Cross-Lingual Keyword Recommendation Using Latent Topics

Atsuhiro Takasu
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda, Tokyo, Japan
takasu@nii.ac.jp

ABSTRACT
Multi-lingual text processing is important for content-based and hybrid recommender systems. It helps recommender systems extract content information from broader sources. It also enables systems to recommend items in a user’s native language. We propose a cross-lingual keyword recommendation method, which is built on an extended latent Dirichlet allocation model, for extracting latent features from parallel corpora. With this model, the proposed method can recommend keywords from text written in different languages. We evaluate the proposed method using a cross-lingual bibliographic database that contains both English and Japanese abstracts and keywords and show that the proposed method can recommend keywords from abstracts in a cross-lingual environment with almost the same accuracy as in a monolingual environment.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Information Filtering

General Terms
Algorithms, Experimentation

Keywords
cross-lingual recommendation, LDA, text tagging

1. INTRODUCTION
Content-based and hybrid recommender systems use various kinds of information to extract content. For example, to recommend items such as movies and CDs, recommender systems exploit reviews, titles, and persons related to the items such as actors, directors, and composers. In social tagging to Web pages, systems need to extract information from pages describing a wide range of topics. Because content information is usually described in various languages, the ability to exploit multi-lingual text enriches recommender systems in the following ways. First, it enables recommender systems to extract content information from wider sources. For example, movie recommender systems with multi-lingual processing ability can extract more information about movies by exploiting reviews in various languages. Second, it can provide information to users in their native language when recommending items. For example, when recommending keywords for academic papers, these systems enable users to attach keywords in various languages. This is especially useful for novice users, such as undergraduate students, who do not have sufficient knowledge in certain academic fields. Recommender systems can help them find papers in English by keywords in their native language as well as to attach English keywords to their papers written in their native language.

We propose a cross-lingual keyword recommendation method and discuss how to exploit text in a cross-lingual environment. Keyword recommendation is a type of tag recommendation problem. There are several approaches to solve this problem. For example, Walter et al. tackled this problem by using social network [8]. We discuss the problem from the cross-lingual aspect. Cross-lingual and multi-lingual text processing has been studied in natural language processing and information retrieval communities. Machine translation (MT) is a typical technology to solve the cross-lingual problem. Some researchers applied MT to cross-language information retrieval (CLIR) (e.g., TREC CLIR track [6]). However, query words and keywords are too short to solve the word sense disambiguation by using MT [6]. Therefore, some researchers adopted corpus-based approaches. The latent semantics analysis (LSA) is one method for handling cross-lingual text [5, 3, 10] where text written in multiple languages are converted into the same low dimensional feature space by singular value decomposition.

We adopted a similar approach to LSA, i.e., mapping cross-lingual text into the same feature space. Recently developed latent Dirichlet allocation (LDA) [1] provides a similar function to LSA. Krestel et al. applied LDA to the tag recommendation problem in a mono-lingual environment [4]. We extended LDA to extract latent features from cross-lingual text as well as keywords. The proposed model regards cross-lingual texts as co-occurrence of texts drawn from different information sources. We evaluated the proposed method using a cross-lingual bibliographic database containing both English and Japanese abstracts and keywords and show that the proposed method can recommend
keywords from abstracts in a cross-lingual environment with almost the same accuracy as in mono-lingual environment.

2. LATENT FEATURE MAPPING

2.1 Outline of Method

To recommend keywords in a cross-language setting, we map map both text and keywords in two languages into a single latent feature space. We propose a generative model for this task.

The proposed model is learned from a parallel corpus containing the following four types of information written in two languages \( A \) and \( B \):
- \( x_{wa} \): text describing an article in language \( A \).
- \( x_{ka} \): a set of keywords associated with the article in language \( A \).
- \( x_{wb} \): text describing the article in language \( B \).
- \( x_{kb} \): a set of keywords associated with the article in language \( B \).

We denote the entry \( x := (x_{wa}, x_{ka}, x_{wb}, x_{kb}) \) in the corpus as a quad-tuple.

We represent text as a bag of words, i.e., a text \( w \) is represented by a set \( \{w_i\}_{i=1}^{\|w\|} \), where \( w_i \) (1 \( \leq i \leq \|w\| \)) is a word included in the text \( w \). Note that \( \|w\| \) stands for the number of elements included in a set or a sequence of \( w \).

We assume a parallel texts and keywords are generated by a generative model based on latent topics \( T := \{t_i\}_{i=1}^{\|T\|} \), as in LDA. As a result, we obtain a probability distribution over the topic set \( T \) for each quad-tuple in the corpus. In this way, both words in the text and keywords are mapped to latent topics in terms of probability distribution. Intuitively, words and keywords co-occurring in the parallel corpus tend to be mapped into the same topic.

When given a document \( d \) in either or both languages, we
1. estimate the probability distribution \( \theta_d \) of \( d \) over the latent topics,
2. calculate the probability that each keyword \( w \) is generated by \( \theta_d \), and
3. rank the keywords in the corpus according to the probability.

2.2 Generative Model for Parallel Corpus

For parameters \( \theta \) and \( \alpha \), \( \text{Multi}(\theta) \) and \( \text{DIR}(\alpha) \) respectively denote the multinomial and Dirichlet distributions.

We use two types of multinomial distributions for corpus generation. For each quad-tuple, a multinomial distribution \( \text{Multi}(\theta) \) is used to generate topics for each word and keyword in the quad-tuple. We use one multinomial distribution for each quad-tuple. Let a parallel corpus be \( C := \{x_i\}_{i=1}^{\|C\|} \). The model then uses a set \( \Theta := \{\theta_i\}_{i=1}^{\|T\|} \) of multinomial distributions. These multinomial parameters are generated by a single Dirichlet distribution \( \text{DIR}(\alpha) \) as in LDA.

For each topic \( t \in T \), we use four multinomial distributions \( \text{Multi}(\phi_{wa}), \text{Multi}(\phi_{wb}), \text{Multi}(\phi_{ka}), \text{Multi}(\phi_{kb}) \) to generate words in language \( A \), words in language \( B \), keywords in language \( A \), and keywords in language \( B \), respectively. We denote the list of multinomial parameters for word generation as
\[
\Phi := \{\phi_{wa}, \phi_{wb}, \phi_{ka}, \phi_{kb}\}.
\]

Note that these multinomial distributions are corpus-wide distributions, i.e., the same distribution is used for all quad-tuples in the corpus. Let \( W_{wa}, W_{wb}, W_{ka}, \) and \( W_{kb} \) denote the sets of words and keywords appearing in the corpus. For each topic \( t \in T \), the parameter of the distribution for word generation is defined as \( \phi_{wa,t} := (\phi_{wa,t,w})_{w \in W_{wa}} \). These multinomial distributions are generated by a Dirichlet distribution \( \text{DIR}(\beta_{wa}) \). The other three multinomial parameters are generated similarly according to Dirichlet distributions \( \text{DIR}(\beta_{wb}), \text{DIR}(\beta_{ka}), \text{DIR}(\beta_{kb}) \), respectively. We denote the list of Dirichlet priors’ parameters as \( \beta := (\alpha, \beta_{wa}, \beta_{wb}, \beta_{ka}, \beta_{kb}) \).

For a set \( T \) of topics, Dirichlet priors \( \beta_{wa}, \beta_{wb}, \beta_{ka}, \beta_{kb} \), and \( \beta_{}\), the generative process is given as follows:
1. for each topic \( t \in T \), generate corpus-wide multinomial parameters
   \[
   \phi_{wa,t} \sim \text{DIR}(\beta_{wa}), \quad \phi_{wb,t} \sim \text{DIR}(\beta_{wb}),
   \phi_{ka,t} \sim \text{DIR}(\beta_{ka}), \quad \phi_{kb,t} \sim \text{DIR}(\beta_{kb})
   \]
2. for each quad-tuple \( x := (x_{wa}, x_{ka}, x_{wb}, x_{kb}) \in C \)
   (a) generate a multinomial parameter \( \theta_t \sim \text{DIR}(\alpha) \)
   (b) for each word in \( x_{wa} \)
      i. generate a topic \( t \sim \text{Multi}(\theta_t) \)
      ii. generate a word \( w \sim \text{Multi}(\phi_{wa,t,w}) \)
   (c) generate latent topics and keywords in \( x_{ka}, x_{kb} \), similarly according to the corresponding multinomial distributions.

Figure 1: Graphical model for parallel corpus. \( D \) denotes number of quad-tuples. \( D_{wa} \) and \( D_{wb} \) denote numbers of words included in abstracts in languages \( A \) and \( B \), respectively. \( T \) denotes number of latent topics.
Figure 1 shows a graphical model of this generative process. The multinomial distribution $\theta_t$ represents a feature vector for a quad-tuple $x$, whereas $\phi_{w_k}$, $\phi_{k\alpha}$, and $\phi_{w_t}$ represent the relationship between a topic $t$ and words and keywords in language A and B, respectively.

Although LDA generates a single text, in this model, LDA is extended to generate four types of text, i.e., $x_{wa}$, $x_{ka}$, $x_{wa}$, and $x_{ka}$. Note that co-occurred words and keywords are likely to be generated from the same topic because the model uses the same multinomial distribution for the same quad-tuple.

This model is a sort of the record model we proposed in a previous study [7]. The record model is originally designed for a record matching problem. We modified the model for keyword recommendation in a cross-lingual environment. Wang et al. recently proposed a cross-lingual information retrieval framework by extracting latent topics [9]. Their model is similar to the one we propose, but their model merges words in two languages to a single word set and uses a single word distribution probability for each latent topic. On the other hand, our proposed model uses separated word distributions for each language.

### 2.3 Parameter Estimation

The parameters are estimated using Gibbs sampling similar to LDA. For the $i$th quad-tuple $z_i$, let us denote the topic assignment to each word and keyword in the text and keyword in each language as $z_i := \{z_{iw_a}, z_{ik_a}, z_{iw_b}, z_{ik_b}\}$. For the topic assignment, the complete-data likelihood is given as

$$
\prod_{i=1}^{[C]} \Pr(x_i, z_i, \Theta, \Phi; \Lambda) = \prod_{f \in F} \Pr(\phi_f, \beta_f) \prod_{i=1}^{[C]} \Pr(\Theta, \alpha) \\
\left[ \prod_{i=1}^{[C]} \prod_{f \in F} \Pr(x_{if}, z_{if} | \Theta, \phi_f) \right].
$$

(1)

where $F$ is a set of fields of a quad-tuple, i.e., $\{w_a, k_a, w_b, k_b\}$. The terms $A$ and $B$ in Eq. (1) respectively represent the probabilities for multinomial parameters of word and topic generation. Note that multinomial parameters are generated according to the corresponding Dirichlet distributions.

The term $C$ in Eq. (1) represents the probability to generate words and keywords in languages A and B. For each field $f \in \{w_a, k_a, w_b, k_b\}$ in a quad-tuple, the probability to generate $x_{if}$ is given by

$$
\Pr(x_{if}, z_{if} | \Theta, \phi_f) = \prod_{j=1}^{[x_{if}]} \Pr(z_{ij} | \theta_j) \Pr(x_{ij} | z_{ij}, \phi_f).
$$

(5)

Because Dirichlet priors are a conjugate distribution of multinomial distributions, we can calculate the marginal distribution

$$
\prod_{i=1}^{[C]} \Pr(x_i, z_i; \Lambda) = \int \int \prod_{i=1}^{[C]} \Pr(x_i, z_i, \Theta, \Phi, \Lambda) d\Theta d\Phi.
$$

as in LDA. We omit the derivation for brevity.

Let $Z_{ij}$ denote the random variable for a latent topic generating the $j$th word in a field $f$ in the $i$th quad-tuple. Furthermore, let $Y_{ij}$ denote the topic assignment except for the $j$th word in the field $f$ in the $i$th quad-tuple.

$$
\Pr(Z_{ij} = \hat{t} | Y_{-ij}; \Lambda) \propto (\alpha_t + N_{ij}^{Y_{-ij}})(\beta_{\hat{t}f_{w_{ij}}} + N_{w_{ij}}^{Y_{-ij}}) / \sum_{w_f} (\beta_{w_f} + N_{w_f}^{Y_{-ij}}),
$$

(2)

where

- $N_{ij}^{Y_{-ij}}$ denotes the number of times a topic $\hat{t}$ is assigned to a word in the $i$th quad-tuple except for the $j$th word in the field $f$, and
- $N_{w_{ij}}^{Y_{-ij}}$ denotes the number of times a topic $\hat{t}$ is assigned to a word $w$ in the field $f$ in the $i$th quad-tuple except for the $j$th word.

We repeat the topic assignment to each word and keyword in the corpus according to Eq. (2) until convergence to obtain the topic generation probability for each quad-tuple and topic

$$
\theta_t \propto (\alpha_t + N_{ij}^{Y_{t}})
$$

(3)

and, word generation probability for each topic $t \in T$, field $f \in \{w_a, k_a, w_b, k_b\}$, and keyword $w \in W_f$

$$
\phi_{tw} = (\beta_{tw} + N_{w}^{Y_{tw}}) / \sum_{w'} (\beta_{w'} + N_{w'}^{Y_{tw}})
$$

(4)

as in [2].

### 2.4 Keyword Recommendation

When given an abstract $x_{wa}$ in language A or $x_{wa}$ in language B, we re-run Gibbs sampling for a quad-tuple $x_{wa}$, $\{\}, \{\}$ or $\{\}, \{\}, \{\}$, respectively. We then obtain the topic assignment $Y := \{Y_{wa}, \{\}, \{\}\}$ or $\{\}, \{Y_{wa}, \{\}, \{\}\}$ to the given abstract. Finally we derive a topic probability distribution $\theta$ of the given abstract with Eq.(3) using obtained $Y_{wa}$ and $Y_{wa}$ instead of $Y_{wa}$. Because the proposed model is designed to independently generate abstracts in language A and B using the same topic distribution $\theta$, an effective topic distribution is expected to be derived from either an abstract $Y_{wa}$ or $Y_{wa}$ if it contains sufficient words for topic estimation. Note that we can also derive a topic distribution using both $x_{wa}$ and $x_{wa}$.

For each keyword $w \in W_{ka}$ of language A, the keyword score is calculated by

$$
\sum_{t \in T} \hat{\theta}_{tw}.
$$

(5)

Similarly, for each keyword $w \in W_{ka}$ of language B, the keyword score is calculated by

$$
\sum_{t \in T} \hat{\theta}_{tw}.
$$

(6)

We recommend top-$k$ keywords in the ranked list using Eq. (5) or (6) where $k$ is the number of recommended keywords.

Because the keyword generation probability is calculated from the latent topic distribution $\hat{\theta}$, that is independent of languages, we can recommend keywords in a language different from that of the abstract used for deriving $\hat{\theta}$.
3. EXPERIMENTAL RESULTS

In our experiment we used a bibliographic database containing 126,496 academic papers each attached with both English and Japanese abstracts as well as keywords in both languages. We first chose keywords that appears in more than 50 articles. As a result, we obtained 651 English and 926 Japanese keywords. Then, we chose papers that contain at least one keyword of the chosen English and Japanese keywords. As a result, 53,119 papers were used in this experiment.

We applied a morphological analyzer, MeCab, \(^1\) to the Japanese abstracts and chose 5,000 words according to the document frequencies. As a result, the average number of words included in a Japanese abstract is 70.5. As for English abstract, we removed stop words and chose 5,000 words according to the document frequencies. The average number of words included in an English abstract is 61.8.

We conducted 5-fold cross validation for evaluation where 53,119 papers are split into five groups of equal size. Then, the corpus-wide probability distributions \(\phi_{\text{en}}\), \(\phi_{\text{m}}\), \(\phi_{\text{n}}\), and, \(\phi_{\text{k}}\), were estimated from four of the groups. The remaining group was used to evaluate keyword recommendation accuracy.

Some keywords have broader meanings and others are very specific. To fit in this situation, we divided both English and Japanese keywords into three groups of almost equivalent size according to the document frequencies. These groups are referred to as high-frequent, middle-frequent, and low-frequent keywords. The number of Japanese keywords of each of these groups were 310, 311, and 305, respectively, whereas those of English keywords were 214, 212, and 225.

We calculated the probability given by Eqs. (5) and (6), then ranked each keyword group according to these probabilities. We then measured the performance of keyword recommendation by recall, i.e., the ratio of author-given keywords included in the top-k ranked recommended keywords over the total number of author-given keywords.

Because the recommendation accuracy is affected by the number of latent topics used in the model, we first tuned the number of latent topics. We compared the accuracy of keyword recommendation to the model consisting of 4, 8, 16, and 24 topics. The topic probability distribution \(\Theta\) for Eqs. (5) and (6) were estimated using both English and Japanese abstracts in the test set. Fig. 2 shows the accuracies for high-frequent English keyword recommendations with respect to the number of topics labeled “EN 4”, “EN 8”, “EN 16”, and “EN 24”, respectively. As shown in the graph, the accuracy improved as the number of topics increased. The best performance is when the number of topics is 16; however performance decreases when there are more than 16 topics. We obtained the same results for all groups of keywords, i.e., high-, middle-, and low-frequent Japanese and English keywords. Therefore, the number of latent topics was set to 16 in the following experiments.

Next, we evaluated the keyword recommendation in a cross-lingual environment. We first estimated the model using both English and Japanese abstracts as well as keywords to obtain conditional probability distributions for each latent topic, i.e., \(\phi_{\text{en}}, \phi_{\text{m}}, \phi_{\text{n}}, \phi_{\text{k}}\). Then, for each article in a test set, we estimated three types of topic distributions, i.e., \(\Theta_{\text{m}}\) using both English and Japanese abstracts, \(\Theta_{\text{m}}\) using only English abstracts, and \(\Theta_{\text{n}}\) using only Japanese abstracts. We then generated the ranked lists of high-, middle-, and low-frequent English and Japanese keywords using these three distributions according to Eqs. (5) and (6). Fig. 3(a) and (b) respectively show the accuracies for the English and Japanese keyword recommendations. In the graph, “multi” means that the topic distribution was estimated using both English and Japanese abstracts, i.e., \(\Theta_{\text{m}}\). On the other hand, “cross” means the topic distribution was estimated using abstracts written in different languages.

\(^1\)http://mecab.sourceforge.net
from recommended keywords. For example, in Fig. 3(a), “high multi” (resp. “low cross”) stands for the accuracy of high-frequent (resp. low-frequent) English word recommendation using the topic distribution estimated from both English and Japanese abstracts (resp. only Japanese abstract). By comparing the accuracies of multi and cross, we can see that the recommendation accuracies for multi are slightly better than those for cross for high-, middle-, and low-frequent keywords in both English and Japanese; however, the difference is very small. From these results we believe that a topic distribution estimated from an abstract in a single language is similar to that from two languages. Therefore, the proposed method is suitable for cross-lingual keyword recommendation.

One drawback of the proposed method is that we need to calculate the topic distribution of a paper to recommend keywords. This process includes Gibbs sampling, which may require considerable number of iterations until convergence. We checked the required number of iterations for convergence and computation time to calculate the topic distribution. Fig. 4 shows the average processing time for Gibbs sampling per quad-tuple in a test dataset and the average Kulback-Leibler divergence between the topic distribution at the $r$th iteration and the distribution at the 300th iteration. As shown in this graph, Gibbs sampling quickly converges within a few iterations. The required processing time is only a few milliseconds. These results show that the proposed method is practical in terms of processing time.

4. CONCLUSIONS

We proposed a keyword recommendation method in a cross-lingual environment. The main idea is to map both text and keywords in different languages into a single feature space, i.e., a probability distribution over latent topics. To achieve this, we proposed an extended LDA that utilizes cross-lingual texts. By conducting experiments using English and Japanese bibliographic database, we showed that the proposed method can recommend keywords in a cross-lingual environment with almost the same accuracy as in a mono-lingual environment.

This study is still in the preliminary stage and should be extended into various research directions. First, we need to incorporate recent progress in natural language processing techniques language into the model to improve its accuracy. Second, although we split keywords into three groups in an ad hoc manner, the structure of a keyword set should be utilized more carefully. Third, collaborative filtering techniques should be added to enhance the recommendation accuracy.

5. REFERENCES