Leveraging Interlingual Classification to Improve Web Search

Jagadeesh Jagarlamudi
University of Maryland
College Park, MD
jags@umiacs.umd.edu

Paul N. Bennett
Microsoft Research
Redmond, WA
pauben@microsoft.com

Krysta M. Svore
Microsoft Research
Redmond, WA
ksvore@microsoft.com

ABSTRACT

In this paper we address the problem of improving accuracy of web search in a smaller, data-limited search market (*search* language) using behavioral data from a larger, data-rich market (*assist* language). Specifically, we use interlingual classification to infer the search language query's intent using the assist language click-through data. We use these improved estimates of query intent, along with the query intent based on the search language data, to compute features that encode the similarity between a search result (URL) and the query. These features are subsequently fed into the ranking model to improve the relevance ranking of the documents. Our experimental results on German and French languages show the effectiveness of using assist language behavioral data – especially, when the search language queries have small click-through data.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.5.4 [**Pattern Recognition**]: Applications—*Text Processing*

General Terms

Algorithms, Experimentation

Keywords

Web Search, Cross-lingual Information Retrieval, Text Classification, Machine Translation

1. INTRODUCTION

Since commercial search engines accumulate user behavior data over time, the amount of behavioral information available in different languages varies significantly in size and quality. In this paper, we address the problem of improving web search ranking in a relatively data-scarce *search language* (*e.g.*, German or French) using data from an *assisting language* (*e.g.*, English). There are two general approaches to address this problem based on the availability of user queries and tools in both languages. When the user queries are not available, transferring knowledge between languages can be posed as a transfer learning problem [4]. On the other hand, when the actual queries are available we can use bilingual query pairs (pairs of queries in both languages with the same information need) to improve search language ranking [3]. In this paper, we improve search language ranking by identifying bilingual query pairs between search and assisting languages (*e.g.*, 'Kleider' in German and 'dresses' in English), and then transferring query intent (*e.g.*, 'Shopping') from the assisting language to the search language. We encourage the reader to refer to [5] for more details on the approach and experimentation.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

2. LANGUAGE-ASSISTED SEARCH

We represent a query's intent as a distribution over a predefined set of classes [1]. We use aligned class labels from ODP as an interlingua to transfer query intent from the assisting language to the search language. Our approach involves three stages: URL classification, Transfer of query intent and Training/Ranking.

URL classification: Documents in both search and assisting languages are classified into the same set of class labels. We use the publicly available ODP categories for our classification taxonomy, which are aligned in multiple languages. The ODP categories are available in a broad variety of languages (over 85). We truncate the ODP classification taxonomy to the top two levels which cover all the sub-classes of the hierarchy ranging from 'Society/Issues' to 'Kids and Teens/School Time' and 'Sports/Baseball' to 'Computers/Virtual Reality' and so on. The truncated taxonomy has 219 English class labels, of which 204 and 197 of them have a corresponding aligned class in German and French, respectively. We train a different classifier in each language. For both the languages, the training data for each class is obtained from crawling the links provided by ODP. These classifiers are used to derive class distributions, $P(c|u)$, for result documents in both languages. The classification of each document is performed offline and the class labels can be stored in the index for efficient retrieval.

Transfer of query intent: Search language queries are translated into the assisting language using Bing's publicly available machine translation (MT) API. If the translated query exists in the assisting language query log then we keep the query pair (*bilingual* query pair) otherwise we ignore it. Let c, q_s (q_a), and u_s (u_a) represent a class, search (assisting) language query, and the search (assisting) language candidate URL, respectively. Then, we derive two class distributions for the search language query, $P_s(c|q_s)$ and $P_a(c|q_s)$, using the search and assisting language click-through data:

$$P_a(c|q_s) = \sum_{q_a, u_a} P(c|u_a)P(u_a|q_a)P(q_a|q_s) \qquad (1)$$

$$P_s(c|q_s) = \sum_{u_s} P(c|u_s)P(u_s|q_s) \quad (2)$$

where $P(c|u_a)$ is the class probability of a given URL, obtained using the assisting language URL classifier, $P(u_a|q_a)$ is the query-url weight (see [1] for more details) and $P(q_a|q_s)$ is the query translation probability. The variables are defined similarly for the search language as well.

Training/Ranking: Subsequently, the similarity between a query class distribution and a document class distribution is expressed through several features. We use the same feature set as used in Bennett *et al.* [1]. These features are appended to the original query-document feature vector and used as input when training or in applying the model.

3. EXPERIMENTS

In our experiments, we consider the search language to be either French or German, and the assisting language to be English. For both the search languages, we took a random sample of ($\sim$ 40K) queries from a commercial search engine, translated them using publicly available MT system, and retained only the queries whose translation existed in a larger sample of English query log. Finally, for German and French languages, we used 2117 and 2359 queries with an average of 54 and 55 results per query, respectively. Each query-URL pair is represented by a 5-level-scale human relevance label and a feature vector of several hundred features.

We train our ranking models using a state-of-the-art boosted-decision-tree algorithm [2]. We perform 10-fold cross validation and report the average of the individual runs. We also perform extensive parameter sweeps and choose the best parameter combination based on validation data and report the accuracy on the test set. We measure accuracy using normalized discounted cumulative gain (NDCG), and optimize in particular for NDCG@3. Our baseline model ('Baseline') is trained on the search language training data using traditional IR features including BM25F and other query-document match features some of which are derived from click-through data. The Baseline model is very competitive with state-of-the-art ranking models. The 'Search' and 'Assist' models are trained using class based features derived using the search and assisting language query class distributions, respectively, in addition to the traditional IR features available for Baseline model.

Table 1 shows the average NDCG scores of the various models. For both the languages, deriving the query class distribution from English (Assist) outperforms the Baseline significantly (as measured by the paired t-test), achieving, for example, NDCG@10 gains of 0.26 and 0.22 for French and German languages, respectively. When we compare the use of class-derived features from the search language versus the assisting language, Assist wins in the case of French, and ties with the Search model in the case of German. This indicates that both the models carry useful information.

Our examination of the per query NDCG scores on one train/test fold of the French data reveals that the Assist model performs best on queries for which the total click volume in search language is small while the Search model wins otherwise. This suggests an ensemble model ('Assist+Search') of using the assisting model when the total click volume for a query falls below a threshold, and using the search model otherwise. Using the validation data, we chose the optimum threshold based on NDCG@1 and applied this to the test data. Table 2 shows the improvement of the ensemble model over Search model. Notice that these gains are in addition to those already achieved by the Search model over the Baseline model. The combined ensemble approach achieves significant improvements of NDCG over both the Search and the Baseline models.

	N@1	N@3	N@5	N@10
French				
Baseline	66.32	66.07	67.07	69.68
Search	66.4	**66.36***	67.3*	69.89*
Assist	**66.51**	66.27*	**67.32***	**69.94***
Δ over Baseline	0.19	0.2	0.25	0.26
German				
Baseline	63.73	64.46	66.11	69.28
Search	**64.22***	**64.75***	**66.43***	**69.52***
Assist	63.93	64.72*	66.35*	69.5*
Δ over Baseline	0.2	0.26	0.24	0.22

Table 1: Test NDCG(N) results for German and French as search language. Bold indicates winning model and * indicates significant improvement compared to baseline at p-value of 0.05.

	N@1	N@3	N@5	N@10
Assist+Search	0.7917	0.4172	0.5267	0.3403
Δ over Search				

Table 2: NDCG(N) improvements of the ensemble over using class information from Search alone.

4. DISCUSSION

In this paper, we have introduced a novel method of transferring information from an assisting language to a search language, using publicly available resources. A key contribution of our approach is its ability to address the challenge in accurately ranking an infrequent or tail query by transferring the query intent from its assisting-language translation – which may have sufficient behavioral data in the assisting language. Our experimental results validate the effectiveness of using assisting language to improve accuracy in a search language, and confirm the potential benefits of combining search and assisting language class features.

5. REFERENCES

[1] P. N. Bennett, K. Svore, and S. T. Dumais. Classification-enhanced ranking. WWW '10. ACM.

[2] C. J. Burges, K. M. Svore, P. N. Benett, A. Pastusiak, and Q. Wu. Learning to rank using an ensemble of lambda-gradient models. *JMLR*, 14:25–35, 2011.

[3] M. K. Chinnakotla, K. Raman, and P. Bhattacharyya. Multilingual PRF: english lends a helping hand. In *SIGIR '10*, pages 659–666. ACM, 2010.

[4] J. Jagarlamudi and P. N. Bennett. Fractional similarity: Cross-lingual feature selection for search. In *ECIR 2011*. Springer, 2011.

[5] J. Jagarlamudi, P. N. Bennett, and K. M. Svore. Leveraging interlingual classification for web search. Technical Report MSR-TR-2012-11, Microsoft Research, 2012.