. Using monolingual clickthrough data to build cross lingual search systems. In *Proceedings of New Directions in Multilingual Information Access, Workshop of SIGIR'06*.

Cao, Y.-B.; and Li, H. 2002. Base noun phrase translation using web data and the EM algorithm. In *Proceedings of COLING'02*.

Fujii, A.; Ishikawa, T. 2000. Applying machine translation to two-stage cross-language information. In *Proceedings of AMTA'00*, 13-24.

Fung, P.; and Yee, L. 1998. An IR for translating new words from nonparallel, comparable texts. In *Proceedings of COLING-ACL*, 414-420.

Gao, W.; Niu, C.; Nie, J.; Zhou, M.; Hu, J.; Wong, K. and Hon, H. 2007. Cross-lingual query suggestion using query logs of different languages. In *Proceedings of SIGIR'07*, 463-470.

Huang, F.; Zhang, Y.; and Vogel, S. 2005. Mining key phrase translations from web corpora. In *Proceedings of the Human Language Technologies Conference* (HLT-EMNLP 2005), 226-234.

Lavrenko, V., Choquette, M.; and Croft, W. B. 2002. Cross-lingual relevance models. In *Proceedings of SIGIR'02*, 175-182.

Lu, W.; Chien, L.; and Lee, H. 2001. Anchor text mining for translation of web queries. In *Proceedings of ICDM '01*, 401-408.

Nie, J.; Simard, M.; Isabelle, P.; and Durand, R. 1999. Cross-Language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *Proceedings of SIGIR'99,* 74-81.

Paşca, M.; Lin, D.; Bigham, J.; Lifchits, A.; and Jain, A. 2006. Organizing and searching the world wide web of facts-step one: the one-million fact extraction challenge. In *Proceedings of AAAI'06*, 1400-1405.

Pirkola, A.; Hedlund, T.; Keskustalo, H.; and Järvelin, K. 2001. Dictionary-based cross-language information retrieval: problems, methods, and research findings. *Information Retrieval*, 4(3/4), 209-230.

Resnik, P.; and Smith, N. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3), 349-380.

Riloff, E.; and Jones, R. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of AAAI'99*, 474-479.

Shen, D.; Qin, M.; Chen, W.-Z.; Yang, Q.; and Chen, Z. 2007. Mining web query hierarchies from clickthrough data. In *Proceedings of AAAI'07*, 341-346.