

# Fractional Similarity: Cross-lingual Feature Selection for Search

Jagadeesh Jagarlamudi<sup>1</sup> and Paul N. Bennett<sup>2</sup>

<sup>1</sup> University of Maryland, Computer Science, College Park MD 20742, USA,  
jags@umiacs.umd.edu

<sup>2</sup> Microsoft Research, One Microsoft Way, Redmond WA 98052, USA,  
paul.n.bennett@microsoft.com

**Abstract.** Training data as well as supplementary data such as usage-based click behavior may abound in one search market (*i.e.*, a particular region, domain, or language) and be much scarcer in another market. Transfer methods attempt to improve performance in these resource-scarce markets by leveraging data across markets. However, differences in feature distributions across markets can change the optimal model. We introduce a method called Fractional Similarity, which uses query-based variance *within* a market to obtain more reliable estimates of feature deviations *across* markets. An empirical analysis demonstrates that using this scoring method as a feature selection criterion in cross-lingual transfer improves relevance ranking in the foreign language and compares favorably to a baseline based on KL divergence.

## 1 Introduction

Recent approaches [1–4] pose ranking search results as a machine learning problem by representing each query-document pair as a vector of features with an associated graded relevance label. The advantage of this approach is that a variety of heterogeneous features – such as traditional IR content-based features like BM25F [5], web graph based features like PageRank, and user behavioral data like aggregated query-click information [6] – can be fed into the system along with relevance judgments, and the ranker will learn an appropriate model. While ideally this learned model could be applied in any market (*i.e.*, a different region, domain, or language) that implements the input features, in practice there can be many challenges to *transferring* learned models.

For example, user behavior-based features, which have been shown to be very helpful in search accuracy [7, 6, 8, 9], may vary in the amount of signal they carry across markets. This may occur when more behavioral data is available in one market than another – either because the number of users varies across markets or simply because of a difference in the amount of time a search engine has been available in those markets. Furthermore, the distributions of content-based features as well as their significance for relevance may vary across languages because of differences in parsing (*e.g.*, dealing with compound words, inflection, tokenization). In addition, the amount of labeled data available to use in model

transfer – a key ingredient in model performance – can vary across markets since acquiring the graded relevance labeled data can be costly. Thus, in order to facilitate the use of labeled data from different markets, we would like an automatic method to identify the commonalities and differences across the markets.

In this paper, we address the problem of using training data from one market (*e.g.*, English) to improve the search performance in a foreign language market (*e.g.*, German) by automatically identifying the ranker’s input features that deviate significantly across markets. Different flavors of this problem have been attempted by other researchers [10–14]. At a broader level, there are two possible approaches to transfer the useful information across languages depending on the availability of original queries in both the languages. When the queries are available we can identify a pair of queries (*e.g.*, “dresses” in English and “Kleider” in German) with the same intent and transfer the query level knowledge across markets [10]. Such an approach uses only aligned queries and discards many English queries which doesn’t have translation in German. In this paper, we devise a general approach that doesn’t rely on the availability of aligned queries and instead uses *only* the information available in the feature vector in all query-document pairs.

We take a machine learning approach and pose the knowledge transfer across languages as a feature selection problem. Specifically, we aim to identify features which have similar distribution across languages and hence their data from English can be used in training the foreign language ranker. We propose a technique called Fractional Similarity (Sec. 4.2) to identify a feature’s similarity across English and foreign language training data. This method addresses two shortcomings of statistical tests in the IR setting (Sec. 3). First, as has been noted elsewhere [15], variance in an observed statistic over an IR collection is primarily dependent on differences in the query set. This mismatch, which we refer to as query-set variance, is especially problematic in the transfer setting since the query sets from the two markets are generally different – often markedly so. Second, the document feature vectors for a specific query are often correlated (at least in some features), this query-block correlation can exacerbate the first problem by causing large shifts in a statistic for a feature based on whether the query is considered or not. Fractional Similarity addresses these shortcomings (Sec. 4.2) by adapting statistical tests to a query-centric sampling setting and thus enables a fair comparison between different features. In Sec. 5 we describe how to use Fractional Similarity in IR scenario to rank all the features based on their similarity value providing a way for feature selection. In our experiments (Sec. 6), we found that a simple approach of dropping the mismatched features from English data and using the rest of the training data helps in improving the search accuracy in the foreign language.

## 2 Related Work

Here we briefly describe the most relevant work before delving further into the problem. Gao *et al.* [10] uses English web search results to improve the rank-

ing of non-ambiguous Chinese queries (referred to as Linguistically Non-local queries). Other research [11, 12] uses English as an assisting language to provide pseudo-relevant terms for queries in different languages. The generality of these approaches is limited either by the type of queries or in the setting (traditional TREC style) they are explored. In [13], English training data from a general domain has been used to improve the accuracy of the English queries from a Korean market. But in this particular case both in-domain and out-of-domain data are from English, hence the set of features used for the learning algorithm remain same. Here we study the problem in a general scenario by mapping it as a transfer learning problem. The main aim is to use the large amounts of training data available in English along with the foreign language data to improve the search performance in the foreign language.

In terms of tests to measure deviation of feature distributions, there have been some measures proposed in domain adaptation literature [16, 17] to compute distance between the source and target distributions of a feature. But they are primarily intended to give theoretical bounds on the performance of the adapted classifier and are also not suitable to IR – both because they do not adequately handle query-set variance and query-block correlation and for the reasons mentioned in Sec. 3.

In addition to addressing both the IR specific issues, our method is more efficient because it only involves off-line computations, computing feature deviations across search markets and training the ranker with the additional English language data, and doesn't involve any computation during the actual querying.

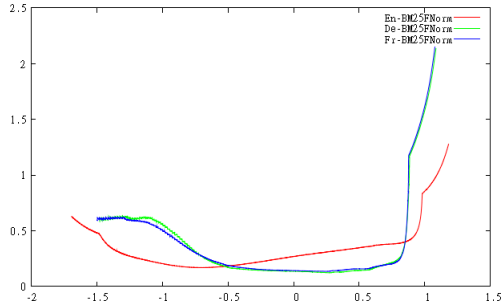
### 3 Challenges in Computing Feature Similarity

In this section we describe some of the challenges in computing the similarity of feature distributions across languages. While some of these challenges are specific to Information Retrieval the rest of them exist in other scenarios as well.

Note that in general, determining whether a feature distribution is similar across markets can be a challenging problem. For example, Figure 1 depicts the distributions of BM25F [5] in English, French, and German language training data (normalization was done to account for trivial differences).<sup>3</sup> From the figure, it is clear that the distribution of BM25F scores in French and German languages bear a closer resemblance than to that of English, but is this difference significant enough to negatively impact transfer learning or is it a minor difference that additional normalization might correct? Even if manual inspection was straightforward, a formal method of characterizing the deviation would be advantageous in many settings.

While statistical tests to measure divergence in distributions exist, they are ill-suited to this problem for two reasons stemming from query-based variance. The first of these effects on variance results from the query-block correlation that occurs among related documents. That is, in the IR learning to rank setting,

<sup>3</sup> We refer to this type of probability density function (pdf) as the feature distribution of a feature in the rest of the paper.



**Fig. 1.** Feature distribution of BM25F in English, German, and French language data. The x-axis denotes the normalized feature value and the y-axis denotes the probability of the feature estimated using Kernel probability estimation technique.

we typically have a training/test instance for each query-document pair, and for each query, a block of instances exist corresponding to differing documents relationship to the query. Because these documents are often related to the query in some way (*e.g.* top 100) that made them a candidate for final ranking, their features will often be correlated. For example, consider a slightly exaggerated case for a feature, “Number\_of\_query\_words\_found\_in\_document”. By the very nature of the documents being initial matches, all of them are likely to have similar values (probably near the query length). While for a different query, the same feature takes a different value but also highly correlated within the query. While the reader may think that a proper normalization (*e.g.*, normalization by query length) can alleviate this problem, we argue that this problem is more general both in this case and also may occur for other features such as those based on link analysis or behavioral data where there is no obvious normalization. The net effect is that even with in a language, statistics such as the mean or variance of a feature can often shift considerably based on the inclusion/exclusion of a small set of queries and their associated blocks.

A related challenge is that of query-set variance. That is, when comparing two different query sets, a feature can appear superficially different ultimately because the queries are different and not because how the feature characterizes the query-document relationship is different. Within a single language this can be a formidable challenge of its own and arises in large part because of query-block correlation, but across languages, this problem can be even worse since the query sets can differ even more (*e.g.* the most common queries, sites, *etc.* may differ across languages, cultures, regions).

As a result, a reasonable method should consider the impact of query-block correlation and query-set variance within the market whose data is to be transferred and use that as a reference to determine what deviations in the foreign language are actually significant. This is exactly what our method does. We now move on to the formal derivation of our approach.

## 4 Computing Feature Similarity

Our proposed method to compute the similarity of a feature’s distribution across languages builds on the well known T-test [18]. Thus, we start by briefly reviewing the T-test and then move on to the details of Fractional Similarity. Though we discuss it mainly in the context of T-test and continuous features, it can be extended to other relevant significance tests (*e.g.*,  $\chi^2$  test) and to discrete features as well.

### 4.1 T-test

Given two samples ( $X_1$  and  $X_2$ ) from an unknown underlying distributions, the T-test [18] can be used to verify if the means of both the samples are equal or not. Since the variance of a feature may differ across languages, we use the more general form that assumes a possibly different variance for the two random variables to be compared. Let  $\mu_1$  and  $\mu_2$  denote the means of both the samples and  $\sigma_1^2$  and  $\sigma_2^2$  denote the variances of both the samples, then the t-statistic and the degrees of freedom are given by:

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \quad \text{and} \quad d.f. = \frac{(\sigma_1^2/n_1 + \sigma_2^2/n_2)^2}{(\sigma_1^2/n_1)^2/(n_1 - 1) + (\sigma_2^2/n_2)^2/(n_2 - 1)}$$

where  $n_1$  and  $n_2$  are the respective sample sizes. Both these statistics along with the Students t-distribution [19] can be used to get the probability (referred to as the “p-value” function in our pseudo code) of observing the result under the null hypothesis that the means of the samples are equal. We would like to remind the reader that in statistical significance testing we are typically interested in showing that a new result is different from the baseline so we want the p-value to lower than the critical value. But in the case of transfer, we are interested in finding the similarity of two samples, so we want the means of both the samples to be same which means that we want p-value to be higher.

We will also need the result below, that the mean and variance of a convex combination of the above samples ( $X^\alpha = (1 - \alpha)X_1 + \alpha X_2$ ) with  $\alpha \in [0, 1]$  is given by:

$$\mu^\alpha = (1 - \alpha)\mu_1 + \alpha\mu_2 \quad \text{and} \quad \sigma_\alpha^2 = (1 - \alpha)(\sigma_1^2 + \mu_1^2) + \alpha(\sigma_2^2 + \mu_2^2) - (\mu^\alpha)^2 \quad (1)$$

We will use these expressions in Sec. 4.2 to compute the mean and variance of an arbitrary convex combination of two random samples.

### 4.2 Fractional Similarity

A direct application of the T-test to our problem would, for each feature, select random samples from both English and foreign language data sets and verify if the means of both the samples are equal or not. However, the impact of query-set variance across markets, typically yields a t-test value close to zero probability

in almost all cases rendering it practically useless to rank the features. In reality, the simple t-test indicates that the sets are composed of different queries (an obvious fact we know) and not that the relationship the feature characterizes between the documents and the queries is significantly different (what we are interested in knowing).

We use the following key idea to compute the similarity between two distributions ( $P$  and  $Q$ ). If both the given distributions are the same (*i.e.*  $P \equiv Q$ ), then, with high probability, any two random samples  $P_s$  (drawn from  $P$ ) and  $Q_s$  (drawn from  $Q$ ) are statistically indistinguishable among themselves and also from a convex combination  $((1 - \alpha)P_s + \alpha Q_s)$  of the samples. When the underlying distributions are indeed different ( $P \neq Q$ ), then for some value of  $\alpha$  the convex combination starts looking different from both the samples and we leverage on this intuition to compute the similarity. If we treat the convex combination as if we are replacing  $\alpha$  fraction<sup>4</sup> of examples from  $P_s$  with those of  $Q_s$ , then as we replace more and more instances the resulting sample starts looking different from the original sample  $P_s$ . We use this *fraction of examples* to be replaced to make it look different from  $P_s$  as indicative of the similarity between both the distributions (hence the name Fractional Similarity). The value of Fractional Similarity lies in the range 0 and 1, with a higher value indicating better similarity of the underlying distributions.

Let  $P_s^\alpha$  be the convex combination of the samples  $P_s$  and  $Q_s$ , *i.e.*,  $P_s^\alpha \leftarrow (1 - \alpha)P_s + \alpha Q_s$  and  $P_s^r$  be another random sample (superscript  $r$  stands for reference) drawn from  $P$ . Now the objective is to compare both  $P_s$  and the convex combination with the reference sample. If  $p_{pp}$  and  $p_{pq}^\alpha$  denote the p-values obtained by the T-test on pairs of samples  $(P_s, P_s^r)$  and  $(P_s^\alpha, P_s^r)$  respectively, then we propose to use the following statistic:

$$\text{frac}^\alpha = \frac{p_{pq}^\alpha}{p_{pp}} \quad (2)$$

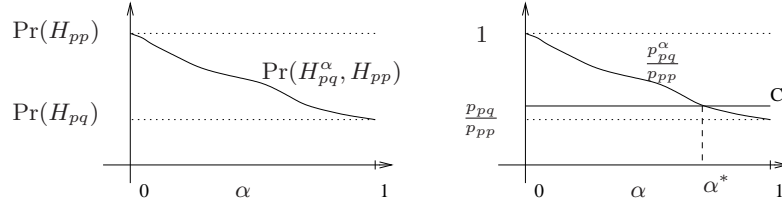
and we define Fractional Similarity as the maximum value of  $\alpha$  such that  $\text{frac}^\alpha$  is greater than a critical value ( $C$ ), *i.e.* Fractional Similarity =  $\arg \max_\alpha \text{frac}^\alpha > C$ . Our statistic explicitly takes within language variability into account, via the denominator of Eqn. 2, and as result it can be understood as a normalized version of across language variability.

Formally, the proposed statistic is inspired by the following observation. Let  $H_{pp}$  denote the event that the means of both the samples  $P_s$  and  $P_s^r$  are equal and also  $H_{pq}^\alpha$  denote the event that the means of  $P_s^\alpha$  and  $P_s^r$  are equal. Since we want to account for the variance of a feature in its original distribution, we assume the truth of  $H_{pp}$  event and find the probability of the event  $H_{pq}^\alpha$ . That is, we are interested in the following quantity:

$$\Pr(H_{pq}^\alpha | H_{pp}) = \frac{\Pr(H_{pp}, H_{pq}^\alpha)}{\Pr(H_{pp})} = \frac{\Pr(H_{pp} | H_{pq}^\alpha) \Pr(H_{pq}^\alpha)}{\Pr(H_{pp})} \quad (3)$$

---

<sup>4</sup> In our further notation, we use  $\alpha$  as a superscript when indicating an  $\alpha$  combined sample.



**Fig. 2.** When  $\alpha = 0$ ,  $P_s^\alpha$  becomes  $P_s$  as a result the joint probability (left image) becomes  $\Pr(H_{pp})$  and hence the  $\text{frac}^\alpha$  (right image) becomes 1. And as  $\alpha$  approaches 1, we introduce more and more noisy instances and eventually the joint probability reduces to  $\Pr(H_{pq})$  and  $\text{frac}^\alpha$  reaches its minimum value of  $\frac{p_{pq}}{p_{pp}}$ .

---

**Algorithm 1** FractionalSimilarity( $P_s^r, P_s, Q_s, C$ )

---

```

1:  $p_{pp} \leftarrow \text{p-value}(P_s^r, P_s)$ . //With in language variability
2:  $\text{frac} \leftarrow \frac{1.0}{p_{pp}}$ 
3: if  $\text{frac} \leq C$  then
4:   return 0
5: end if
6:  $p_{pq} \leftarrow \text{p-value}(P_s^r, Q_s)$ . //Across language variance
7:  $\text{frac} \leftarrow \frac{p_{pq}}{p_{pp}}$ 
8: if  $\text{frac} > C$  then
9:   return 1
10: end if
11: Set  $\alpha \leftarrow 1$  //Prepare for binary search over  $\alpha$ 
12: Let  $P_s^\alpha \leftarrow (1 - \alpha)P_s + \alpha Q_s$  and  $\text{frac}^\alpha \leftarrow \frac{\text{p-value}(P_s^r, P_s^\alpha)}{p_{pp}}$ .
13: return  $\alpha^* = \arg \max_\alpha \text{s.t. } \text{frac}^\alpha > C$  // Do a binary search over  $\alpha$ 

```

---

Now consider both the events in the numerator,  $H_{pq}^\alpha$  indicates the truth that  $P_s^r$  has the same mean as that of a  $\alpha$  noisy sample of  $P_s$  which automatically implies that it also has the same mean as the original, noiseless, sample  $P_s$  resulting in the truth of the event  $H_{pp}$ . That is to say  $\Pr(H_{pp} | H_{pq}^\alpha) = 1$ . Thus the conditional probability reduces to Eqn. 2.

The hypothetical behavior of the joint probability (numerator in Eqn. 3) and our statistic are shown in Fig. 2. As the value of  $\alpha$  increases, the conditional probability in Eqn. 3 reduces and Fractional Similarity is the value ( $\alpha^*$ ) at which the fraction falls below the critical value ( $C$ ). The pseudo code to compute this value is shown in Algorithm 1. The code between lines 1-10 checks if the samples are either too dissimilar or very similar to each other. While lines 11-13 suggest a binary search procedure to find the value of required  $\alpha^*$ . During the binary search procedure, for any given arbitrary  $\alpha$  we don't need to explicitly combine instances from  $P_s$  and  $Q_s$  to obtain  $P_s^\alpha$ . Instead, since the T-test requires only mean and variance, we use the analytically derived values (Eqn. 1) of the combined sample. This relaxation makes the search process both efficient and also more accurate as the corresponding curve becomes a non-increasing function.

---

**Algorithm 2** FeatureSimilarity( $E, F$ )

---

**Input:** Feature values in English ( $E$ ) and the foreign language ( $F$ )

**Output:** Similarity of this feature in both these distributions

- 1: **for**  $i = 1 \rightarrow n$  **do**
  - 2:   Generate random samples  $E_s^r, E_s \sim E$  and  $F_s \sim F$
  - 3:   Estimate probability density function (pdf) from  $E - \{E_s^r \cup E_s\}$ .
  - 4:   Let  $L_e^r, L_e$  &  $L_f$  be the average log-likelihood of the queries in each sample
  - 5:    $\alpha_i^* \leftarrow \text{FractionalSimilarity}(L_e^r, L_e, L_f)$
  - 6: **end for**
  - 7: **return** Average( $\alpha_1^*, \dots, \alpha_n^*$ ).
- 

## 5 Cross-lingual Feature Selection

This section describes how we use Fractional Similarity to identify features that have a similar distribution in both English and foreign language training data. Algorithm 2 gives the pseudo code.

Let there be a total of  $m$  queries in our English training data ( $E$ ). We first generate two random samples ( $E_s^r$  and  $E_s$ ) from  $E$  with 10% of queries in each sample (line 2 of the pseudo code) leaving 80% of queries for training a pdf. We then generate a third sample ( $F_s$ ) of approximately the same number of queries from foreign language training data ( $F$ ). If we choose to include a query in any sample then we include feature values corresponding to all the results of the query. At this point, an instance in any of the above three samples corresponds to the feature value of a query-document pair.

However, simply comparing the means of the feature values would not test if the entire distributions are similar. If we have a pdf of the feature, though, we can use the value of the likelihood of a sample under that pdf to transform the problem of verifying the equality of means to the problem of verifying if the underlying distributions are same. This holds because probability at any point is the area of an infinitesimal rectangle considered around that point. So by comparing the average log-likelihood values we compare the area under the trained English pdf at random points drawn from English and foreign language distributions. Thus the similarity between these log-likelihood values is an indicator of the similarity between the underlying distributions from which the samples are drawn. Another advantage of using the likelihood is that it makes the approach amenable to different types of features with appropriate choice of pdf estimate.

Since we assume that English has more training data, we estimate the probability density function (pdf) of this feature from the remaining 80% of English training data (line 3) using kernel density estimation [20]. Now we compute log-likelihood of the feature values in all the three samples and then compute the average of all the instances per query (line 4). Let the average log-likelihood values of all the three samples be stored in  $L_e^r, L_e$  and  $L_f$  respectively. Here, each log-likelihood value is a property of a query which is an aggregate measure, the joint probability, over all the results of this query. Next, we use Fractional Similarity described in Sec. 4.2 to compute similarity between the distributions

that generated the log-likelihood values (line 5). We repeat this process multiple times and return the average of all the similarity values (line 7) as the similarity of this feature between English and foreign language training data.

There are two main key insights in the way we apply Fractional Similarity to IR. The first one is that, our sampling criterion is based on queries and not on query-document pairs, this implies while computing Fractional Similarity we try to find the fraction of English queries that can be replaced with foreign language queries. Because we are sampling queries as a unit, this enables us to deal with query-specific variance more directly. Secondly, we also deal with query-set variance better because the denominator of Eqn. 2 includes normalization by an estimate of the within language query-set variance.

Though we have developed our approach in the context of Information Retrieval, it is applicable in any scenario in which correlation arises among groups of instances. Furthermore, with appropriate choice of pdf estimation technique, it also straightforwardly extends to different types of features and to multivariate setting.

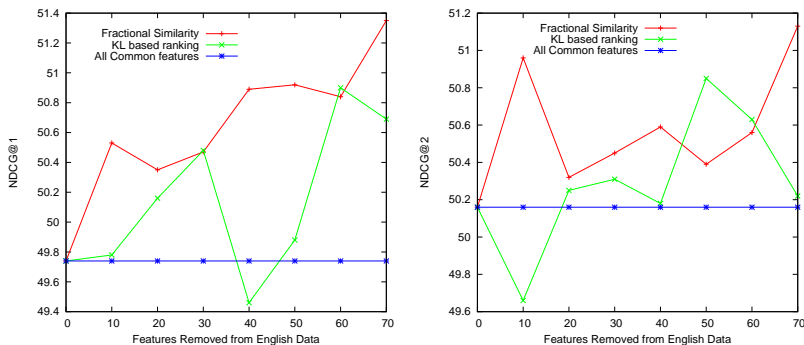
## 6 Experiments

We experimented with the task of improving a German language web search ranking model by using English training data. For both the languages, we took random query samples from the logs of a commercial search engine. Both these query sets differ in terms of the time spans and the markets they cover. English queries are from U.S. market while German queries are from German market. The English data set contains 15K queries with an average of 135 results per query while the German data set has 7K queries with an average of 55 results per query. The German data set has a total of 537 features, including some click-through features, of which 347 are common to the English data set. Our aim is to select features from these common features whose training data from English can also be used in training the German ranker. We only use a subset of features from English and always use all the 537 features in German data set. For each of those common features, we use Alg. 2 to compute its similarity across English and German data sets and then rank features according to their similarity and use them selectively. We also compute feature similarity using KL-divergence and use it as a baseline system to compare with Fractional Similarity. Note that KL-divergence based feature ranking doesn't take query-specific variance into account. We use a state-of-the-art learning algorithm, LambdaMART [3], to train the ranker.

In the first experiment, we compare different types of combination strategies to verify if English training data can improve the relevance ranking in German. On a smaller data set (of approximately half size), we try two adaptation strategies: the first one is to consider the data corresponding to all the common features from the English data set and simply add it to the German training data to learn the ranker ('Eng\_align+German' run). This strategy discards the fact that the new data is from a different language and treats it as if it were coming

	NDCG@1	NDCG@2	NDCG@10
German only	48.69	49.42	56.52
Eng_adapt+German	49.87	49.91	56.91
Eng_align+German	<b>50.77</b>	<b>50.63</b>	<b>57.47</b>

**Table 1.** Performance with different combination strategies.



**Fig. 3.** Performance obtained by gradually removing noisy features from English data. X-axis denotes the number of features removed and Y-axis denotes the NDCG values.

from German but with some missing feature values. We use model adaptation [13] as another strategy. Here, we first learn a base ranker on the English data (out-of-domain) with only the common features and then adapt it to the German training data (in-domain). We also report the results of a baseline ranker learned only on the German training data. The NDCG scores [21] for these different runs are shown in Table 1. Though we also report NDCG@10, we are more interested in the performance among the first few positions where improved results often impact the user experience. The results show that both strategies improved upon the German only baseline – indicating the benefit of using English training data in training the German ranker. Also, we observe that simply selecting the common features from the English data and appending it to the German data (‘\_align’ strategy) showed considerable improvements compared to the adapted version. So in our further experiments we report results only using the ‘\_align’ strategy.

The previous experiment showed that using English training data improves the German ranker, but it is not clear if all the common features are needed or only a subset of them are really useful. To explore this further, we first identify the mismatched features using Fractional Similarity and then repeat the experiment multiple times while gradually removing the mismatched features. We also use KL-divergence to identify the mismatched features and compare it with the ranking obtained by Fractional Similarity. We divide the corpus into two random splits and then subdivide each split into 80%, 10% and 10% for training, validation and test sets respectively. The results reported in Fig. 3 are averaged

over both the test sets. Each time we remove the least similar ten features returned by the two feature ranking schemes and retrain two separate rankers. Here, the baseline system uses all the common features from the English data. Note that the baseline number is slightly different from the Eng\_align+German number in Table 1 as the latter was only over a subset of the data. The red line in Fig. 3 indicates removing the mismatched features found by Fractional Similarity while the green line uses feature rankings obtained by KL-divergence. Though removing features based on KL-divergence also improves the performance compared to the baseline system, its performance fell short sometimes while our method always performed better than the baseline system. Except at one point, our method always performed better than KL-divergence feature selection. Removing approximately 25% of the mismatched features gave an improvement of 1.6 NDCG points at rank 1 (left image of Fig. 3) which is on the order of notable improvements in competitive ranking settings reported by other researchers [2, 3]. From the graphs, it is very clear that our method identifies the noisy features better than the KL-divergence based method. We observed this consistently at higher ranks, but as we go down (rank  $\geq 5$ ), feature selection using either method has, on average, either a neutral or negative effect. We believe this is because as we go deeper, the task becomes more akin to separating the relevant from the non-relevant rather than identifying the most relevant. While many of these feature distributions differ, they still provide a useful signal for general relevance, and thus transferring all of the data may prove superior when identifying general relevance is the goal.

Although feature selection can improve performance, naturally we expect that being too aggressive hurts performance. In fact, we observed that if we use only the 100 most similar features then the performance drops below the baseline ('Eng\_align+German'). In our experiments, we found that transferring 70-75% of the most similar features yielded the greatest improvements.

## 7 Discussion

Because of the high variance of a feature within a language and the challenges of comparing samples from two different query sets, traditional statistical significance tests output negligible probabilities – failing to correctly discriminate between feature distributions that truly differ and those that do not across languages. Fractional Similarity overcomes this problem by explicitly accounting for the within language variance of a feature by sampling in query blocks and normalizing by a within language factor. This increases the robustness of our approach and enables a fair comparison of the similarity scores between different features. Furthermore, empirical results demonstrate notable wins using a state-of-the-art ranker in a realistic setting.

## References

1. Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G.: Learning to rank using gradient descent. In: ICML '05, New York, NY,

- USA, ACM (2005) 89–96
2. Burges, C.J.C., Ragno, R., Le, Q.V.: Learning to rank with nonsmooth cost functions. In: NIPS, MIT Press (2006) 193–200
  3. Wu, Q., Burges, C.J., Svore, K.M., Gao, J.: Adapting boosting for information retrieval measures. *Information Retrieval* **13**(3) (2010) 254–270
  4. Gao, J., Qi, H., Xia, X., Yun Nie, J.: Linear discriminant model for information retrieval. In: Proceedings of the 28th international ACM SIGIR conference, ACM Press (2005) 290–297
  5. Robertson, S., Zaragoza, H., Taylor, M.: Simple BM25 extension to multiple weighted fields. In: CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management, New York, NY, USA, ACM (2004) 42–49
  6. Agichtein, E., Brill, E., Dumais, S.: Improving web search ranking by incorporating user behavior information. In: SIGIR '06, New York, NY, USA, ACM (2006) 19–26
  7. Broder, A.: A taxonomy of web search. *SIGIR Forum* **36**(2) (2002) 3–10
  8. Rose, D.E., Levinson, D.: Understanding user goals in web search. In: WWW '04: Proceedings of the 13th international conference on World Wide Web, New York, NY, USA, ACM (2004) 13–19
  9. Xue, G.R., Zeng, H.J., Chen, Z., Yu, Y., Ma, W.Y., Xi, W., Fan, W.: Optimizing web search using web click-through data. In: CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management, New York, NY, USA, ACM (2004) 118–126
  10. Gao, W., Blitzer, J., Zhou, M.: Using english information in non-english web search. In: iNEWS '08: Proceeding of the 2nd ACM workshop on Improving non english web searching, New York, NY, USA, ACM (2008) 17–24
  11. Chinnakotla, M.K., Raman, K., Bhattacharyya, P.: Multilingual pseudo-relevance feedback: performance study of assisting languages. In: ACL '10: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Morristown, NJ, USA (2010) 1346–1356
  12. Chinnakotla, M.K., Raman, K., Bhattacharyya, P.: Multilingual PRF: english lends a helping hand. In: SIGIR '10, New York, NY, USA, ACM (2010) 659–666
  13. Gao, J., Wu, Q., Burges, C., Svore, K., Su, Y., Khan, N., Shah, S., Zhou, H.: Model adaptation via model interpolation and boosting for web search ranking. In: EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Morristown, NJ, USA (2009) 505–513
  14. Bai, J., Zhou, K., Xue, G., Zha, H., Sun, G., Tseng, B., Zheng, Z., Chang, Y.: Multi-task learning for learning to rank in web search. In: CIKM '09, New York, NY, USA, ACM (2009) 1549–1552
  15. Carterette, B., Pavlu, V., Kanoulas, E., Aslam, J.A., Allan, J.: Evaluation over thousands of queries. In: SIGIR '08, New York, NY, USA, ACM (2008) 651–658
  16. Mansour, Y., Mohri, M., Rostamizadeh, A.: Domain adaptation: Learning bounds and algorithms. *CoRR* **abs/0902.3430** (2009)
  17. Ben-david, S., Blitzer, J., Crammer, K., Sokolova, P.M.: Analysis of representations for domain adaptation. In: NIPS, MIT Press (2007)
  18. Welch, B.L.: The generalization of ‘student’s’ problem when several different population variances are involved. *Biometrika* **34**(1/2) (1947) 28–35
  19. A., F.R.: Applications of “student’s” distribution. *Metron* **5** (1925) 90104
  20. Parzen, E.: On estimation of a probability density function and mode. *The Annals of Mathematical Statistics* **33**(3) (1962) 1065–1076
  21. Järvelin, K., Kekäläinen, J.: IR evaluation methods for retrieving highly relevant documents. In: SIGIR '00, New York, NY, USA, ACM (2000) 41–48