Modeling Click-through based Word-pairs for Web Search

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

Jianfeng Gao
Microsoft Research
Redmond, WA, U.S.A.
jfgao@microsoft.com

ABSTRACT

Statistical translation models and latent semantic analysis (LSA) are two effective approaches to exploit click-through data for web search ranking. This paper presents two document ranking models that combine both approaches by explicitly modeling word-pairs. The first model, called PairModel, is a monolingual ranking model based on word pairs that are derived from click-through data. It maps queries and documents into a concept space spanned by these word pairs. The second model, called Bilingual Paired Topic Model (BPTM), uses bilingual word pairs and jointly models a bilingual query-document collection. This model maps queries and documents in multiple languages into a lower dimensional semantic subspace. Experimental results on web search task show that they significantly outperform the state-of-the-art baseline models, and the best result is obtained by interpolating PairModel and BPTM.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.6 [**Artificial Intelligence**]: Learning; I.5.4 [**Pattern Recognition**]: Applications—*Text Processing*

General Terms

Learning, Algorithms, Experimentation

Keywords

Clickthrough Data, Latent Semantic Analysis, Topic Models, Multilingual IR, Translation Model, Web Search

1. INTRODUCTION

Click-through data has been used to address the lexical gap problem that frequently arises in web search task. For example, statistical translation based approaches, first, learn pairwise associations between query and document words and then use these translation probabilities to rank the documents [1]. On the other hand, topic modeling based approaches treat a document and the set of queries for which it has been clicked as an *aligned query-document pair* and learn a shared topic distribution [2]. They use the shared topic distribution to rank documents. In this paper, we aim to design models which use pairwise word associations and the alignment information between a document and its clicked query stream to jointly model a query-document collection. We achieve this by using word-pairs (referred to as concepts) as latent variables.

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

2. PAIR MODEL

Our first model, called PairModel, is a monolingual model. It uses monolingual word pairs to model a query-document collection in a language (say French). Formally, we assume that an aligned query-document pair $(\mathbf{q}, \mathbf{d}) = \big((q_1, \cdots, q_{N_q}), (d_1, \cdots, d_{N_d})\big)$ share the concept distribution $\psi_{(\mathbf{q},\mathbf{d})}$ – a multinomial distribution over conecpts (c). For each word in the query-document pair, we first draw a concept (c) from the concept distribution and then draw a word from the words associated with the concept. Let $|C|$ be the total number of concepts, then we assume the following generative story for generating a document and its query stream.

1. For each concept $c = 1 \cdots |C|$, choose $\phi_c \sim \text{Dir}(\beta)$
2. For each aligned query-document pair
   (a) Choose $\psi_{(\mathbf{q},\mathbf{d})} \sim \text{Dir}(\gamma)$
   (b) For each document term $i = 1 \cdots N_d$
      - Select a concept $c_i \sim \text{Mult}(\psi_{(\mathbf{q},\mathbf{d})})$
      - Select a word $w_i \sim \text{Mult}(\phi_{c_i})$
   (c) For each query term $i = 1 \cdots N_q$
      - Select a concept $c_i \sim \text{Mult}(\psi_{(\mathbf{q},\mathbf{d})})$
      - Select a word $q_i \sim \text{Mult}(\phi_{c_i})$

where β and γ are hyperparameters. We use Expectation Maximization algorithm and derive MAP estimates for ϕ and ψ [2].

3. BILINGUAL PAIRED TOPIC MODEL

Our second model is a bilingual model called Bilingual Paired Topic Model (BPTM). It uses bilingual word-pairs (say between French and English) to model a *bilingual* query-document collection. Let $\big\{(\mathbf{q}_i^e, \mathbf{d}_i^e), (\mathbf{q}_j^f, \mathbf{d}_j^f)\big\}$ for $i{=}1 \cdots m$ and $j = 1 \cdots n$ represent aligned query-document collections. The queries and documents across different languages are assumed to be comparable (e.g., from the same time period) but not necessary to be translations of each other. We assume that a query-document pair share the topic distribution $\theta_{(\mathbf{q},\mathbf{d})}$ which is a multinomial distribution over T bilingual topics and is drawn from a Dirichlet symmetric prior (α). Each of these bilingual topics (ψ_k) is a multinomial distribution over the concepts (bilingual word-pairs) [3].

	English			German			French		
	ndcg@1	ndcg@3	ndcg@10	ndcg@1	ndcg@3	ndcg@10	ndcg@1	ndcg@3	ndcg@10
JMLM	28.7	37.54	48.66	34.59	43.79	56.2	38.22	46.36	60.57
WTM	31.79	40.77	51.31	36.54	46.26	58.63	40.09	48.89	*63.06*
BLTM	34.7	43.16	53.03	37.26	46.5	58.33	40.03	48.36	62.49
PairModel	**35.02**	**43.46**	**53.26**	*39.55*	*47.64*	*58.98*	*40.94*	*48.95*	62.58
BPTM	NA	NA	NA	37.38	46.58	58.48	40.85	49.01	62.73
BPTM+PairModel	NA	NA	NA	**39.66**	**47.73**	**59.04**	**42.25**	**50.01**	**63.42**

Table 1: Comparison of PairModel, BPTM and their combination (BPTM+PairModel) with state-of-the-art baseline systems. In each column, the best system is bolded and the second best system is italicized.

Finally, depending on the language of the query-document pair ($l_{(\mathbf{q},\mathbf{d})}$), these concepts generate words. Given a concept (bilingual word-pair) and the language, there is only one option for choosing the word and is deterministic. We follow Jagarlamudi and Daumé III [3] and add dummy translations to handle out-of-vocabulary words. Let $\mathbb{I}(c_i, l_d)$ denote a binary indicator variable that denotes whether the concept c_i can generate a word from the language l_d, then the generative process of BPTM is as follows.

1. For each topic $k = 1 \cdots T$, choose $\psi_k \sim \text{Dir}(\gamma)$
2. For each aligned query-document pair
   (a) Choose $l_{(\mathbf{q},\mathbf{d})} \sim \text{Bin}(\frac{1}{2})$.
   (b) Choose $\theta_{(\mathbf{q},\mathbf{d})} \sim \text{Dir}(\alpha)$
   (c) For each document term $i = 1 \cdots N_d$
   - Select a topic $z_i \sim \text{Mult}(\theta_{(\mathbf{q},\mathbf{d})})$
   - Select a concept $c_i \sim \text{Mult}(\psi_{z_i}) \cdot \mathbb{I}(c_i, l_{(\mathbf{q},\mathbf{d})})$
   - Select a word $w_i \sim P(w_i|c_i, l_{(\mathbf{q},\mathbf{d})})$

   (d) For each query term $i = 1 \cdots N_q$
   - Select a topic $z_i \sim \text{Multi}(\theta_{(\mathbf{q},\mathbf{d})})$
   - Select a concept $c_i \sim \text{Mult}(\psi_{z_i}) \cdot \mathbb{I}(c_i, l_{(\mathbf{q},\mathbf{d})})$
   - Select a word $q_i \sim P(q_i|c_i, l_{(\mathbf{q},\mathbf{d})})$

Like in the case of PairModel, we use Expectation Maximization to learn the MAP estimates of the parameters.

4. EXPERIMENTS AND DISCUSSION

We compare our models with a unigram language model with Jelenek-Mercer smoothing (JMLM) and state-of-the-art baseline systems, such as word based translation model (WTM) [1] and Bilingual Topic Model (BLTM) [2], in three languages: English, German and French. For evaluation, we use a random sample of approximately 5K queries from Bing query logs with at least 10 results per query. Query-document pairs are manually judged on a scale of 0 to 5. We use 128K, 133K and 2.1M query-document pairs in German French and English languages, respectively, to learn monolingual word-pairs for PairModel and WTM, and to learn semantic subspace for BPTM and BLTM. The bilingual dictionaries required for BPTM, in English-German and English-French language pairs, are learnt using Giza++ on parallel data used for Bing machine translation system. Similar to [1, 2], we use only document titles in our experiments and use the probability estimates returned by our models to smooth the document unigram language model. The interpolation parameters in all the systems are estimated using two-fold cross validation. We use normalized discounted cumulative gain (ndcg) to evaluate the ranking against the human judgements.

Table 1 shows the results of different systems. Across all the languages, JMLM smoothing does poorly because of the lexical gap problem. As expected, WTM outperforms JMLM demonstrating the effectiveness of addressing the term mismatch problem using co-occurrence statistics. BLTM performs significantly better than WTM on English but is indistinguishable on German and French. This is largely due to the smaller training data used in German and French. The order of the baseline systems is consistent with what has been reported in the previous literature [2].

Our PairModel, which uses both the word-pairs and the alignment information between query and document pairs, outperforms all the baseline systems except for ndcg@10 in French. The next row shows the results of BPTM, which uses English data to improve web search ranking in German and French languages. On its own, BPTM is marginally better than BLTM and performs poorly compared to PairModel. But the interpolated model ('BPTM+PairModel') outperforms all the models and gives notable improvements, especially for French. This is because PairModel exploits monolingual query-document collection while BPTM transfers useful information from the English collection and combining both these models leads to a superior model than the individual ones. The improvements are smaller for German because of the rich morphology and the compound word phenomenon of German which limits the coverage of the English-German bilingual dictionary.

In this paper, we propose two models which exploit two forms of evidence mined from click-through data, *i.e.* word-pair associations and the alignment information. Our models achieve promising improvements for emerging markets such as French and German where the click stream is small. Our second model uses bilingual dictionaries to map queries and documents of both the languages into a common lower dimensional subspace and thus uses training data from a data rich assisting language to improve ranking in a resource poor language. Our models can operate on both aligned and unaligned data, hence they can potentially work without the click-through data, though, in this case, the word-pairs need to be derived from another source like Wordnet.

5. REFERENCES

[1] J. Gao, X. He, and J.-Y. Nie. Clickthrough-based translation models for web search: from word models to phrase models. In *CIKM 10*, pages 1139–1148. ACM.

[2] J. Gao, K. Toutanova, and W.-t. Yih. Clickthrough-based latent semantic models for web search. In *SIGIR '11*, pages 675–684. ACM.

[3] J. Jagarlamudi and H. Daumé III. Extracting multilingual topics from unaligned comparable corpora. In *ECIR '10*, volume 5993, pages 444–456. Springer.