

Cross-language Context-Aware Citation Recommendation in Scientific Articles

Xuwei Tang, Xiaojun Wan* and Xun Zhang

Institute of Computer Science and Technology, Peking University, Beijing 100871, China

Key Laboratory of Computational Linguistics (Peking University), MOE, China

{tangxuwei, wanxiaojun, zhangxunah}@pku.edu.cn

ABSTRACT

Adequacy of citations is very important for a scientific paper. However, it is not an easy job to find appropriate citations for a given context, especially for citations in different languages. In this paper, we define a novel task of cross-language context-aware citation recommendation, which aims at recommending English citations for a given context of the place where a citation is made in a Chinese paper. This task is very challenging because the contexts and citations are written in different languages and there exists a language gap when matching them. To tackle this problem, we propose the bilingual context-citation embedding algorithm (i.e. BLSRec-I), which can learn a low-dimensional joint embedding space for both contexts and citations. Moreover, two advanced algorithms named BLSRec-II and BLSRec-III are proposed by enhancing BLSRec-I with translation results and abstract information, respectively. We evaluate the proposed methods based on a real dataset that contains Chinese contexts and English citations. The results demonstrate that our proposed algorithms can outperform a few baselines and the BLSRec-II and BLSRec-III methods can outperform the BLSRec-I method.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – *retrieval models*

Keywords

Cross-language Citation Recommendation, Context-aware Citation Recommendation, Bilingual Information.

1. INTRODUCTION

Adequacy of citations is very important for a scientific paper. Firstly, it demonstrates that authors are aware of relevant prior research by citing related papers. Secondly, the space of a paper is limited, authors cannot put all the detail into a paper, so adequate citations can provide more background information and make a paper easier to understand. However, finding appropriate citations is not an easy task for researchers, particularly for the junior ones. There are millions of papers out there and lots of new papers are published every year. Even in a specific research field, researchers do not have enough time to read all the related papers. What's worse, potential citations are probably written in a different language.

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

Copyright 2014 ACM 978-1-4503-2257-7/14/07 ...\$15.00.

<http://dx.doi.org/10.1145/2600428.2609564>

In order to address the above problems and alleviate researchers' burden of finding citations, a few researches on citation recommendation have been conducted in recent years [12, 13, 14]. Generally speaking, the citation recommendation researches can be coarsely categorized as global citation recommendation and local citation recommendation. Global citation recommendation aims to recommend a list of citations (references) for a given query paper, while local citation recommendation aims to recommend citations for specific context of each place where a citation should be made in a paper, which is also called context-aware citation recommendation and is the focus of this study.

Text of a Chinese paper and its English translation	协同过滤算法是目前最受欢迎的推荐技术，它利用用户爱好之间的相似性来进行推荐 [3]，不依赖于物品的实际内容，而是需要用户对物品的偏好信息，通常以评价或者打分的形式 [2]。然而这种经典的协同过滤方法不能直接应用于社交网络的好友推荐根据用户过去喜欢的物品，为用户推荐和他过去喜欢的相似的物品 [4]。基于内容相似性的方法可以很好地应用在社交网络的好友推荐。(English version: Collaborated filtering is the most popular algorithm for recommender systems, which takes advantage of the similarity of users' interest [3]. It is based on users' preference rather than contents of items. Users' preferences are denoted as ratings [2]. However, the classical collaborative filtering algorithm cannot be applied to friends recommendation in social network,.... Based on what user liked in the past, we can recommend similar items to him [4]. Content similarity based algorithms are suitable for friends recommendation in social network)
English Citations	[2] Analysis of recommendation algorithms for e-commerce. (2000) [3] Amazon.com recommendations: Item-to-item collaborative filtering. (2003). [4] Content-based recommendation systems (2007)

Figure 1: An example of a snippet of a Chinese paper and its citations

Most existing citation recommendation algorithms work in a single language, which means the contexts and citations are written in the same language. However, cross-language citations

are very common in scientific papers, especially the non-English papers (e.g. Chinese papers). For example, Figure 1 is a snippet of text and several example citations in a Chinese paper, and the translated English text for the Chinese text is placed below the Chinese text. As it shows, this paper is written in Chinese and some of the citations are written in English. The citation placeholders in the text are marked as “[]”, we consider the words surrounding the placeholder *citation context* (*context* for short). The Chinese contexts usually describe the main content of the corresponding English citations.

In this paper, we define a novel task, *cross-language context-aware citation recommendation*, which aims at recommending English citations for a given Chinese context. There are several challenges for this task.

Firstly, we cannot compute the relevance between Chinese contexts and English citations directly, because they are in different languages. One reasonable method is to use machine translation (MT) to translate contexts or citations, and then the problem is reduced to the monolingual context-aware citation recommendation. However, machine translation could be not accurate. As a result, this straightforward method would suffer from noises brought by machine translation.

Another straightforward method is to find the most similar contexts to a given context and recommend the corresponding citations. However, it is necessary to compare the contexts in the whole dataset with the given context, which is computationally expensive. This method also suffers from the cold-start problem, if one paper has not been cited by others, it can never be recommended. More seriously, different authors usually use different words and expressions in their contexts when they cite a same paper, which results in poor performance for existing similarity metrics.

Last but not the least, as lots of new papers are published every year, a good algorithm should have the capability to be incrementally updated.

In order to address the above challenges, we propose the bilingual context-citation embedding algorithm (i.e. BLSRec-I), which can learn a low-dimensional joint embedding space for both contexts and citations. Moreover, two advanced algorithms named BLSRec-II and BLSRec-III are proposed by enhancing BLSRec-I with translation results and abstract information, respectively. We evaluate the proposed methods based on a real dataset that contains Chinese contexts and English citations. The results demonstrate that our proposed algorithms can outperform a few baselines and the BLSRec-II and BLSRec-III methods can outperform the BLSRec-I method. Our contributions are summarized as follows:

- 1) We propose a novel task of cross-language citation recommendation in order to help researchers find cross-language citations.
- 2) We propose three methods to address this new task, and BLSRec-I is a bilingual context-citation embedding model, which can capture the latent semantics of contexts and citations. The training method is an on-line algorithm, which is suitable for incremental updating. BLSRec-II and BLSRec-III enhance BLSRec-I with translation results and abstract information.
- 3) Experiments are performed on a real dataset, and the results show the efficacy of our proposed methods.

The rest of this paper is organized as follow. We discuss related work in section 2. The problem is defined in section 3. Our proposed methods are formulated and described in section 4. The details of the experiments are discussed in section 5. In section 6 we conclude this paper and discuss some future work.

2. RELATED WORK

2.1 Document Recommendation

There are several efforts on recommending scientific documents. Two common approaches for paper recommendation are content-based recommendation and collaborative recommendation.

Content-based recommendation infers users’ interests based upon the contents of the papers and a profile of the user’s interests. Chandrasekaran et al. [1] present a profile-based method to recommend papers to users according to their profiles stored in CiteSeer. The profile-based recommendation compares candidate item’s content with user’s profile. Sugiyama et al. [2] describe a method that models user’s profile based on past publications and its neighboring papers such as citation and reference papers.

Collaborative recommendation algorithms infer users’ interests based on preference list, or partial list of citation. In particular, for a document d , they extract the citation list from the document. Give a proportion of citations in a list, the goal is to recover the rest of citations. For example, McNee et al. [4] explore the use of collaborative filtering to recommend research papers. They build the rating matrix based on the citation web. Zhou et al. [5] apply multiple graphs model on the document recommendation problem. Their model jointly combines multiple graphs including citation, author and venue information.

Torres et al. [8] introduce a hybrid recommender algorithm which combines collaborative filtering and content-based filtering to recommend research papers to users. Wang et al. [9] introduce Collaborative Topic Modeling (CTR) for recommending scientific articles; their approach combines the merits of collaborative filtering and probabilistic topic modeling [10]. Gori et al. [11] propose a scholarly paper recommendation algorithm based on the citation graph and random-walker properties.

Most document recommendation algorithms need some additional information, such as users’ profile, preference list, or partial list of citation. In our method, we only require a context or a context with an abstract as a query to recommend a list of citation candidates, which alleviates the users’ burden for providing additional profiles.

2.2 Topic-based Citation Link Prediction

Topic model is a powerful tool to discover “topics” in a document set. This model is also used to predict citations for bibliographies. Tang and Zhang [7] present a study of topic-based citation recommendation. Two-layer restricted Boltzmann machine model is used for modeling paper contents and citation relationships. Nallapati et al [15] jointly model the text and the citation relationship under a framework of topic model. The proposed model Pair-Link-LDA is too expensive to scale to large digital libraries. They also introduce a simpler model, Link-PLSA-LDA, where citations are modeled as a sample from a probability distribution associated with a topic. Kataria et al [16] extend Link-PLSA-LDA by combing context information. They assume that context information can help in improving the topic identification for words and document. Kataria et al [17] propose a method to model the influence of cited authors along with the interests of citing authors. They hypothesize that contexts provide extra topical information about cited authors. Jointly modeling context,

cited author and interest of citing authors can learn better topic distribution.

Overall, Models under the framework of topic model require long training process. When new documents come, the models have to be retrained.

2.3 Citation Recommendation

2.3.1 Local Citation Recommendation

Local citation recommendation aims to recommend citations for specific context of each place where a citation should be made in a paper, which is also called context-aware citation recommendation. He et al. [12] use contexts of a citation and its abstract to represent a paper. Then they propose a probabilistic model to compute the relevance score. The shortcoming of the method is that it requires citation contexts of a paper, however, lots of papers do not have any citation. What's more, citation contexts of a paper probably have similar semantics, but different words. This model cannot capture such similar semantics. Lu et al. [13] regard context and citation as parallel corpus and take advantage of translation model to bridge the gap between citations and context. They translate one word in context to one word in citation. Huang et al. [14] regard a paper as a new "word" in another language, and they estimate the probability of translating a context to a citation. The aforementioned work in [7] formalizes citation recommendation as a topic discovery task. Citation candidates are ranked based on their topic similarity to the context. However, all the above researches on local citation recommendation focus on a monolingual setting, and cross-language local citation recommendation has not been investigated yet. In this study, we aim to propose new models for addressing the new task, and our proposed models are related to Supervised Semantic Indexing (SSI) [19]. SSI can be used for document-document or query-document retrieval. However, their model requires that queries and documents are in a same feature space and it has not been attempted for cross-language tasks.

2.3.2 Global Citation Recommendation

Global citation recommendation aims to recommend a list of citations for a given query paper. Strohman, et al. [6] find that simple text similarity computation is not enough for global citation recommendation. They take paper content and author information into their evaluation model and use bibliography similarity and Katz centrality measurement to rank citation candidates. Meng et al [3] propose a unified graph-based model with random walk that incorporates various types of information (e.g. content, authorship, citation and collaboration network), which can provide personal global citation recommendation. He et al. [18] seek to not only recommend citations, but also identify candidate locations in a query manuscript where citations are needed.

2.4 Cross-Language Retrieval

The cross-language citation recommendation task is related to the cross-language retrieval task [20, 21, 22]. Cross-language retrieval addresses the situation where the query, that a user presents to an IR system, is not in the same language as the documents being searched. Translation-based models are common for this task. Some models require some resources such as a bilingual dictionary, machine translation tools, or a parallel corpora to map terms in source language to terms in target language ([20], [22], [23], [24]). Ballesteros et al [25] describe a phrasal translation and query expansion techniques for cross-language information retrieval. Dumais et al. [26] propose a method which constructs a multi-lingual semantic space using Latent Semantic Indexing (LSI

[27]). Potthast et al. [28] introduce CL-ESA, which exploits the multilingual alignment of Wikipedia.

3. PROBLEM DEFINITION

In this section, we define concepts and problems used in this paper.

DEFINITION 3.1 *Let d represent a scientific document written in Chinese. In document d , a **context** is a bag of words, which surround a citation placeholder. The citation placeholder includes one or several English citation papers (short for **citation**). We use the title and abstract of the corresponding English paper to denote the content of a citation.*

The context with a few Chinese sentences is considered a query, and a citation recommendation system is required to return a list of English citation candidates. We define this new task as follows:

DEFINITION 3.2 (cross-language context-aware citation recommendation) *Given a few Chinese sentences as a context in paper d , the task aims to return a ranked list of English citations. The citations in the ranked list are recommended to users and users can select one or more citations for the place of the context in d .*

Besides the given context, a user can provide the Chinese abstract of document d as an input, and then the system is required to return a few citation candidates. The task can be defined as follows:

DEFINITION 3.3 (cross-language context-aware citation recommendation with abstract) *Given a Chinese context and a Chinese abstract, the task aims to return a ranked list of English citations. The citations in the ranked list are recommended to users as candidates.*

Both the above tasks are very helpful when researchers write papers. The task of cross-language context-aware citation recommendation task requires only a few Chinese sentences as a context in a Chinese paper to recommend English citation candidates, without using any global information of the Chinese paper. The task of cross-language context-aware citation recommendation with abstract corresponds to the situation that when the title and the abstract of the Chinese paper have been written down, and researchers want to get some English citation candidates for a specific context in the paper.

Since there exists a language gap between the Chinese contexts and English citations, machine translation is usually used for eliminating the language gap and the Chinese contexts and the English citations can be matched after translation. However, due to the unsatisfactory machine translation quality and the different expressions and word usages in different papers written by different researchers, existing content similarity based methods usually result in poor performances. In order to address the big challenges mentioned in Section 1, we have to propose more suitable methods in order to achieve satisfactory performances.

Note that though in this study we focus on Chinese-to-English cross-language citation recommendation, our system framework and proposed methods are language independent. That's to say, our system and methods can also be used for recommending cross-language citations between any other pair of languages.

4. OUR PROPOSED APPROACHES

In this section, firstly, we introduce a joint embedding model proposed by Weston et al [30]. We then extend this model to

tackle the cross-language citation-aware recommendation problem. Furthermore, we enhance our model by introducing translated corpus and abstract information.

4.1 Background

Weston et al [30] propose a joint word-image embedding model to find annotations for images. In the model, bags-of-visual terms are used to represent images. We will introduce this model, and then extend it to address our problem.

It starts with a representation of an image $q \in R^k$ and a representation of an annotation term $i \in \{1, \dots, Y\}$, which indicates a possible annotation. Then the model tries to learn a mapping from the image feature space to a joint space R^n :

$$\Phi_I(q): R^k \rightarrow R^n$$

The model also tries to jointly learn a mapping for annotations:

$$\Phi_A(i): \{1, \dots, Y\} \rightarrow R^n$$

These mappings are chosen to be linear, i.e. $\Phi_I(q) = q^T W$ and $\Phi_A(i) = F_i$, where $F_i \in R^n$ is the i^{th} row of a $Y \times n$ matrix, W is a $k \times n$ matrix. W and F will be learnt in the training process. n is the dimension of the embedding space where annotations and images will be represented.

We then use the following function to compute the relevance score between a query and an annotation term:

$$f(q, i) = \Phi_I(q) \Phi_A(i)^T = q^T W F_i^T$$

The goal is to rank the possible annotations according to the relevance scores such that the highest annotations best match the semantic content of the given image.

In the model, finding annotations for images is reduced to find some annotation terms for the given bags-of-visual terms, which is similar to our cross-language citation recommendation problem.

4.2 BLSRec-I: Using Bilingual Latent Semantics for Recommendation

In this section, we will extend the above joint word-image embedding model to address our problem. Table 1 summarizes the notations used in our models. In the cross-language citation recommendation, we start with the TF-IDF feature vector of a Chinese context $q_c \in R^k$ and the TF-IDF feature vector of an English citation $d_e \in R^g$. Then we learn a mapping from the TF-IDF feature space to the joint space R^n :

$$\Phi_Q(q_c): R^k \rightarrow R^n$$

We also try to learn a mapping for citations jointly:

$$\Phi_D(d_e): R^g \rightarrow R^n$$

We adopt linear mappings for both Φ_Q and Φ_D , i.e. $\Phi_Q(q_c) = q_c^T W$ and $\Phi_D(d_e) = d_e^T F$, where W is a $k \times n$ parameter matrix, and F is a $g \times n$ parameter matrix. W , F will be learnt in the training process. n is the dimension of the embedding space where contexts and citations will be represented (this is a hyper-parameter, typically $n \ll k$ and $n \ll g$).

Our goal is to rank the possible citations such that the highest citation best matches a given context. The score function for

measuring the relevance between a context and a citation is modeled as follows:

$$f(q_c, d_e) = \Phi_Q(q_c) \Phi_D(d_e)^T = q_c^T W (d_e^T F)^T \quad (1)$$

The intuitive explanation of our model is that it maps the contexts (via $q_c^T W$) and the citations (via $d_e^T F$) into the same low dimensional space for computing their similarity. The contexts and the citations are mapped by different functions, hence our model is suitable if contexts and citations are in different languages.

Table 1: Main notations in this paper

SYMBOL	DESCRIPTION
$q_c \in R^k$	TF-IDF feature vector of a Chinese context
$d_e \in R^g$	TF-IDF feature vector of an English citation
$u_c \in R^k$	TF-IDF feature vector of a Chinese abstract
$q_e \in R^g$	TF-IDF feature vector of an English context (translated from Chinese context)
$d_c \in R^k$	TF-IDF feature vector of a Chinese citation (translated from English citation)
$u_e \in R^g$	TF-IDF feature vector of an English abstract (translated from Chinese abstract)
I_c, I_e	identity matrices
n	dimension of the joint space
W_e, F_e, F, W_e^u	$g \times n$ parameter matrices
$W_c, F_c, W, W_m^u, W_c^u$	$k \times n$ parameter matrices
Q, A, D	context set, abstract set, citation set
T	Training set

One basic assumption is that if a context cites several papers, these papers should have similar topic and be mapped into similar low-dimensional vectors. Given a context, our model aims to give a high score to any positive citation, even if the positive ones use synonymy and polysemy to describe a same topic. Moreover, if a paper has been cited in several different contexts, these contexts should be similar and be mapped into similar low-dimensional vectors. In other words, our model can capture the semantics of citations by embedding synonymy and polysemy into a similar position in the low dimensional space. For the same reason, the semantics in the contexts also can be captured by our model.

4.3 BLSRec-II: Enhancing BLSRec-I with Translated Corpus.

In this section, we try to make use of the translated corpus to enhance BLSRec-I. We can use machine translation to translate contexts and citations and get two views (Chinese-Chinese, English-English). Let $q_e \in R^g$ be the TF-IDF feature vector of the

English version of q_c , $d_e \in R^k$ be the TF-IDF feature vector of the Chinese version of d_e .

For monolingual context and citations (Chinese-Chinese or English-English), we adopt Supervised Semantic Index (SSI) [19] to model their relevance score. Here, we use English corpus as an example:

$$f(q_e, d_e) = q_e^T (W_e F_e^T + I_e) d_e$$

where W_e , F_e are $g \times n$ matrices, I_e is an identity matrix. W_e , F_e will be learnt in the training process. $W_e F_e^T + I_e$ is a diagonal persevering approximation of a $g \times g$ dense matrix. The latent semantics of citations and contexts can be embedded in the dense matrix.

We can consider a joint model that takes into account both cross-language data and monolingual data. In this case, our score function can be written as follows:

$$f(q_c, d_e) = q_c^T W (d_e^T F)^T + q_e^T (W_e F_e^T + I_e) d_e + q_c^T (W_c F_c^T + I_c) d_e \quad (2)$$

where F are $g \times n$ matrices. W_c , F_c , W are $k \times n$ matrices and I_e , I_c are two identity matrices.

In Equation (2), the first term captures the cross-language context-citation relevance. The second term captures the relevance of English contexts to English citations and the last term captures the relevance of Chinese contexts to Chinese citations. The relevance scores are combined in a smooth way.

4.4 BLSRec-III: Enhancing BLSRec-II with Abstract Information

An abstract describes the general idea of a paper, which is very useful for citation recommendation. We propose BLSRec-III to incorporate the abstract information. Let $u_c \in R^k$ be the TF-IDF feature vector of Chinese abstract, $u_e \in R^g$ be the TF-IDF feature vector of the English version of u_c via machine translation. The score function of our proposed method can be written as follows:

$$f(u_c, q_c, d_e) = q_c^T W (d_e^T F)^T + q_e^T (W_e F_e^T + I_e) d_e + q_c^T (W_c F_c^T + I_c) d_e + u_c^T W_m^u (d_e^T F)^T + u_e^T (W_e^u F_e^u + I_e) d_e + u_c^T (W_c^u F_c^u + I_c) d_e \quad (3)$$

where W_e^u is a $g \times n$ matrix. W_m^u , W_c^u are $k \times n$ matrices. W_e^u , W_m^u , W_c^u will be learnt in the training process.

In equation (3), the first three terms capture the context-citation relatedness; the last three terms capture the abstract-citation relatedness.

4.5 Model Training

We now introduce the training methods for BLSRec-I, BLSRec-II and BLSRec-III.

4.5.1 Training BLSRec-I

In the cross-language context-aware citation recommendation task, we are interested in learning a ranking score function $f(q_c, d_e)$ for a given Chinese context q_c and $d_e \in D$, where the top k ranking documents will be presented to the user. Let Q denote the set of Chinese contexts, let D denote the set of English citations.

Suppose we have a set of tuples T (training data), where each tuple contains a Chinese context q_c , a corresponding English citation d_e^+ . For any non-relevant English paper d_e^- , our score function should satisfy the condition $f(q_c, d_e^+) > f(q_c, d_e^-)$, which means d_e^+ should be ranked higher than d_e^- for the give context q_c .

We use Weighted Approximate Rank Pairwise loss (WARP) [30, 31] as our loss function. WARP focuses on the top of the ranked list of the returned citation papers. WARP has been widely used for information retrieval [29] and recommender system [31]. In our setting, the loss function is defined as:

$$errWARP = \sum_{(q_c, d_e^+) \in T} L(rank(q_c, d_e^+)) \quad (4)$$

where $rank(q_c, d_e^+)$ is the margin-based rank of the labeled tuple (q_c, d_e^+) :

$$rank(q_c, d_e^+) = \sum_{d_e^- \neq d_e^+} \Theta(1 + f(q_c, d_e^-) \geq f(q_c, d_e^+)) \quad (5)$$

where Θ is the indicator function. Function L transforms the rank into a loss:

$$L(r) = \sum_{i=1}^r \alpha_i, \text{ with } \alpha_1 \geq \alpha_2 \geq \dots \geq 0 \quad (6)$$

We could choose different a to define different weights of the relative position of the positive example in the ranked list. $a_i = 1/i$ is a common choice which has been used in [31, 33]. $a_i = 1/i$ prefers the top position and decays its weight for lower positions. It was shown experimentally that the choice of $a_i = 1/i$ could lead to state-of-the-art results when we want to optimize *precision at K* for a variety of different values of K .

The loss function can be optimized by stochastic gradient descent (SGD) following [30, 31] which updates the parameters for samples of each random draw. However, to perform the SGD, we have to compute $rank(q_c, d_e^+)$ for all $(q_c, d_e^+) \in T$ and $f(q_c, d_e)$ for all $d_e \in D$, which is too expensive for large dataset.

We solve the problem with an approximating approach: for a given positive label, we sample negative labels until we find a violating label. The total number of violating labels in D is $v = rank(q_c, d_e^+)$, the number of the trials in our sampling process is N_v which follows a geometric distribution of parameter $\frac{v}{|D|-1}$. Thus $v = \frac{|D|-1}{E[N_v]}$. So the value of Equation (5) can be approximated by:

$$rank(q_c, d_e^+) \approx \frac{|D|-1}{N} \quad (7)$$

where $\lfloor \cdot \rfloor$ is the floor function, N is the number of trials in the sampling step, and $|D|$ is the number of citations in the database.

In this method, many parameters need to be learnt. We constrain the parameters with the following norm:

$$\begin{aligned} \|W_{(i)}\|_2 &\leq C, i = 1, \dots, k \\ \|F_{(i)}\|_2 &\leq C, i = 1, \dots, g \end{aligned} \quad (8)$$

Where C is a parameter and $C > 0$. Equation 8 acts as a regularizer in the same way as is used in lasso [32]. Pseudo code for training with WARP loss is described in Algorithm 1. We choose a fixed learning rate λ . The algorithm is terminated when the performance no longer improves on a validation set.

Algorithm 1 Online WARP Loss Optimization

Input: labeled data T

Repeat

 Pick a labeled example (q_c, d_e^+) randomly from T

 Let $f^+ = f(q_c, d_e^+)$

 Set $N = 0$

 Repeat

 Pick a citation candidate $d_e \in \{D \setminus d_e^+\}$

 Let $f^- = f(q_c, d_e)$

$N = N + 1$

 Until $f^- > f^+ - 1$ or $N \geq |D| - 1$

 If $f^- > f^+ - 1$ then

 Make a gradient step to minimize:

$$L\left(\left[\frac{|D|-1}{N}\right]\right)(1 - f^+ + f^-)$$

 Project weights to satisfy the constraints (8)

 End if

Until performance on validation set does not improve.

4.5.2 Training BLSRec-II and BLSRec-III

For BLSRec-II, most steps of the training process are similar to the steps discussed in section 4.5.1. There are two places which should be modified. The first one is replacing equation (1) with equation (2). The second one is adding the following regularizers into equation (8):

$$\begin{aligned} \|W_{c(i)}\|_2 &\leq C, i = 1, \dots, k \\ \|F_{c(i)}\|_2 &\leq C, i = 1, \dots, k \\ \|W_{e(i)}\|_2 &\leq C, i = 1, \dots, g \\ \|F_{e(i)}\|_2 &\leq C, i = 1, \dots, g \end{aligned} \quad (9)$$

For BLSRec-III, most steps of the training process are similar to the steps discussed in section 4.5.1(BLSRec-I). Based on BLSRec-I, There are two places which should be modified. The first one is replacing equation (1) with equation (3). The second one is adding the following regularizers into equation (8):

$$\begin{aligned} \|W_{c(i)}^u\|_2 &\leq C, i = 1, \dots, k \\ \|W_{m(i)}^u\|_2 &\leq C, i = 1, \dots, k \\ \|W_{e(i)}^u\|_2 &\leq C, i = 1, \dots, g \end{aligned} \quad (10)$$

In our experiments, $n = 100$, $\lambda = 0.01$, the parameter matrices are initialized at random with mean 0, standard deviation $\frac{1}{\sqrt{n}}$.

5. EXPERIMENT

5.1 Dataset

There is no standard benchmark dataset for Chinese-to-English cross-language citation recommendation. So we built our dataset

based on a Chinese Journal - Chinese Journal of Computers¹. This journal is a top-ranking Chinese journal in China, and the papers published in this journal focus on research topics in computer science and technology. Since most research findings in computer science and technology have been published in English, the Chinese papers in this journal usually cite a considerable portion of English papers.

We collected the papers published in this journal from year 2002 to 2012. Then we extracted titles, abstracts, citation contexts and its corresponding citations from the papers. We only considered the English citations in this study. Following [13, 17], three sentences around a citation placeholder were extracted as the context of the citation (including the sentence containing the citation placeholder and the previous and following sentences). And we looked up the titles of the English citations by using the Microsoft academic search service² and got the corresponding abstract information. We removed the citations which did not have abstract information, and combined the abstract and title texts to represent the content of a citation. Note that the contexts are in Chinese, and the contents of citations are in English. We used Google Translation³ to translate the Chinese contexts into English, and translate the content of English citations into Chinese. We used a Chinese segmentation tool⁴ to segment Chinese corpus and removed stop words and the words which appeared only one time.

After preprocessing, our dataset contains 2061 Chinese papers, 30912 context-citation pairs, 21735 Chinese contexts, 17693 English citations. There are 35368 unique words in English corpus and 43747 unique words in Chinese corpus. All 17693 English citations make up the citation set D , and for each query context, the citations in D are ranked. We split the dataset into five sets randomly, and conducted training and testing on our dataset like 5-fold cross validation. At each time, four sets were used as training set and the remaining one set was used as test set. We further split a small portion of examples in the training dataset as a validation set. The training and testing processes are performed five times and the performance values are then averaged.

5.2 Evaluation Metrics

For the cross-language context-aware citation recommendation task, we are given a context q_c and compute $f(q_c, d_e)$ for all $d_e \in D$ and rank them in decreasing order of $f(q_c, d_e)$. For cross-language context-aware citation recommendation with abstract, the setting is the similar: given an abstract u_c , a context q_c , we compute $f(u_c, q_c, d_e)$ for all $d_e \in D$ and rank them in decreasing order of $f(u_c, q_c, d_e)$. The evaluation metrics aim to assess the positions of the right citations in the ranking list S for each given context. In this study, we adopt three common metrics as follows:

Recall @K: It means the recall score for top K results in the ranking list. It is computed by using Equation (11) for each context and then we average the scores across all the contexts in the test set. R means the number of positive (correct) citations in

¹ <http://cjc.ict.ac.cn/eng/kwjse/>

² <http://datamarket.azure.com/dataset/mrc/microsoftacademic>

³ <http://translate.google.cn/?hl=en#zh-CN/en/>

⁴ <https://github.com/fixsjy/jieba>

top K results and $|C|$ means the total number of correct citations provided by researchers.

$$Recall@K = \frac{R}{|C|} \quad (11)$$

Mean Average Precision (MAP): Recall@K considers the top K ranking results, but it does not consider the exact ranking position. Hence we use **MAP** as a metric to evaluate the performance by considering exact ranking positions. Let T_q be the set of all the test contexts. For a context q_c in T_q , the positive (correct) citation set is C , and the recommendation system returns a list S . Note that we only consider the top 100 results in the ranking list. In other words, $|S|=100$. Let $r \in C$ be a positive citation. If r is in the list S , let $rank(r)$ be its position in the list S , otherwise let $rank(r)=0$. $pos_num(r)$ is the number of the positive citations ranking higher than r . The MAP score is then computed as follows:

$$MAP = \frac{1}{|T_q|} \sum_{q_c \in T_q} \frac{1}{|C|} \sum_{r \in C, rank(r) \neq 0} \frac{pos_num(r) + 1}{rank(r)} \quad (12)$$

Mean Reciprocal Rank (MRR): Usually the first correctly recommended citation is important. So we adopt MRR [34] to evaluate based on the position of the first correct one. Let T_q be the set of the test contexts and $first_rank(q)$ be the position of the first correct citation in the ranking list for $q \in T_q$. The MRR score is then computed as follows:

$$MRR = \frac{1}{|T_q|} \sum_{q \in T_q} \left(\frac{1}{first_rank(q)} \right) \quad (13)$$

5.3 Baselines

We compare our proposed methods with several popular methods as follows:

Similarity-based method: The cosine similarity is a basic method in information retrieval. After we translate the original contexts and citations, we can compute the cosine similarity between each Chinese context and each translated Chinese citation, and compute the cosine similarity between each translated English context and each English citation, and compute the cosine similarity by using both Chinese and English information. We denote them as **Cosine (C)**, **Cosine (E)**, and **Cosine (E and C)** respectively.

Context-aware Relevance Model (CRM) [12]: we apply CRM on both Chinese corpus and English corpus, which are denoted as **CRM (E)** and **CRM (C)** respectively. We follow the setting mentioned in [12].

Translation Model (TM) [13]: Lu et al. [13] proposed a translation model to overcome the language gap between contexts and citation papers. For the context-aware citation recommendation task, we follow the setting in [13]. For context-aware citation recommendation with abstract, we merge abstract and context as a new context. We tune the parameters as the authors suggested in [13].

TM (with translation): To investigate the effect of machine translation, we extend the TM model [13]. We combine the original contexts and citations with the corresponding translated contexts and citations, respectively. Then we apply the TM model [13] on the new dataset.

Citation Translation Model (CTM) [14]: Different from TM [13] which aims to “translate” contexts to content of citations, CTM aims to “translate” contexts to references. We follow the setting mentioned by Huang et al. [14].

Note that all the above baselines can be applied to both the two recommendation tasks: cross-language context-aware citation recommendation and cross-language context-aware citation recommendation with abstract information. For the task with abstract information, the abstracts can be simply combined with the contexts and then used by any of the above baselines. We will compare the performances of the baselines with and without using the abstract information.

5.4 Cross-Language Citation Recommendation Results

5.4.1 Results for Cross-Language Context-aware Citation Recommendation

The comparison results are shown in Table 2 and Figure 2. Our proposed methods, BLSRec-I and BLSRec-II, significantly outperform all the baselines. We can also see that BLSRec-II outperforms BLSRec-I significantly. The standard deviation values of 5-fold cross-validation show the robustness of the proposed methods.

Table 2: Evaluation results for cross-language context-aware citation recommendation

	MAP	MRR
Cosine (E)	0.177±0.002	0.189±0.002
Cosine (C)	0.147±0.003	0.158±0.003
Cosine (E and C)	0.189±0.003	0.200±0.004
CRM (E)	0.192±0.005	0.208±0.005
CRM (C)	0.174±0.003	0.190±0.003
TM	0.213±0.005	0.236±0.004
TM(with Translation)	0.209±0.003	0.232±0.003
CTM	0.198±0.003	0.218±0.004
BLSRec-I	0.297±0.003	0.304±0.003
BLSRec-II	0.331±0.002	0.338±0.003

In more details, our proposed methods, BLSRec-I and BLSRec-II outperform the translation-based models (i.e. TM, TM (with translation information), CTM) significantly on the MAP and MRR metrics and $Recall@N$ when N is below 10, and thus our methods beat them overall. Translation-based models utilize co-occurrence of words to overcome the vocabulary gap between citation contexts and citations. However, the translation-based models also bring noise at the same time, which hurts their performance. Compared to our proposed method, we can see that TM, TM (with translation), CTM show low recall value when N is smaller than 10. In practice, the top 10 returned citations are more important than the rest to users. This is more significant when it comes to the MAP and MRR metrics. Because MAP and MRR focus more on the top ranked items and decay its score quickly as the ranking position becomes larger. Our proposed methods outperform the baselines significantly on both metrics. This is due

to our models' capability of capturing the semantics in contexts and citations.

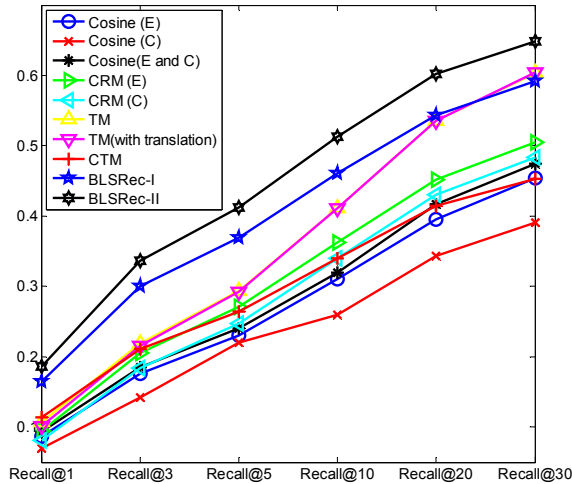


Figure 2: Recall@K for cross-language context-aware citation recommendation

5.4.2 Results for Cross-Language Context-aware Citation Recommendation with Abstract Information

The comparison results are shown in Table 3 and Figure 3. As we can see, for the cross-language context-aware citation recommendation task with abstract information, our proposed methods, BLSRec-I and BLSRec-III, outperform all the baselines significantly over all evaluation metrics.

Comparing Table 3 with Table 2 and comparing Figure 3 with Figure 2, BLSRec-III performs better than BLSRec-II and BLSRec-III is the best-performing one among all methods with or without using abstract information. The reason is that the paper abstract can provide additional background information for a context.

Almost all the methods with abstract information perform better than the corresponding ones utilizing only the context information, except TM and TM (with translation). We believe this is due to the noise that brought by machine translation. In contrast, our proposed method BLSRec-III shows its robustness and it can smoothly combine the abstract information and the context information.

Overall, our proposed methods, BLSRec-I, BLSRec-II, BLSRec-III, are very competitive methods for the cross-language context-aware citation recommendation tasks. Among the three proposed methods, BLSRec-II is more reliable than BLSRec-I, and BLSRec-III is more reliable than BLSRec-II if the abstract information is available.

5.5 Parameters Analysis

We tune the dimension n from 50 to 300 to observe its influence on recommendation performance. The larger n is, the more flexibility the model has. However, as n goes up, the model complexity goes up as well. Hence we need to find an appropriate n for our problem.

Figures 4 and 5 demonstrate the influence of parameter n on BLSRec-II and BLRec-III over different metrics. As we can see, when n goes up from 50 to 100, the performances of both methods have significant improvement. When n is larger than 200, the performances almost do not change any more.

Table 3: Evaluation results for cross-language context-aware citation recommendation with abstract

	MAP	MRR
Cosine (E)	0.188±0.003	0.199±0.003
Cosine (C)	0.151±0.002	0.162±0.002
Cosine (E and C)	0.191±0.003	0.201±0.003
CRM (E)	0.271±0.005	0.289±0.005
CRM (C)	0.261±0.003	0.279±0.005
TM	0.203±0.002	0.225±0.002
TM (with Translation)	0.196±0.002	0.216±0.002
CTM	0.273±0.002	0.295±0.002
BLSRec-I	0.347±0.01	0.350±0.01
BLSRec-III	0.374±0.005	0.375±0.005

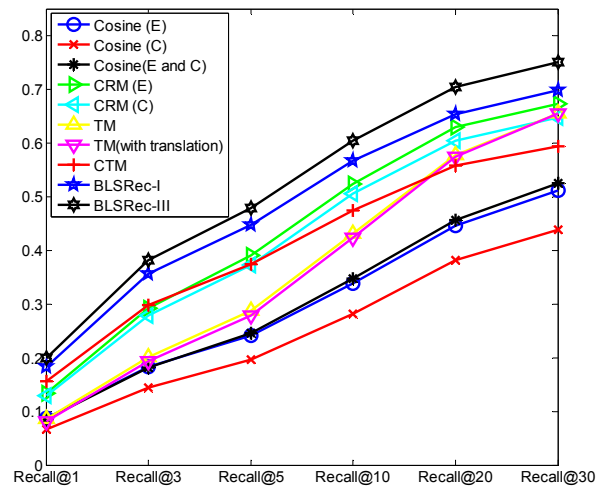


Figure 3: Recall@K for cross-language context-aware citation recommendation with abstract

5.6 Efficiency of the System

Our proposed methods are very efficient in practice. Contexts and Citations can be mapped into the low dimension space. Hence, each context or citation is associated with a low dimensional vector, respectively. Measuring the relevance score between a citation and a context is reduced to dot product. Additionally, our methods are trained by an on-line learning algorithm. It is convenient to update incrementally when newly published papers arrive.

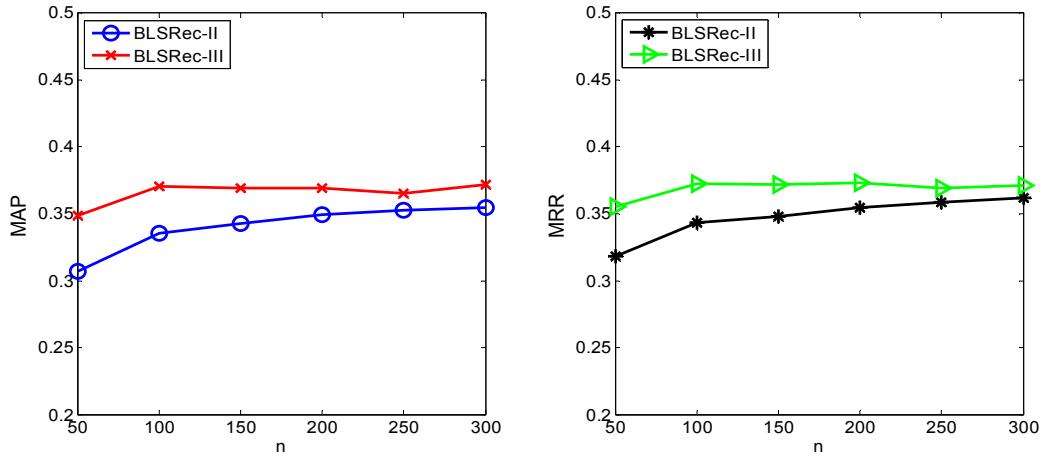


Figure 4: MAP and MRR comparison by varying dimension n .

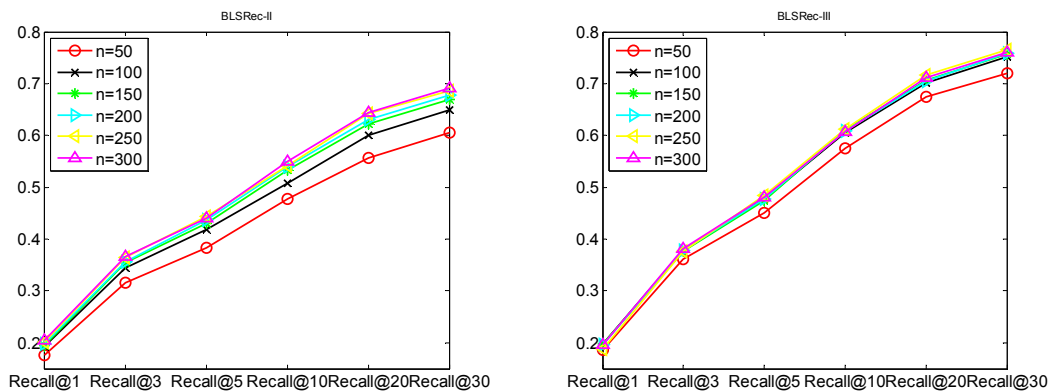


Figure 5: Recall@K comparison by varying dimension n .

6. CONCLUSION AND FUTURE WORK

In this paper, we define a novel task called cross-language context-aware citation recommendation. This task is very challenging because the contexts and citations are written in different languages and there exists a language gap when matching them. To tackle this problem, we propose the bilingual context-citation embedding algorithm (i.e. BLSRec-I), which can learn a low-dimensional joint embedding space for both contexts and citations. Moreover, two advanced algorithms named BLSRec-II and BLSRec-III are proposed by enhancing BLSRec-I with translation results and abstract information, respectively. We evaluate the proposed methods based on a real dataset that contains Chinese contexts and English citations. The results demonstrate that our proposed algorithms can outperform a few baselines and the two advanced algorithms are more reliable than BLSRec-I.

In the future, we will try to simultaneously recommend both Chinese citations and English citations in a Chinese paper, which is a more challenging task. We will also test our proposed methods in other language pairs to show their robustness.

Furthermore, we will implement a real citation recommendation system based on the proposed methods in this paper and evaluate the system in practice through user studies.

7. ACKNOWLEDGMENTS

This work was supported by NSFC (61170166, 61331011), Beijing Nova Program (2008B03) and National High-Tech R&D Program (2012AA011101). We thank the anonymous reviewers for their helpful comments.

8. REFERENCES

- [1] Chandrasekaran, Kannan, Susan Gauch, Praveen Lakkaraju, and Hiep Phuc Luong. Concept-based document recommendations for citeseer authors. In *Adaptive Hypermedia and Adaptive Web-Based Systems*, page 83-92, 2008.
- [2] Sugiyama, K., Kan, M. Y. Scholarly paper recommendation via user's recent research interests. In *Proceedings of the 10th annual joint conference on Digital libraries*, page 29-38, 2010.
- [3] Meng, Fanqi, Dehong Gao, Wenjie Li, Xu Sun, and Yuexian Hou. A unified graph model for personalized query-oriented reference paper recommendation. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, page 1509-1512, 2013.
- [4] McNee, Sean M., Istvan Albert, Dan Cosley, Prateep Gopalkrishnan, Shyong K. Lam, Al Mamunur Rashid, Joseph A. Konstan, and John Riedl. On the recommending of

- citations for research papers. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work*, page 116-125, 2002.
- [5] Zhou, Ding, Shenghuo Zhu, Kai Yu, Xiaodan Song, Belle L. Tseng, Hongyuan Zha, and C. Lee Giles. Learning multiple graphs for document recommendations. In *Proceedings of the 17th international conference on World Wide Web*, page 141-150, 2008.
- [6] Strohman, Trevor, W. Bruce Croft, and David Jensen. Recommending citations for academic papers. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, page 705-706, 2007.
- [7] Tang, Jie, and Jing Zhang. A discriminative approach to topic-based citation recommendation. In *Advances in Knowledge Discovery and Data Mining*, page 572-579, 2009.
- [8] Torres, Roberto, Sean M. McNee, Mara Abel, Joseph A. Konstan, and John Riedl. Enhancing digital libraries with TechLens+. In *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, page 228-236, 2004.
- [9] Wang, Chong, and David M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the SIGKDD '11*, page 448-456, 2011.
- [10] Steyvers, Mark, and Tom Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis* 427, no. 7 page 424-440, 2007.
- [11] Gori, Marco, and Augusto Pucci. Research paper recommender systems: A random-walk based approach. In *Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on*, page 778-781, 2006.
- [12] He, Qi, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. Context-aware citation recommendation. In *Proceedings of the 19th international conference on World wide web*, page 421-430, 2010.
- [13] Lu, Yang, Jing He, Dongdong Shan, and Hongfei Yan. Recommending citations with translation model. In *Proceedings of the CIKM '11*, page 2017-2020, 2011.
- [14] Huang, Wenyi, Saurabh Kataria, Cornelia Caragea, Prasenjit Mitra, C. Lee Giles, and Lior Rokach. Recommending citations: translating papers into references. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, page 1910-1914, 2012.
- [15] Nallapati, Ramesh M., Amr Ahmed, Eric P. Xing, and William W. Cohen. Joint latent topic models for text and citations. In *Proceedings of the SIGKDD '08*, page 542-550, 2008.
- [16] Kataria, Saurabh, Prasenjit Mitra, Cornelia Caragea, and C. Lee Giles. Context sensitive topic models for author influence in document networks. In *Proceedings of the AAAI '11*, page 2274-2280, 2011.
- [17] Kataria, Saurabh, Prasenjit Mitra, and Sumit Bhatia. Utilizing Context in Generative Bayesian Models for Linked Corpus. In *AAAI*, vol. 1, no. 9, page 36, 2010.
- [18] He, Qi, Daniel Kifer, Jian Pei, Prasenjit Mitra, and C. Lee Giles. Citation recommendation without author supervision. In *Proceedings of the fourth ACM international conference on Web search and data mining*, page 755-764, 2011.
- [19] Bai, Bing, Jason Weston, David Grangier, Ronan Collobert, Kunihiko Sadamasa, Yanjun Qi, Olivier Chapelle, and Kilian Weinberger. Supervised semantic indexing. In *Proceedings of the 18th ACM conference on Information and knowledge management*, page 187-196, 2009.
- [20] Grefenstette, Gregory. *The problem of cross-language information retrieval*. Springer US, 1998.
- [21] Oard, Douglas W., and Anne R. Diekema. Cross-language information retrieval. *Annual review of Information science* 33, 1998.
- [22] Oard, Douglas W., and Bonnie J. Dorr. A survey of multilingual text retrieval. 1998.
- [23] Nie, Jian-Yun, Michel Simard, Pierre Isabelle, and Richard Durand. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In *Proceedings of SIGIR '99*, page 74-81, 1999.
- [24] Gollins, Tim, and Mark Sanderson. Improving cross language retrieval with triangulated translation. In *Proceedings of the SIGIR '01*, page 90-95, 2001.
- [25] Ballesteros, Lisa, and W. Bruce Croft. Phrasal translation and query expansion techniques for cross-language information retrieval. In *ACM SIGIR Forum*, vol. 31, no. SI, page 84-91, 1997.
- [26] Dumais, Susan T., Todd A. Letsche, Michael L. Littman, and Thomas K. Landauer. Automatic cross-language retrieval using latent semantic indexing. In *AAAI spring symposium on cross-language text and speech retrieval*, vol. 15, page 21, 1997.
- [27] Dumais, S., G. Furnas, T. Landauer, S. Deerwester, and S. Deerwester. Latent semantic indexing. In *Proceedings of the Text Retrieval Conference*. 1995.
- [28] Potthast, Martin, Benno Stein, and Maik Anderka. A Wikipedia-based multilingual retrieval model. In *Advances in Information Retrieval*, page 522-530. Springer Berlin Heidelberg, 2008.
- [29] Usunier, Nicolas, David Buffoni, and Patrick Gallinari. Ranking with ordered weighted pairwise classification. In *Proceedings of the 26th annual international conference on machine learning*, page 1057-1064, 2009.
- [30] Weston, Jason, Samy Bengio, and Nicolas Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning* 81, no. 1, page 21-35, 2010.
- [31] Weston, Jason, Chong Wang, Ron Weiss, and Adam Berenzweig. Latent collaborative retrieval. *arXiv preprint arXiv:1206.4603* 2012.
- [32] Tibshirani, Robert. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 1996: 267-288.
- [33] Joachims, Thorsten. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, page 133-142, 2002.
- [34] Voorhees, Ellen M. The TREC-8 Question Answering Track Report. In *TREC*, vol. 99, page 77-82.