Position-Aligned Translation Model for Citation Recommendation

Jing He[1], Jian-Yun Nie[1], Yang Lu[2], and Wayne Xin Zhao[2]

[1] Université de Montréal
{hejing,nie}@iro.umontreal.ca
[2] Peking University
{luyang,zhaoxin}@pku.edu.cn

Abstract. The goal of a citation recommendation system is to suggest some references for a snippet in an article or a book, and this is very useful for both authors and the readers. The citation recommendation problem can be cast as an information retrieval problem, in which the query is the snippet from an article, and the relevant documents are the cited articles. In reality, the citation snippet and the cited articles may be described in different terms, and this makes the citation recommendation task difficult. Translation model is very useful in bridging the vocabulary gap between queries and documents in information retrieval. It can be trained on a collection of query and document pairs, which are assumed to be parallel. However, such training data contains much noise: a relevant document usually contains some relevant parts along with irrelevant ones. In particular, the citation snippet may only mention only some parts of the cited article's content. To cope with this problem, in this paper, we propose a method to train translation models on such noisy data, called position-aligned translation model. This model tries to align the query to the most relevant parts of the document, so that the estimated translation probabilities could rely more on them. We test this model in a citation recommendation task for scientific papers. Our experiments show that the proposed method can significantly improve the previous retrieval methods based on translation models.

1 Introduction

Users often use terms in their queries that are different from those in the documents. Similar situations appear in a recommendation system: the recommended element and the recommendation context (query) may be described by different terms. The phenomenon leads to what we call the vocabulary gap, or term mismatch problem, which is crucial to solve in information retrieval and recommendation system.

Many efforts have been devoted to address this problem by mining relationships between the terms from the document collection [1,2]. The mined relationships can be used to expand the query by adding related terms. Although the document collection is a valuable resource for relation mining, one can only create relationships between terms in the documents. However, the vocabularies

L. Calderón-Benavides et al. (Eds.): SPIRE 2012, LNCS 7608, pp. 251–263, 2012.

used in queries are substantially different from those used in documents, making the effect of the above approach limited.

More recent attempts exploited data that connect queries to documents such as user click-though. This allows us to create relationships between terms capable of bridging documents and queries. In [3] a standard statistical translation model (IBM model 1) is trained by assuming that a query is parallel to the title of the document clicked by the user. Despite the fact that the query and the title is not parallel in the sense of translation as in machine translation (MT), the term relations extracted are shown to be highly useful in IR. Similar approaches have been used successfully in several IR applications such as cross-linguistic retrieval [4,5], question answering [6,7], ad-hoc and Web retrieval [8,2], information flow tracking [9] and citation recommending [10]. In the latter applications, even noisier training data are used such as pairs of queries and relevant documents.

We notice that the previous studies used the same approach as in MT to train translation models (typically IBM model 1). There is however an important difference on the training data in the above IR-related applications: the data are no longer truly parallel sentences, but related texts. One may argue that applying the same training process on related data can still result in useful "translation" relations between related terms. This is true to some extent. When the proportion of noise (i.e. unrelated parts) increases in the training data, the resulting translation model may be highly prone to noise and its usefulness can be significantly reduced. Consider, for example, the case of query and relevant document pairs, a query is usually much shorter than the relevant documents. This leads to a translation model that "translates" a query term by many documents terms. If all the parts of the document were relevant to the query, this could produce a desirable effect for IR. However, in practice, even when one document is relevant to a query, usually only some parts of it are relevant and the other parts are not. Assuming that the whole document is "parallel" to the query will unavoidably lead to a very noisy translation model, i.e. many query terms are translated from unrelated document terms. Unfortunately, this phenomenon has been hardly considered in the context of IR.

In machine translation, noisy data have also been used to train translation models. In most cases, some filtering is performed to select the parallel parts from the data [11,12]. With query-document pairs, we can also resort to some selection process to create a better training data. For example, one may select the most similar passages using cosine similarity or any retrieval score. However, the fact that we rely on the query to perform a selection will result in a subset of data that share words with the query. This may limit the ability of the resulting translation model to effectively bridge different query and document terms.

In this paper, we will use a different approach by adding a position parameter in the alignment: A document is considered to be composed of different passages. Each passage is intended to describe a specific topic. It is aligned to the query to some extent. The stronger is the alignment, the more will the translation model rely on it. The above idea can be naturally incorporated into the translation

model's training process, i.e., Expectation-maximization (EM) process. We call such a model Position-aligned Translation Model (PTM).

We carried out experiments on the proposed methods in a citation recommendation task, i.e. given a context in a scientific paper, we want to identify the correct reference for it. This task can be a real task, but we use it more as a testbed to evaluate our position-aligned translation model: We will see if it is more reasonable to assume that passages of the cited document correspond to the citing context to different degrees. Our results show that the position-aligned translation model performs clearly better than the one trained with the entire document.

In the remainder of the paper, we will first describe some related work (Section 2). We will then describe our position-aligned translation model (Section 3). Experiments will be presented in Section 4 and finally conclusions are drawn in Section 5.

2 Related Work

Translation Model was introduced to be used for information retrieval by Berger et al. [3]. The main idea is that the translation model can translate the terms from the documents to the terms in the queries, so it can bridge the vocabulary gap between the query and the document. In machine translation, a translation model is trained using a parallel corpus.

The translation model can be naturally applied in the cross-language retrieval [4,5]. Furthermore, in many other applications where queries and documents use different sets of vocabularies, the translation model can be used to bridge the vocabulary gap. Murdock et al. [7] and Xue et al. [6] employed the translation model to retrieve sentences and questions in a frequently asked question (FAQ) archive. Metzler et al. [9] used translation models as a similarity measure for the information flow tracking task. More recently, [2] trained translation model with mutual information of the cooccurent terms, and used it for improving ad-hoc retrieval. Gao et al. [8] extended the translation model to translate between phrases, and used it bridge the vocabulary gap between the Web search query and the page title.

However, in all the above studies, translation models are trained using the same tools as in machine translation. Although the noisy nature of the training data used has been widely recognized, one often looked at the positive expansion effect than the possible negative topic drift effect. In fact, if a translation model is very noisy, the resulting translations will not be strongly related to the original terms. There is then a high risk of topic drift, leading to matching a document to a very different query. The risk of topic drift is much increased when one uses query-document pairs for training: As we stated earlier, even when a document is relevant to a query, it is usually the case that parts of it are relevant, and there are still other irrelevant parts. Using different parts of the document indistinctly for model training will unavoidably result in a very noisy model. To avoid the noise, some research instead trained the translation model on the parallel data with less noise, such as query-query pairs or query-title pairs [8].

A more reasonable approach is to segment the whole document into segments, and to rely more on the relevant ones. Similar ideas have been successfully used in passage retrieval [13,14]. The idea to use passages is intuitive. Indeed, although a document contains several topics, we can assume that descriptions on different topics do not follow randomly. An author usually talks on a topic in a continuous part before moving to a different topic. It is reasonable to assume a topical consistency within a segment. The idea is further extended in local context analysis [1], in which related passages are used for query expansion rather than the whole document. It is shown that this passage-based pseudo-relevance feedback is more effective than document-based feedback.

In this paper, the Position-aligned translation model bears some similarity to this family of approaches using passages. We extend the idea to the training process of translation models.

3 Position-Aligned Translation Model

In this section, we first briefly introduce the translation model for information retrieval, and then we propose the position-aligned translation model and its application in document ranking.

3.1 Translation Model for IR

A translation model defines the probability of translating terms in a source language into terms in another target language. When it is applied to information retrieval, we usually assume that the language used in the queries is different from that of the documents, and we can connect a query q and a document d by translating the terms in the document ($t_D \in d$) to the terms in the query ($t_Q \in q$) [3].

To estimate a translation model, we usually need a parallel corpus of two languages as the training dataset. In information retrieval, we can assume a query q and one document d that is relevant to q as a record in the parallel corpus. Therefore, the parallel corpus C is a collection of such query and relevant document pairs. To estimate the translation probabilities, we can assume that the query is generated from the relevant document.

The likelihood of generating a query q from a document d can be formulated as:

$$p(q|d) = \prod_{t_Q \in q} \sum_{t_D \in d} p(t_Q|t_D;\psi)p(t_D;\theta_d) \tag{1}$$

where $p(t_Q|t_D;\psi)$ is the translation probability from a document term t_D to a query term t_Q, and $p(t_D;\theta_d)$ is the probability of generating a document term t_D from the document language model θ_d.

The translation probabilities are usually estimated by the maximum likelihood EM estimation for all (query,document) pairs in the training dataset:

$$\hat{\psi} = \arg\max_{\psi} \prod_{(q,d)\in C} \prod_{t_Q \in q} \sum_{t_D \in d} p(t_Q|t_D;\psi)p(t_D;\theta_d) \tag{2}$$

In retrieval, the translation model can be combined together with the query likelihood language model. It has been found that the trained translation model usually underestimates the self-translation probability (translating one term to itself), so it usually boosts the self-translation in retrieval [2]. The ranking function can be formulated as:

$$P(q|d) = \prod_{t_Q \in q} \frac{|d| \cdot p_{\text{TM}}(t_Q|d) + \mu \cdot p(t_Q|C)}{|d| + \mu} \tag{3}$$

$$p_{\text{TM}}(t_Q|d) = \beta \cdot p(t_Q; \theta_d) + (1 - \beta) \cdot \sum_{t_D} p(t_Q|t_D; \psi) p(t_D; \theta_d) \tag{4}$$

where $p(t_Q|C)$ is the probability of term t_Q in the collection model, and μ is its coeffient, and β is the weight for the self-translation boosting. In Eq 3, it smooths the document model with the collection model using the Dirichlet smoothing method.

3.2 Position-Aligned Translation Model

As described in the previous section, we usually assume that the query is translated from terms of the relevant document. However, different from a parallel corpus in machine translation, the query and relevant document pairs in information retrieval are imbalanced and not strictly aligned. Compared to the relatively short queries, the document usually is very long and covers several topics. Even when a document is relevant to the query, one cannot assume that each part of the document is related to the query.

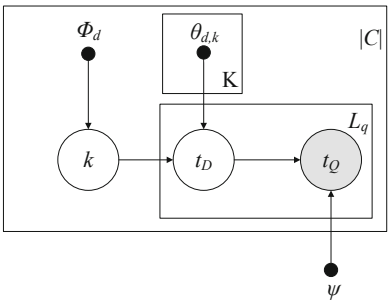

Fig. 1. Position-aligned Translation Model

Intuitively, it would help improve the translation probability estimation, if we can align a query to some highly relevant parts of the document only. This is however difficult to do: we do not know exactly which parts of the document are relevant. Relevant parts may appear anywhere. Fortunately, topics in a document do not change randomly. Authors usually describe about one topic at some length before moving to other (related) topics. Therefore, in a document,

the terms about one specific topic are likely to be clustered together. This suggests an approach based on document passages. Some previous work on topic segmentation in a document [15,16,17] and passage-based retrieval [18,19] has validated this idea. Here we use it in translation model training.

In the new translation model training process, we take into account the alignment strength between a query and parts of the document. Since this translation model is estimated by position alignment, we name it Position-aligned Translation Model (PTM). For a query and document pair (q, d), we assume a generation story of a query q from a document d as follows (Figure 1):

1. choose a position $k \sim \text{Multinomial}(\phi_d)$;
2. for each query term $t_Q \in q$:
   (a) choose a document term $t_D \sim \text{Multinomial}(\theta_{d,k})$
   (b) choose a query term $t_Q \sim \text{Multinomial}(\psi_{t_D})$

where ϕ_d is the prior position distribution of generating query q, and it can reflect the prior importance of different positions, e.g., the beginning positions of a document or the positions in some fields are more important. In this paper, we simply set is as a uniform distribution. $\theta_{d,k}$ is a position-specific language model, and ψ is the translation model. The position-specific language model of a position is determined by its surrounding terms. It can be presented by either a window of the surrounding terms (as fixed-length passage) [18,19], or a model whose terms are weighted decreasingly along with the distance to the position [13]. The generating process is depicted in Figure 1. Accordingly, the likelihood of generating a query q from a document d can be formulated as:

$$p(q|d) = \prod_{t_Q \in q} \sum_{t_D \in d} p(t_Q|t_D; \psi) p(t_D; \theta_{d,k}) p(k; \phi_d) \tag{5}$$

Similar to the original translation model, the parameters can be estimated by EM algorithm. As Figure 1 shows, the query term variables t_Q are observed, and the document positional language model parameters $P(t_D; \theta_{d,k})$ can be explicitly estimated for each position in a document. In the model, the generative position k for each document, and the generative term t_D for each query terms t_Q are latent variables. We can use EM algorithm to estimate the translation parameters as follows.

E-step

EM algorithm is an iterative algorithm, and we discuss the update process in the i-th iteration. In E-step, we estimate the posterior distribution of the latent variables given the current estimation of the parameters $\psi^{(i)}$.

We can update the posterior distribution of the document terms translated to a query term $p(t_D|t_Q, q, d; \psi^{(i)}, \phi_d)$ based on the current estimation of translation model $\psi^{(i)}$:

$$p(t_D|t_Q, q, d; \psi^{(i)}, \phi_d) = \frac{\sum_k p(k|q, d; \phi_d) p(t_Q, t_D; \theta_{d,k}, \psi^{(i)})}{\sum_{t'_D} \sum_k p(k|q, d; \phi_d) p(t_Q, t'_D; \theta_{d,k}, \psi^{(i)})} \tag{6}$$

where

$$p(t_Q, t_D; \theta_{d,k}, \psi^{(i)}) = p(t_Q|t_D; \psi^{(i)})p(t_D; \theta_{d,k}) \tag{7}$$

We also need to update the posterior position distribution $p(k|q, d; \psi^{(i)}, \phi_d)$ for each (query, document) pair as follows:

$$p(k|q, d; \psi^{(i)}, \phi_d) = \frac{p(k|q, d; \phi_d) \prod_{t_Q} \sum_{t_D} p(t_Q, t_D; \theta_{d,k}, \psi^{(i)})}{\sum_{k'} p(k'|q, d; \phi_d) \prod_{t_Q} \sum_{t_D} p(t_Q, t_D; \theta_{d,k'}, \psi^{(i)})} \tag{8}$$

This equation determines the importance of each position in a query-document pair. The positions that are more likely to generate the query are weighted more in the training phase, and hence they play a more important role than other parts in the translation model.

M-step

In M-step, we have to estimate the parameters so that the expected likelihood can be maximized. Here we can get the translation probability for $(i + 1)$-th iteration as follows:

$$p(t_Q|t_D; \psi^{(i+1)}) = \frac{\sum_{(q,d)\in C} p(t_D|t_Q, q, d; \psi^{(i)}, \phi_d)}{\sum_{t'_Q} \sum_{(q,d)\in C} p(t_D|t'_Q, q, d; \psi^{(i)}, \phi_d)} \tag{9}$$

Compared to a traditional translation model, the above PTM has a higher complexity. For IBM-1 model, the complexity is $\mathbf{O}(MNL_dL_q)$, where M is the number of EM iterations, N is the number of query-document pairs in the training collection, L_d and L_q are document length and query length respectively. For position-aligned translation model, it needs to calculate the joint distribution of t_D and t_Q for each position k (Eq 7), so the complexity is $\mathbf{O}(MNKL_pL_q)$, where K is the number of positions considered in a document, and L_p is the number of surrounding terms considered for each position. Since KL_p is usually larger than L_d, the complexity of the position-aligned translation model is higher. It is extremely expensive if we consider all positions in a document ($K = L_d$ in this case). Alternatively, we can pre-segment the document into some overlapped fixed-length passages, take the center position of each such passage as a candidate position, and only consider the surrounding terms in this passage. This is consistent with the assumption that topics in a document follow logically. The complexity is dependent on the overlap between the passages. A larger overlap leads to a higher complexity, since one document term is considered more times in different passages. Assuming the overlap is L_o, there are approximately $\frac{L_d}{L_p - L_o}$ passages in a document. Therefore, the complexity ratio between IBM-1 model and the position-aligned translation model is:

$$\frac{C(\text{TM})}{C(\text{PTM})} \approx \frac{MNL_dL_q}{MNL_p\frac{L_d}{L_p - L_o}L_q} = 1 - \frac{L_o}{L_p}$$

It shows that the complexity is determined by the ratio of overlap length and the passage length. In the extreme case, the position-aligned translation model

has the same complexity as IBM-1 model if there is no overlap between neighbor passages. If the overlap is half of the passage length, the complexity is doubled. By choosing appropriate passage length, we can avoid a large increase in complexity.

4 Experiments

4.1 Task and Data Set

In this paper, we experiment our models for a citation recommending task. In our experiments, rather than building such an application, we use the task to compare different translation models and their different utilizations. The input of the citation recommending task is an article snippet that needs a reference, and the output is a ranked list of recommended references. The citation recommending problem can be cast as an IR problem, in which the query is the snippet from an article, and the documents are the cited articles.

We collected 29,353 computer science papers from 1988 to 2010 as follows: We first sample 5,000 papers from the DBLP dataset, and then crawl the full text of these papers as well as the papers cited by them. For each paper, we extract all citation placeholders (places containing citations to other literatures) and the citation contexts. In our experiment, we simply take the sentence of the citation placeholder as the citation context. From the dataset, we extracted 96,873 citation placeholders with at least one cited paper in our corpus (some citation placeholders have more than one cited reference). One citation context can be considered as a query, and the cited papers of the corresponding placeholder can be considered as the relevant documents for the query. We randomly select 200 queries as the test data, and the remaining 96,673 queries with their cited documents as the training data. From the training set, we use the queries (snippets) and the corresponding documents to train a position-aligned translation model.

We can use standard retrieval evaluation measures to evaluate the performance of our retrieval models. In this paper, we employ the standard Mean Average Precision (MAP) measure for the evaluation.

4.2 Experimented Methods

To experiment the position-aligned translation model, we set the passage length as 500, 1000, 2000, 3000, and 4/5 of a passage overlapped with the following passage (if there is any). We denote a setting of passage length L_p as PTM(L_p) (e.g., PTM(2000) is position-aligned translation model with passage length 2000). We use the traditional translation model (TM) and query likelihood language model (LM) as our baselines.

For the smoothing method for the collection language model, we have examined both Dirichlet smoothing method and Jelinek-Mercer smoothing method for document model smoothing with the collection model, and found that Dirichlet smoothing method performs consistently better for this task. Therefore, in

this section, we present and discuss the results of the Dirichlet prior smoothing method.

Since the goal of this paper is to evaluate the effectiveness of retrieval methods with only textual features, we do not compare with other existing citation recommending methods [20,21,22,23,24], which use other features such as fields and the link structure; nor do we use pseudo-relevance feedback, which is commonly used to enhance the retrieval effectiveness.

4.3 Results

The results of the experiments for different alignment methods are presented in Table 1. Each row in the table presents the results of a version of translation model (the number in the parentheses indicates the length of the passage), and each column presents the results for a specific smoothing method. The cases with $*$ indicate a statistically significant difference between the PTM and the baseline model TM. As we can see, all the results included in the table using a translation model are better than the basic language modeling method.

Table 1. Results for Alignment and Smoothing Methods in Translation Models

Translation Models	MAP
LM	0.4938
TM	0.5829
PTM(500)	0.5868*
PTM(1000)	0.5919*
PTM(2000)	0.5865*
PTM(3000)	0.5844

We can see the position-aligned translation models generally help improve the retrieval performance. From the results, we can see the optimal passage length is about 1000. Actually, passage length selection is a trade-off between the translation precision and coverage. In the position-aligned translation model, we assume that the query is generated by some positions (passages). A smaller passage length restricts the translation of the query terms from a smaller piece of snippet (and thus fewer document terms). This leads to a more focused translation model in which a query term can be translated from less document terms. On the other hand, a longer passage contains more document terms. Thus a query term can be translated from more document terms.

4.4 Smoothing Parameter Tuning

In this section, we examine the impact of different smoothing parameter settings. μ controls the weight of prior from the collection smoothing method (Eq 3), and β controls the weight of self-translation (Eq 4).

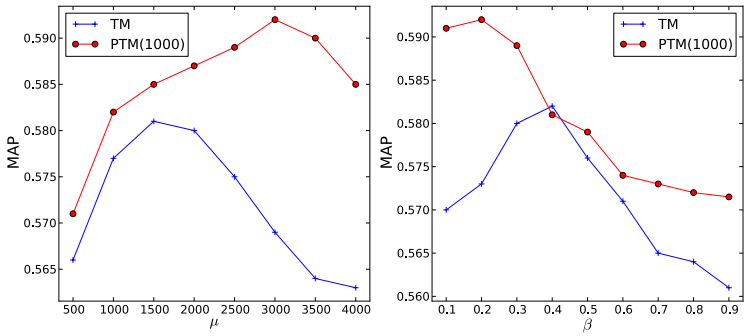

Fig. 2. Smoothing Parameter Tuning

Due to the space constraints, we only present the results for the positional-aligned translation model PTM(1000) and the translation model without alignment TM, shown in Fig 2. We examine different values of these two parameters and present the optimal result for each assignment of a single parameter.

The parameter β smooths translation-based query term likelihood with the document language model. We can see that the retrieval model performs better at small β value (around 0.2) for PTM(1000), and it performs better at relative larger β value (around 0.4-0.6) for TM. The β value determines the importance of self-translation take place in the retrieval. A smaller optimal β value for PTM(1000) means that the translation model trained by position alignment is more accurate, and relative small amount of document language model smoothing is required.

For the Dirichlet collection smoothing parameter μ, we can find PTM translation model needs more collection smoothing than TM translation model since the optimal μ value (2500-3500) for PTM translation model is larger than that of TM translation model (1000-2000). As we discussed earlier, the small number of candidate document terms leads to relatively large posterior probability $p(t_D|t_Q, q, d)$. For this reason, some popular query terms are likely assigned with a larger translation probability according to Eq 9. It leads to a larger generative likelihood of a popular query term $p_{\text{TM}}(t_Q|d)$ by PTM translation model, and it needs a larger collection smoothing.

4.5 Effect of the Size of Training Dataset

The training dataset plays an important role in the estimation of translation model. In this section, we investigate the effect of different training dataset sizes. We keep the query dataset described in Section 4.1, and randomly select a subset of the remaining query-document pairs as the new training dataset. Namely, we select four small training datasets by randomly sampling (without replacement) 5000, 10000, 20000, and 50000 query-document pairs for our experiment. The results of different training datasets are presented in Table 2.

Table 2. Results on Different Sizes of Training Dataset

Size	Model				
	TM	PTM(500)	PTM(1000)	PTM(2000)	PTM(3000)
5000	0.5711	0.5705	0.5726	**0.5798**	0.5741
10000	0.5722	0.5711	0.5733	**0.5809**	0.5797
20000	0.5736	0.5779	**0.5865**	0.5834	0.5763
50000	0.5823	0.5857	**0.5899**	0.5889	0.5842

From the results, it is clear that the performance of different methods increases when the training dataset becomes larger. Again, we find that in most cases, the position-aligned translation model perform better than the traditional translation model.

Another interesting observation is that the optimal passage length becomes smaller when a larger training dataset is available. The optimal passage length becomes 2000 when there is only a limited training dataset (i.e., 5000 and 10000) available. It can be interpreted as a trade-off between translation precision and coverage. When the training dataset is smaller, the coverage of the translation model is quite low, so it can be improved more from adding more translation relations. The position-aligned translation model with longer passages can consider passages containing more terms, so it can help expand the translation model even with a relatively small training set. However, as the training dataset becomes larger, the translation model has collected important term relations, and the translation precision becomes more important than the coverage. Thus we can benefit more from aligning the query to a short but more relevant snippet of the document.

5 Conclusion and Future Work

We have studied the problem of improving the performance of citation recommendation with translation model trained on noisy data. We propose a position-aligned translation model to make the estimated translation probabilities more accurate and more robust to noise. It attempts to align the query to a highly relevant position in the document, and the translation probability is estimated by aligning the query terms and the surrounding terms of the highly relevant position. The experiment shows that this method can help estimate more accurate translation probabilities, and the model trained in this way is more helpful for the retrieval task. It is especially useful when the dataset is relative large, since in this case, the retrieval effectiveness is more affected by translation precision rather than coverage.

There are several interesting future directions to explore further. First, there are some other alternatives of position specific language model rather than the arbitrary passage language model used in our experiment. For example, we can use the positional language model [13], in which the terms are weighted according to the distances to the position. Second, the parameter selection in the model

is important, and the optimal parameter selection depends on the property of the training dataset and the document collection in the retrieval phase. One interesting problem is to determine the parameters automatically according to the dataset. Third, in many IR applications, the document has many fields (e.g., title, abstract, anchor text, query in the clickthrough, etc.), each of which can be considered as a special position or passage, but there can be some prior importance for each field. So it is interesting to integrate the field information with the position information to train a more informative translation model. Finally, we can further examine the effectiveness of the positional-aligned translation model in other information retrieval applications such as question answering and ad hoc Web retrieval.

References

1. Xu, J., Croft, W.B.: Query expansion using local and global document analysis. In: SIGIR 1996, pp. 4–11 (1996)
2. Karimzadehgan, M., Zhai, C.: Estimation of statistical translation models based on mutual information for ad hoc information retrieval. In: Proceeding of SIGIR 2010, pp. 323–330 (2010)
3. Berger, A., Lafferty, J.: Information retrieval as statistical translation. In: Proceedings of SIGIR 1999, pp. 222–229 (1999)
4. Nie, J.Y., Simard, M., Isabelle, P., Durand, R.: Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In: Proceedings SIGIR 1999, pp. 74–81 (1999)
5. Lavrenko, V., Choquette, M., Croft, W.B.: Cross-lingual relevance models. In: Proceedings of SIGIR 2002, pp. 175–182 (2002)
6. Xue, X., Jeon, J., Croft, W.B.: Retrieval models for question and answer archives. In: Proceeding of SIGIR 2008, pp. 475–482 (2008)
7. Murdock, V., Croft, W.B.: A translation model for sentence retrieval. In: Proceedings of HLT 2005, pp. 684–691. Association for Computational Linguistics, Stroudsburg (2005)
8. Gao, J., He, X., Nie, J.Y.: Clickthrough-based translation models for web search: from word models to phrase models. In: Proceedings of CIKM 2010, pp. 1139–1148 (2010)
9. Metzler, D., Bernstein, Y., Croft, W.B., Moffat, A., Zobel, J.: Similarity measures for tracking information flow. In: CIKM 2005, pp. 517–524 (2005)
10. Lu, Y., He, J., Shan, D., Yan, H.: Recommending citations with translation model. In: Proceedings of CIKM 2011, pp. 2017–2020 (2011)
11. Fung, P., Cheung, P.: Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and em. In: Proceedings of EMNLP 2004, pp. 57–63 (2004)
12. Zhao, B., Vogel, S.: Adaptive parallel sentences mining from web bilingual news collection. In: Proceedings of ICDM 2002, p. 745 (2002)
13. Lv, Y., Zhai, C.: Positional language models for information retrieval. In: Proceedings of SIGIR 2009, pp. 299–306 (2009)
14. Wang, M., Si, L.: Discriminative probabilistic models for passage based retrieval. In: Proceedings of SIGIR 2008, pp. 419–426. ACM, New York (2008)
15. Hearst, M.A., Plaunt, C.: Subtopic structuring for full-length document access. In: Proceedings of SIGIR 2003, pp. 59–68 (1993)

16. Bestgen, Y.: Improving text segmentation using latent semantic analysis: A reanalysis of Choi, Wiemer-hastings, and Moore (2001); Comput. Linguist. 32, 5–12 (2006)
17. Misra, H., Yvon, F., Cappé, O., Jose, J.: Text segmentation: A topic modeling perspective. Inf. Process. Manage. 47, 528–544 (2011)
18. Callan, J.P.: Passage-level evidence in document retrieval. In: Proceedings SIGIR 1994, pp. 302–310 (1994)
19. Zobel, J., Moffat, A., Wilkinson, R., Sacks-Davis, R.: Efficient retrieval of partial documents. Inf. Process. Manage. 31, 361–377 (1995)
20. He, Q., Pei, J., Kifer, D., Mitra, P., Giles, L.: Context-aware citation recommendation. In: Proceedings of WWW 2010, pp. 421–430 (2010)
21. McNee, S.M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S.K., Rashid, A.M., Konstan, J.A., Riedl, J.: On the recommending of citations for research papers. In: Proceedings of CSCW 2002, pp. 116–125 (2002)
22. Zhou, D., Zhu, S., Yu, K., Song, X., Tseng, B.L., Zha, H., Giles, C.L.: Learning multiple graphs for document recommendations. In: Proceeding of WWW 2008, pp. 141–150 (2008)
23. Nascimento, C., Laender, A.H., da Silva, A.S., Gonçalves, M.A.: A source independent framework for research paper recommendation. In: Proceedings of JCDL 2011, pp. 297–306 (2011)
24. Kodakateri Pudhiyaveetil, A., Gauch, S., Luong, H., Eno, J.: Conceptual recommender system for citeseerx. In: Proceedings of RecSys 2009, pp. 241–244 (2009)