Exploiting Bilingual Information to Improve Web Search

Wei Gao[1], **John Blitzer**[2], **Ming Zhou**[3], and **Kam-Fai Wong**[1]
[1]The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, China
{wgao,kfwong}@se.cuhk.edu.hk
[2]Computer Science Division, University of California at Berkeley, CA 94720-1776, USA
blitzer@cs.berkeley.edu
[3]Microsoft Research Asia, Beijing 100190, China
mingzhou@microsoft.com

Abstract

Web search quality can vary widely across languages, even for the same information need. We propose to exploit this variation in quality by learning a ranking function on bilingual queries: queries that appear in query logs for two languages but represent equivalent search interests. For a given bilingual query, along with corresponding monolingual query log and monolingual ranking, we generate a ranking on pairs of documents, one from each language. Then we learn a linear ranking function which exploits bilingual features on pairs of documents, as well as standard monolingual features. Finally, we show how to reconstruct monolingual ranking from a learned bilingual ranking. Using publicly available Chinese and English query logs, we demonstrate for both languages that our ranking technique exploiting bilingual data leads to significant improvements over a state-of-the-art monolingual ranking algorithm.

1 Introduction

Web search quality can vary widely across languages, even for a single query and search engine. For example, we might expect that ranking search results for the query 托马斯 霍布斯 (*Thomas Hobbes*) to be more difficult in Chinese than it is in English, even while holding the basic ranking function constant. At the same time, ranking search results for the query *Han Feizi* (韩非子) is likely to be harder in English than in Chinese. A large portion of web queries have such properties that they are originated in a language different from the one they are searched.

This variance in problem difficulty across languages is not unique to web search; it appears in a wide range of natural language processing problems. Much recent work on bilingual data has focused on exploiting these variations in difficulty to improve a variety of monolingual tasks, including parsing (Hwa et al., 2005; Smith and Smith, 2004; Burkett and Klein, 2008; Snyder and Barzilay, 2008), named entity recognition (Chang et al., 2009), and topic clustering (Wu and Oard, 2008). In this work, we exploit a similar intuition to improve *monolingual* web search.

Our problem setting differs from cross-lingual web search, where the goal is to return machine-translated results from one language in response to a query from another (Lavrenko et al., 2002). We operate under the assumption that for many monolingual English queries (e.g., *Han Feizi*), there exist good documents in English. If we have Chinese information as well, we can exploit it to help find these documents. As we will see, machine translation can provide important predictive information in our setting, but we do not wish to display machine-translated output to the user.

We approach our problem by learning a ranking function for *bilingual queries* – queries that are easily translated (e.g., with machine translation) and appear in the query logs of two languages (e.g., English and Chinese). Given query logs in both languages, we identify bilingual queries with sufficient clickthrough statistics in both sides. Large-scale aggregated clickthrough data were proved useful and effective in learning ranking functions (Dou et al., 2008). Using these statistics, we can construct a ranking over *pairs* of documents, one from each language. We use this ranking to learn a linear scoring function on pairs of documents given a bilingual query.

We find that our bilingual rankings have good monolingual ranking properties. In particular, given an optimal pairwise bilingual ranking, we show that simple heuristics can effectively approximate the optimal monolingual ranking. Using

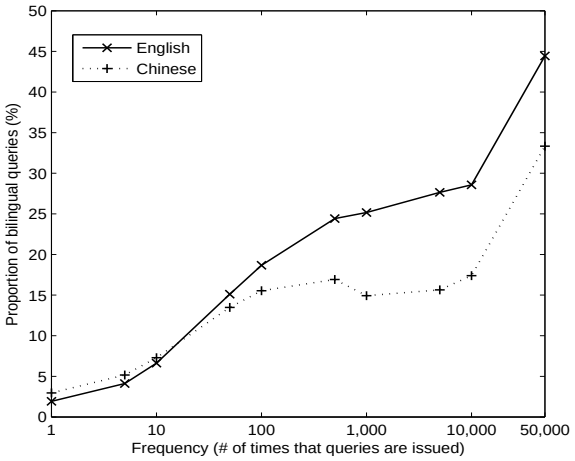

Figure 1: Proportion of bilingual queries in the query logs of different languages.

these heuristics and our learned pairwise scoring function, we can derive a ranking for new, unseen bilingual queries. We develop and test our bilingual ranker on English and Chinese with two large, publicly available query logs from the AOL search engine[1] (English query log) (Pass et al., 2006) and the Sougou search engine[2] (Chinese query log) (Liu et al., 2007). For both languages, we achieve significant improvements over monolingual Ranking SVM (RSVM) baselines (Herbrich et al., 2000; Joachims, 2002), which exploit a variety of monolingual features.

2 Bilingual Query Statistics

We designate a query as bilingual if the concept has been searched by users of both two languages. As a result, not only does it occur in the query log of its own language, but its translation also appears in the log of the second language. So a bilingual query yields reasonable queries in both languages. Of course, most queries are not bilingual. For example, our English log contains *map of Alabama*, but not our Chinese log. In this case, we wouldn't expect the Chinese results for the query's translation, 阿拉巴马地图, to be helpful in ranking the English results.

In total, we extracted 4.8 million English queries from AOL log, of which 1.3% of their translations appear in Sogou log. Similarly, of our 3.1 million Chinese queries from Sogou log, 2.3% of their translations appear in AOL log. By total number of *queries issued* (i.e., counting duplicates), the proportion of bilingual queries is much higher. As Figure 1 shows as the number of times a query is issued increases, so does the chance of it being bilingual. In particular, nearly 45% of the highest-frequency English queries and 35% of the highest-frequency Chinese queries are bilingual.

[1] http://search.aol.com
[2] http://www.sogou.com

3 Learning to Rank Using Bilingual Information

Given a set of bilingual queries, we now describe how to learn a ranking function for monolingual data that exploits information from both languages. Our procedure has three steps: Given two monolingual rankings, we construct a *bilingual* ranking on pairs of documents, one from each language. Then we learn a linear scoring function for pairs of documents that exploits monolingual information (in both languages) and bilingual information. Finally, given this ranking function on pairs and a new bilingual query, we reconstruct a monolingual ranking for the language of interest. This section addresses these steps in turn.

3.1 Creating Bilingual Training Data

Without loss of generality, suppose we rank English documents with constraints from Chinese documents. Given an English log L_e and a Chinese log L_c, our ranking algorithm takes as input a bilingual query pair $q = (q_e, q_c)$ where $q_e \in L_e$ and $q_c \in L_c$, a set of returned English documents $\{e_i\}_{i=1}^{N}$ from q_e, and a set of constraint Chinese documents $\{c_j\}_{j=1}^{n}$ from q_c. In order to create bilingual ranking data, we first generate monolingual ranking data from clickthrough statistics. For each language-query-document triple, we calculate the aggregated click count across all users and rank documents according to this statistic. We denote the count of a page as $C(e_i)$ or $C(c_j)$.

The use of clickthrough statistics as feedback for learning ranking functions is not without controversy, but recent empirical results on large data sets suggest that the aggregated user clicks provides an informative indicator of relevance preference for a query. Joachims et al. (2007) showed that *relative* feedback signals generated from clicks correspond well with human judgments. Dou et al. (2008) revealed that a straightforward use of *aggregated clicks* can achieve a better ranking than using explicitly labeled data because clickthrough data contain fine-grained differences between documents useful for learning an

Table 1: Clickthrough data of a bilingual query pair extracted from query logs.

Bilingual query pair (***Mazda***, 马自达)		
doc	URL	click #
e1	www.mazda.com	229
e2	www.mazdausa.com	185
e3	www.mazda.co.uk	5
e4	www.starmazda.com	2
e5	www.mazdamotosports.com	2
	.....	
c1	www.faw-mazda.com	50
c2	price.pcauto.com.cn/brand.jsp?bid=17	43
c3	auto.sina.com.cn/salon/FORD/MAZDA.shtml	20
c4	car.autohome.com.cn/brand/119/	18
c5	jsp.auto.sohu.com/view/brand-bid-263.html	9
	.....	

accurate and reliable ranking. Therefore, we leverage aggregated clicks for comparing the relevance order of documents. Note that there is nothing specific to our technique that requires clickthrough statistics. Indeed, our methods could easily be employed with human annotated data. Table 1 gives an example of a bilingual query pair and the aggregated click count of each result page.

Given two monolingual documents, a preference order can be inferred if one document is clicked more often than another. To allow for cross-lingual information, we extend the order of individual documents into that of *bilingual document pairs*: given two bilingual document pairs, we will write $\left(e_i^{(1)}, c_j^{(1)}\right) \succ \left(e_i^{(2)}, c_j^{(2)}\right)$ to indicate that the pair of $\left(e_i^{(1)}, c_j^{(1)}\right)$ is ranked higher than the pair of $\left(e_i^{(2)}, c_j^{(2)}\right)$.

Definition 1 $\left(e_i^{(1)}, c_j^{(1)}\right) \succ \left(e_i^{(2)}, c_j^{(2)}\right)$ *if and only if one of the following relations hold:*

1. $C(e_i^{(1)}) > C(e_i^{(2)})$ *and* $C(c_j^{(1)}) \geq C(c_j^{(2)})$
2. $C(e_i^{(1)}) \geq C(e_i^{(2)})$ *and* $C(c_j^{(1)}) > C(c_j^{(2)})$

Note, however, that from a purely monolingual perspective, this definition introduces orderings on documents that should not initially have existed. For English ranking, for example, we may have $\left(e_i^{(1)}, c_j^{(1)}\right) \succ \left(e_i^{(2)}, c_j^{(2)}\right)$ even when $C(e_i^{(1)}) = C(e_i^{(2)})$. This leads us to the following asymmetric definition of $\succ$ that we use in practice:

Definition 2 $\left(e_i^{(1)}, c_j^{(1)}\right) \succ \left(e_i^{(2)}, c_j^{(2)}\right)$ *if and only if* $C(e_i^{(1)}) > C(e_i^{(2)})$ *and* $C(c_j^{(1)}) \geq C(c_j^{(2)})$

With this definition, we can unambiguously compare the relevance of bilingual document pairs based on the order of monolingual documents. The advantages are two-fold: (1) we can treat multiple cross-lingual document similarities the same way as the commonly used query-document features in a uniform manner of learning; (2) with the similarities, the relevance estimation on bilingual document pairs can be enhanced, and this in return can improve the ranking of documents.

3.2 Ranking Model

Given a pair of bilingual queries (q_e, q_c), we can extract the set of corresponding bilingual document pairs and their click counts $\{(e_i, c_j), (C(e_i), C(c_j))\}$, where $i = 1, \ldots, N$ and $j = 1, \ldots, n$. Based on that, we produce a set of bilingual ranking instances $S = \{\mathbf{\Phi_{ij}}, z_{ij}\}$, where each $\mathbf{\Phi_{ij}} = \{\mathbf{x_i}; \mathbf{y_j}; \mathbf{s_{ij}}\}$ is the feature vector of (e_i, c_j) consisting of three components: $\mathbf{x_i} = \mathbf{f}(q_e, e_i)$ is the vector of monolingual relevancy features of e_i, $\mathbf{y_i} = \mathbf{f}(q_c, c_j)$ is the vector of monolingual relevancy features of c_j, and $\mathbf{s_{ij}} = \mathbf{sim}(e_i, c_j)$ is the vector of cross-lingual similarities between e_i and c_j, and $z_{ij} = (C(e_i), C(c_j))$ is the corresponding click counts.

The task is to select the optimal function that minimizes a given loss with respect to the order of ranked bilingual document pairs and the gold. We resort to Ranking SVM (RSVM) (Herbrich et al., 2000; Joachims, 2002) learning for classification on pairs of instances. Compared the baseline RSVM (monolingual), our algorithm learns to classify on *pairs of bilingual document pairs* rather than on pairs of individual documents.

Let f being a linear function:

$$f_{\vec{w}}(e_i, c_j) = \vec{w}_x \cdot \mathbf{x_i} + \vec{w}_y \cdot \mathbf{y_j} + \vec{w}_s \cdot \mathbf{s_{ij}} \quad (1)$$

where $\vec{w} = \{\vec{w}_x; \vec{w}_y; \vec{w}_s\}$ denotes the weight vector, in which the elements correspond to the relevancy features and similarities. For any two bilingual document pairs, their preference relation is measured by the difference of the functional values of Equation 1:

$$\begin{aligned}
&\left(e_i^{(1)}, c_j^{(1)}\right) \succ \left(e_i^{(2)}, c_j^{(2)}\right) && \Leftrightarrow \\
&f_{\vec{w}}\left(e_i^{(1)}, c_j^{(1)}\right) - f_{\vec{w}}\left(e_i^{(2)}, c_j^{(2)}\right) > 0 && \Leftrightarrow \\
&\vec{w}_x \cdot \left(\mathbf{x_i^{(1)}} - \mathbf{x_i^{(2)}}\right) + \vec{w}_y \cdot \left(\mathbf{y_j^{(1)}} - \mathbf{y_j^{(2)}}\right) + \\
&\vec{w}_s \cdot \left(\mathbf{s_{ij}^{(1)}} - \mathbf{s_{ij}^{(2)}}\right) > 0
\end{aligned}$$

We then create a new training corpus based on the preference ordering of any two such pairs: $S' = \{\mathbf{\Phi'_{ij}}, z'_{ij}\}$, where the new feature vector becomes

$$\mathbf{\Phi'_{ij}} = \left\{\mathbf{x_i^{(1)}} - \mathbf{x_i^{(2)}}; \mathbf{y_j^{(1)}} - \mathbf{y_j^{(2)}}; \mathbf{s_{ij}^{(1)}} - \mathbf{s_{ij}^{(2)}}\right\},$$

and the class label

$$z'_{ij} = \begin{cases} +1, & \text{if } \left(e_i^{(1)}, c_j^{(1)}\right) \succ \left(e_i^{(2)}, c_j^{(2)}\right); \\ -1, & \text{if } \left(e_i^{(2)}, c_j^{(2)}\right) \succ \left(e_i^{(1)}, c_j^{(1)}\right) \end{cases}$$

is a binary preference value depending on the order of bilingual document pairs. The problem is to solve SVM objective: $\min_{\vec{w}} \frac{1}{2}\|\vec{w}\|^2 + \lambda \sum_i \sum_j \xi_{ij}$ subject to bilingual constraints: $z'_{ij} \cdot (\vec{w} \cdot \mathbf{\Phi'_{ij}}) \geq 1 - \xi_{ij}$ and $\xi_{ij} \geq 0$.

There are potentially $\Gamma = nN$ bilingual document pairs for each query, and the number of comparable pairs may be much larger due to the combinatorial nature (but less than $\Gamma(\Gamma-1)/2$). To speed up training, we resort to stochastic gradient descent (SGD) optimizer (Shalev-Shwartz et al., 2007) to approximate the true gradient of the loss function evaluated on a single instance (i.e., per constraint). The parameters are then adjusted by an amount proportional to this approximate gradient. For large data set, SGD-RSVM can be much faster than batch-mode gradient descent.

3.3 Inference

The solution $\vec{w}$ forms a vector orthogonal to the hyper-plane of RSVM. To predict the order of bilingual document pairs, the ranking score can be simply calculated by Equation 1. However, a prominent problem is how to derive the full order of monolingual documents for output from the order of bilingual document pairs. To our knowledge, there is no precise conversion algorithm in polynomial time. We thus adopt two heuristics for approximating the true document score:

- **H-1 (max score):** Choose the maximum score of the pair as the score of document, i.e., $score(e_i) = max_j(f(e_i, c_j))$.
- **H-2 (mean score):** Average over all the scores of pairs associated with the ranked document as the score of this document, i.e., $score(e_i) = 1/n \sum_j f(e_i, c_j)$.

Intuitively, for the rank score of a single document, **H-2** combines the "voting" scores from its n constraint documents weighted equally, while **H-1** simply chooses the maximum one. A formal approach to the problem is to leverage rank aggregation formalism (Dwork et a., 2001; Liu et al., 2007), which will be left for our future work. The two simple heuristics are employed here because of their simplicity and efficiency. The time complexity of the approximation is linear to the number of ranked documents given n is constant.

4 Features and Similarities

Standard features for learning to rank include various query-document features, e.g., BM25 (Robertson, 1997), as well as query-independent features, e.g., PageRank (Brin and Page, 1998). Our feature space consists of both these standard monolingual features and cross-lingual similarities among documents. The cross-lingual similarities are valuated using different translation mechanisms, e.g., dictionary-based translation or machine translation, or even without any translation at all.

4.1 Monolingual Relevancy Features

In learning to rank, the relevancy between query and documents and the measures based on link analysis are commonly used as features. The discussion on their details is beyond the scope of this paper. Readers may refer to (Liu et al., 2007) for the definitions of many such features. We implement six of these features that are considered the most typical shown as Table 2. These include sets of measures such as BM25, language-model-based IR score, and PageRank. Because most conventional IR and web search relevancy measures fall into this category, we call them altogether *IR features* in what follows. Note that for a given bilingual document pair (e, c), the monolingual IR features consist of relevance score vectors $\mathbf{f}(q_e, e)$ in English and $\mathbf{f}(q_c, c)$ in Chinese.

4.2 Cross-lingual Document Similarities

To measure the document similarity across different languages, we define the similarity vector $\mathbf{sim}(e, c)$ as a series of functions mapping a bilingual document pair to positive real numbers. Intuitively, a good similarity function is one which maps cross-lingual relevant documents into close scores and maintains a large distance between irrelevant and relevant documents. Four categories of similarity measures are employed.

Dictionary-based Similarity (DIC): For dictionary-based document translation, we use

Table 2: List of monolingual relevancy measures used as IR features in our model.

IR Feature	Description
BM25	Okapi BM25 score (Robertson, 1997)
BM25_PRF	Okapi BM25 score with pseudo-relevance feedback (Robertson and Jones, 1976)
LM_DIR	Language-model-based IR score with Dirichlet smoothing (Zhai and Lafferty, 2001)
LM_JM	Language-model-based IR score with Jelinek-Mercer smoothing (Zhai and Lafferty, 2001)
LM_ABS	Language-model-based IR score with absolute discounting (Zhai and Lafferty, 2001)
PageRank	PageRank score (Brin and Page, 1998)

the similarity measure proposed by Mathieu et al. (2004). Given a bilingual dictionary, we let $T(e, c)$ denote the set of word pairs (w_e, w_c) such that w_e is a word in English document e, and w_c is a word in Chinese document c, and w_e is the English translation of w_c. We define $tf(w_e, e)$ and $tf(w_c, c)$ to be the term frequency of w_e in e and that of w_c in c, respectively. Let $df(w_e)$ and $df(w_c)$ be the English document frequency for w_e and Chinese document frequency for w_c. If n_e (n_c) is the total number of English (Chinese), then the bilingual idf is defined as $idf(w_e, w_c) = \log \frac{n_e+n_c}{df(w_e)+df(w_c)}$. Then the cross-lingual document similarity is calculated by

$$sim(e, c) = \frac{\sum_{(w_e,w_c)\in T(e,c)} tf(w_e,e)tf(w_c,c)idf(w_e,w_c)^2}{\sqrt{Z}}$$

where Z is a normalization coefficient (see Mathieu et al. (2004) for detail). This similarity function can be understood as the cross-lingual counterpart of the monolingual cosine similarity function (Salton, 1998).

Similarity Based on Machine Translation (MT): For machine translation, the cross-lingual measure actually becomes a monolingual similarity between one document and another's translation. We therefore adopt cosine function for it directly (Salton, 1998).

Translation Ratio (RATIO): Translation ratio is defined as two sets of ratios of translatable terms using a bilingual dictionary: RATIO_FOR – what percent of words in e can be translated to words in c; RATIO_BACK – what percent of words in c can be translated back to words in e.

URL LCS Ratio (URL): The ratio of longest common subsequence (Cormen et al., 2001) between the URLs of two pages being compared. This measure is useful to capture pages in different languages but with similar URLs such as `www.airbus.com`, `www.airbus.com.cn`, etc.

Note that each set of similarities above except URL includes 3 values based on different fields of web page: title, body, and title+body.

5 Experiments and Results

This section presents evaluation metric, data sets and experiments for our proposed ranker.

5.1 Evaluation Metric

Commonly adopted metrics for ranking, such as mean average precision (Buckley and Voorhees, 2000) and Normalized Discounted Cumulative Gain (Järvelin and Kekäläinen, 2000), is designed for data sets with human relevance judgment, which is not available to us. Therefore, we use the Kendall's tau coefficient (Kendall, 1938; Joachims, 2002) to measure the degree of correlation between two rankings. For simplicity, let's assume strict orderings of any given ranking. Therefore we ignore all the pairs with ties (instances with the identical click count). Kendall's tau is defined as $\tau(r_a, r_b) = (P - Q)/(P + Q)$, where P is the number of concordant pairs and Q is the number of disconcordant pairs in the given orderings r_a and r_b. The value is a real number within $[-1, +1]$, where -1 indicates a complete inversion, and $+1$ stands for perfect agreement, and a value of zero indicates no correlation.

Existing ranking techniques heavily depend on human relevance judgment that is very costly to obtain. Similar to Dou et al (2008), our method utilizes the automatically aggregated click count in query logs as the gold for deriving the true order of relevancy, but we use the clickthrough of different languages. We average Kendall's tau values between the algorithm output and the gold based on click frequency for all test queries.

5.2 Data Sets

Query logs can be the basis for constructing high quality ranking corpus. Due to the proprietary issue of log, no public ranking corpus based on real-world search engine log is currently available. Moreover, to build a predictable bilingual ranking corpus, the logs of different languages are needed and have to meet certain conditions: (1) they should be sufficiently large so that a good number of bilingual query pairs could be identi-

Table 3: Statistics on AOL and Sogou query logs.

	AOL(EN)	**Sogou(CH)**
# sessions	657,426	5,131,000
# unique queries	10,154,743	3,117,902
# clicked queries	4,811,650	3,117,590
# clicked URLs	1,632,788	8,627,174
time span	2006/03-05	2006/08
size	2.12GB	1.56GB

fied; (2) for the identified query pairs, there should be sufficient statistics of associated clickthrough data; (3) The click frequency should be well distributed at both sides so that the preference order between bilingual document pairs can be derived for SVM learning.

For these reasons, we use two independent and publicly accessible query logs to construct our bilingual ranking corpus: English AOL log[3] and Chinese Sogou log[4]. Table 3 shows some statistics of these two large query logs.

We automatically identify 10,544 bilingual query pairs from the two logs using the Java API for Google Translate[5], in which each query has certain number of clicked URLs. To better control the bilingual equivalency of queries, we make sure the bilingual queries in each of these pairs are bi-directional translations. Then we download all their clicked pages, which results in 70,180 English[6] and 111,197 Chinese documents. These documents form two independent collections, which are indexed separately for retrieval and feature calculation.

For good quality, it is necessary to have sufficient clickthrough data for each query. So we further identify 1,084 out of 10,544 bilingual query pairs, in which each query has at least 10 clicked and downloadable documents. This smaller collection is used for learning our model, which contains 21,711 English and 28,578 Chinese documents[7]. In order to compute cross-lingual document similarities based on machine translation (see Section 4.2), we automatically translate all these 50,298 documents using Google Translate, i.e., English to Chinese and vice versa. Then the bilingual document pairs are constructed, and all the monolingual features and cross-lingual similarities are computed (see Section 4.1&4.2).

[3] http://gregsadetsky.com/aol-data/

[4] http://www.sogou.com/labs/dl/q.html

[5] http://code.google.com/p/google-api-translate-java/

[6] AOL log only records the domain portion of the clicked URLs, which misleads document downloading. We use the "search within site or domain" function of a major search engine to approximate the real clicked URLs by keeping the first returned result for each query.

[7] Because Sogou log has a lot more clicked URLs, for balancing with the number of English pages, we kept at most 50 pages per Chinese query.

Table 4: Kendall's tau values of English ranking. The significant improvements over baseline (99% confidence) are bolded with the p-values given in parenthesis. * indicates significant improvement over IR (no similarity). $n = 5$.

Models	**Pair**	**H-1 (max)**	**H-2 (mean)**
RSVM (baseline)	n/a	0.2424	0.2424
IR (no similarity)	0.2783	0.2445	0.2445
IR+DIC	0.2909	0.2453	0.2496
IR+MT	0.2858	**0.2488*** (p=0.0003)	**0.2494*** (p=0.0004)
IR+DIC+MT	0.2901	0.2481	**0.2514*** (p=0.0009)
IR+DIC+RATIO	0.2946	0.2466	**0.2519*** (p=0.0004)
IR+DIC+MT +RATIO	0.2940	**0.2473*** (p=0.0009)	**0.2539*** (p=1.5e-5)
IR+DIC+MT +RATIO+URL	0.2979	**0.2533*** (p=2.2e-5)	**0.2577*** (p=4.4e-7)

5.3 English Ranking Performance

Here we examine the ranking performance of our English ranker under different similarity settings. We use traditional RSVM (Herbrich et al., 2000; Joachims, 2002) without any bilingual consideration as the *baseline*, which uses only English IR features. We conduct this experiment using all the 1,084 bilingual query pairs with 4-fold cross validation (each fold with 271 query pairs). The number of constraint documents n is empirically set as 5. The results are shown in Table 4.

Clearly, bilingual constraints are helpful to improve English ranking. Our pairwise settings unanimously outperforms the RSVM baseline. The paired two-tailed t-test (Smucker et al., 2007) shows that most improvements resulted from heuristic **H-2** (mean score) are statistically significant at 99% confidence level (p<0.01). Relatively fewer significant improvements can be made by heuristic **H-1** (max score). This is because the maximum score on pair is just a rough approximation to the optimal document score. But this simple scheme works surprisingly well and still consistently outperforms the baseline.

Note that our bilingual model with only IR features, i.e., IR (no similarity), also outperforms the baseline. The reason is that in this setting there are

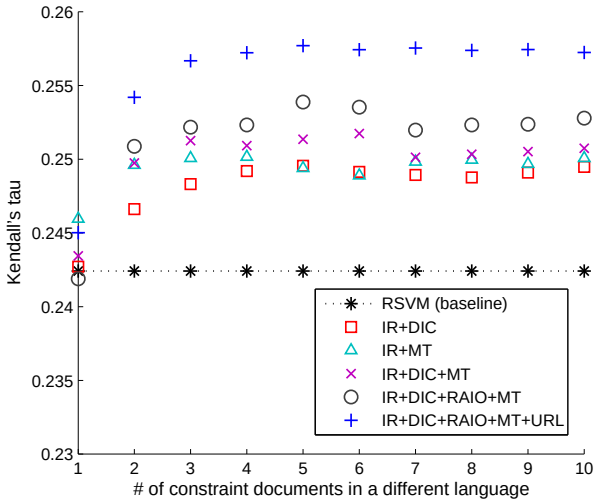

Figure 2: English ranking results vary with the number of constraint Chinese documents.

IR features of n Chinese documents introduced in addition to the IR features of English documents in the baseline.

The DIC similarity does not work as effectively as MT. This may be due to the limitation of bilingual dictionary alone for translating documents, where the issues like out-of-vocabulary words and translation ambiguity are common but can be better dealt with by MT. When DIC is combined with RATIO, which considers both forward and backward translation of words, it can capture the correlation between bilingually very similar pages, thus performs better.

We find that the URL similarity, although simple, is very useful and improves 1.5–2.4% of Kendall's tau value than not using it. This is because the URLs of the top Chinese (constraint) documents are often similar to many of returned English URLs which are generally more regular. For example, in query pair (*Toyota Camry*, 丰田佳美), 9/13 English pages are anchored by the URLs containing keywords "toyota" and/or "camry", and 3/5 constraint documents' URLs also contain them. In contrast, the URLs of returned Chinese pages are less regular in general. This also explains why this measure does not improve much for Chinese ranking (see Section 5.4).

We also vary the parameter n to study how the performance changes with different number of constraint Chinese documents. Figure 2 shows the results using heuristic **H-2**. More constraint documents are generally helpful, but when only one constraint document is used, it may be detrimental to the ranking for some features. One explanation is that the document clicked most often is not necessarily relevant, and it is very likely that no English page is similar to the first Chinese page. Joachims et al. (2007) found that users' click behavior is biased by the rank of search engine at the first and/or second positions (especially the first). More constraint pages are helpful because the pages after the first are less biased and the click counts can reflect the relevancy more accurately.

Table 5: Kendall's tau values of Chinese ranking. The significant improvements over baseline (99% confidence) are bolded with the p-values given in parenthesis. * indicates significant improvement over IR (no similarity). $n = 5$.

Models	Pair	H-1 (max)	H-2 (mean)
RSVM (baseline)	n/a	0.2935	0.2935
IR (no similarity)	0.3201	0.2938	0.2938
IR+DIC	0.3220	**0.2970** (p=0.0060)	**0.2973*** (p=0.0020)
IR+MT	0.3299	**0.2992*** (p=0.0034)	**0.3008*** (p=0.0003)
IR+DIC+MT	0.3295	**0.2991*** (p=0.0014)	**0.3004*** (p=0.0008)
IR+DIC+RATIO	0.3240	**0.2972*** (p=0.0010)	**0.2968*** (p=0.0014)
IR+DIC+MT +RATIO	0.3303	**0.2973*** (p=0.0004)	**0.3007*** (p=0.0002)
IR+DIC+MT +RATIO+URL	0.3288	**0.2981*** (p=0.0005)	**0.3024*** (p=1.5e-6)

5.4 Chinese Ranking Performance

We also benchmark Chinese ranking with English constraint documents under the similar configurations as Section 5.3. The results are given by Table 5 and Figure 3.

As shown in Table 5, improvements on Chinese ranking are even more encouraging. Kendall's tau values under all the settings are significantly better than not only the baseline but also IR (no similarity). This may suggest that English information is generally more helpful to Chinese ranking than the other way round. The reason is straightforward: there are a high proportion of Chinese queries having English or foreign-language origins in our data set. For these queries, relevant information at Chinese side may be relatively poorer, so the English ranking can be more reliable. As far as we can, we manually identified 215 such queries from all the 1,084 bilingual queries (amount to 23.2%).

To shed more light on this finding, we examine top-20 queries improved most by our method

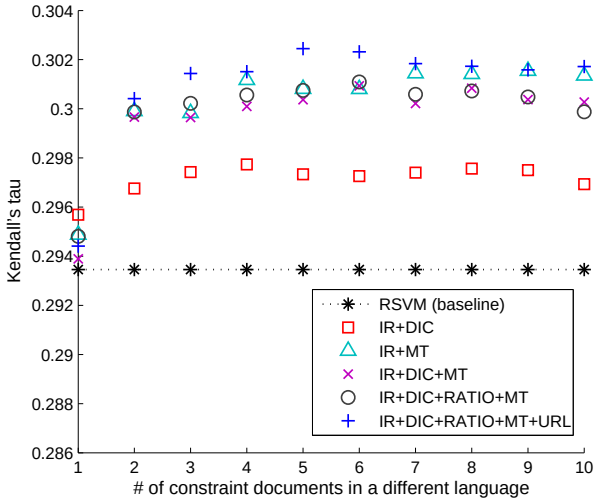

Figure 3: Chinese ranking results vary with the number of constraint English documents.

(with all features and similarities) over the baseline. As shown in Table 6, most of the top improved Chinese queries are about concepts originated from English or other languages, or something non-local (bolded). Interestingly, 政治漫画 (*political catoons*) are among these Chinese queries improved most by English ranking, which is believed as rare (or sensitive) content on Chinese web. In contrast, top English queries are short of this type of queries. But we can still see *Bruce Lee* (李小龙), a Chinese Kung-Fu actor, and *peony* (牡丹), the national flower of China. Their information tends to be more popular on Chinese web, and thus helpful to English ranking. For the exceptions like *Sunrider* (仙妮蕾德) and *Aniston* (安妮斯顿), despite their English origins, we find they have surprisingly sparse click counts in English log while Chinese users look much more interested and provide a lot of clickthrough that is helpful.

6 Conclusions and Future Work

We aim to improve web search ranking for an important set of queries, called bilingual queries, by exploiting bilingual information derived from clickthrough logs of different languages. The thrust of our technique is using search ranking of one language and cross-lingual information to help ranking of another language. Our pairwise ranking scheme based on bilingual document pairs can easily integrate all kinds of similarities into the existing framework and significantly improves both English and Chinese ranking performance.

Table 6: Top 20 most improved bilingual queries. Bold means a positive example for our hypothesis. * marks an exception.

Most improved CH queries	Most improved EN queries
沙门氏菌 (salmonella)	free online tv (免费在线电视)
苏格兰 (scotland)	weapons (武器)
咖啡因 (caffeine)	lily (百合)
墓志铭 (epitaph)	cable (电缆)
英国历史 (british history)	*sunrider (仙妮蕾德)
政治漫画 (political cartoons)	*aniston (安妮斯顿)
免疫系统 (immune system)	clothes (衣服)
葡萄酒瓶 (wine bottles)	*three little pigs (三只小猪)
匈牙利 (hungary)	hair care (护发)
巫术 (witchcraft)	neon (霓虹灯)
爆米花 (popcorn)	**bruce lee (李小龙)**
脓疱疮 (impetigo)	radish (萝卜)
卫生间设计 (bathroom design)	chile (智利)
鸽子 (pigeon)	**peony (牡丹)**
北极熊 (polar bear)	toothache (牙痛)
非洲地图 (map of africa)	free online translation (免费在线翻译)
拉布拉多犬 (labrador retriever)	water (水)
帕米拉安德森 (pamela anderson)	oil (石油)
瑜伽服装 (yoga clothing)	shopping network (购物网)
联邦快递 (federal express)	*prince harry (哈里王子)

Our model can be generally applied to other search ranking problems, such as ranking using monolingual similarities or ranking for cross-lingual/multilingual web search. Another interesting direction is to study the recovery of the optimal document ordering from pairwise ordering using well-founded formalism such as rank aggregation approaches (Dwork et a., 2001; Liu et al., 2007).

Furthermore, we may involve more sophisticated monolingual features that do not transfer cross-lingually but are asymmetric for either side, such as clustering, document classification features built from domain taxonomies like DMOZ.

Acknowledgments

This work is partially supported by the Innovation Technology Fund, Hong Kong (project No.: ITS/182/08). We would like to thank Cheng Niu for the insightful advice and anonymous reviewers for the useful comments.

References

Sergey Brin and Lawrence Page. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of WWW*.

Chris Buckley and Ellen M. Voorhees. 2000. Evaluating Evaluation Measure Stability. In *Proceedings of ACM SIGIR*, pp. 33-40.

David Burkett and Dan Klein. 2008. Two Languages are Better than One (for Syntactic Parsing). In *Proceedings of EMNLP*, pp. 877-886.

Ming-Wei Chang, Dan Goldwasser Dan Roth and Yuancheng Tu. 2009. Unsupervised Constraint Driven Learning for Transliteration Discovery. In *Proceedings of NAACL-HLT*.

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest and Clifford Stein. 2001. *Introduction to Algorithms (2nd Edition)*, MIT Press, pp. 350-355.

Zhicheng Dou, Ruihua Song, Xiaojie Yuan and Ji-Rong Wen. 2008. Are Click-through Data Adequate for Learning Web Search Rankings? In *Proceedings of ACM CIKM*, pp. 73-82.

Cynthia Dwork, Ravi Kumar, Moni Naor and D. Sivakumar. 2001. Rank Aggregation Methods for the Web. In *Proceedings of WWW*, pp. 613-622.

Ralf Herbrich, Thore Graepel and Klaus Obermayer. 2000. Large Margin Rank Boundaries for Ordinal Regression. *Advances in Large Margin Classifiers*, The MIT Press, pp. 115-132.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak 2005. Bootstrapping Parsers via Syntactic Projection across Parallel Texts. *Natural Language Engineering*, 11(3):311-325.

Kalervo Järvelin and Jaana Kekäläinen. 2000. IR Evaluation Methods for Retrieving Highly Relevant Documents. In *Proceedings of ACM SIGIR*, pp. 41-48.

Thorsten Joachims. 2002. Optimizing Search Engines Using Clickthrough Data. In *Proceedings of ACM SIGKDD*, pp. 133-142.

Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski and Geri Gay 2007. Evaluating the Accuracy of Implicit Feedback from Clicks and Query Reformulations in Web Search. *ACM Transaction on Information Systems*, 25(2):7.

M. Kendall. 1938. A New Measure of Rank Correlation. *Biometrika*, 30:81-89.

Victor Lavrenko, Martin Choquette and Bruce W. Croft. 2002. Cross-Lingual Relevance Models. In *Proceedings of ACM SIGIR*, pp. 175-182.

Tie-Yan Liu, Jun Xu, Tao Qin, Wenying Xiong, and Hang Li. 2007. LECTOR: Benchmark Dataset for Research on Learning to Rank for Information Retrieval. In *Proceedings of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval*, pp. 3-10, Amsterdam, The Netherland.

Yiqun Liu, Yupeng Fu, Min Zhang, Shaoping Ma and Liyun Ru. 2007. Automatic Search Engine Performance Evaluation with Click-through Data Analysis. In *Proceedings of WWW*, pp. 1133-1134.

Yu-Ting Liu, Tie-Yan Liu, Tao Qin, Zhi-Ming Ma, and Hang Li. 2007. Supervised Rank Aggregation. In *Proceedings of WWW*, pp. 481-489.

Benoit Mathieu, Romanic Besancon and Christian Fluhr. 2004. Multilingual Document Clusters Discovery. In *proceedings of Recherche d'Information Assistée par Ordinateur (RIAO)*, pp. 1-10.

Greg Pass, Abdur Chowdhury and Cayley Torgeson. 2006. A Picture of Search. In *Proceedings of the 1st International Conference on Scalable Information Systems (INFOSCALE)*, Hong Kong.

S. E. Robertson. 1997. Overview of the OKAPI Projects. *Journal of Documentation*, 53(1):3-7.

S. E. Robertson and K. Sparc Jones. 1976. Relevance Weighting of Search Terms. *Journal of the American Society of Information Science*, 27(3):129-146.

Gerard Salton. 1998. *Automatic Text Processing*. Addison-Wesley Publishing Company.

Shai Shalev-Shwartz, Yoram Singer and Nathan Srebro. 2007. Pegasos: Primal Estimated sub-GrAdient SOlver for SVM. In *Proceedings of ICML*, pp. 807-814.

David A. Smith and Noah A. Smith. 2004. Bilingual Parsing with Factored Estimation: Using English to Parse Korean. In *Proceedings of EMNLP*.

Mark D. Smucker, James Allan, and Ben Carterette. 2007. A Comparison of Statistical Significance Tests for Information Retrieval Evaluation. In *Proceedings of ACM CIKM*, pp. 623-632.

Benjamin Snyder and Regina Barzilay. 2008. Unsupervised Multilingual Learning for Morphological Segmentation. In *Proceedings of ACL*, pp. 737-745.

Yejun Wu and Douglas W. Oard. 2008. Bilingual Topic Aspect Classification with a Few Training Examples. In *Proceedings of ACM SIGIR*, pp. 203-210.

Chengxiang Zhai and John Lafferty. 2001. A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. In *Proceedings of ACM SIGIR*, pp. 334-342.