ABSTRACT

The major limitation in bilingual latent semantic analysis (bLSA) is the requirement of parallel training corpora. Motivated by semi-supervised learning, we propose a cluster-based bLSA training approach to incorporate monolingual corpora. Treating each parallel document pair as centroids of the parallel document clusters, each monolingual document is associated to the closest centroid according to their topic similarity. The resulting parallel document clusters are used as constraints to enforce a one-to-one topic correspondence in variational EM. Slight performance improvement in crosslingual language model adaptation is observed compared to the baseline without monolingual corpora.

Index Terms: monolingual corpora, bilingual LSA, crosslingual word trigger, crosslingual LM adaptation

1. INTRODUCTION

In [1], we had proposed bilingual Latent Semantic Analysis (bLSA) for crosslingual language model (LM) adaptation for statistical machine translation (SMT). bLSA works as a crosslingual word trigger model and is usually trained on parallel documents with bilingual sentence-aligned text. This requirement limits the coverage of other possible crosslingual word triggers in bLSA. In this paper, we attempt to address the limitation via incorporating monolingual non-parallel corpora into training. Incorporating monolingual corpora is attractive since they cover a broad range of topics and vocabulary and are easy to collect. However, blind incorporation of the corpora may destroy a one-to-one topic correspondence in bLSA since the alignment between a source and target monolingual document is unknown or even non-existent.

To work around the issue of unknown document alignment, we employ a semi-supervised learning approach where some parallel seed documents are given. The smoothness assumption [2] says that if two points $x_1$ and $x_2$ are close in a high-density region, the corresponding outputs $y_1$ and $y_2$ should be close as well. In our setting, each parallel source-target document pair is treated as input-output points in some spaces. With the smoothness assumption, we associate each monolingual document to the closest parallel document via a document similarity measure. As a result, a partial alignment between the source and target monolingual document is recovered at the document cluster level. The parallel document clusters which are populated with monolingual documents are served as constraints which can be integrated into the standard bLSA training via the Lagrangian theory.

Related work includes crosslingual LSA based on singular value decomposition (SVD) [3] where bilingual documents are concatenated into a single input supervector before SVD. Incorporation of monolingual documents is performed by filling in zeros in the missing counterparts of a “pseudo”-bilingual supervector. Another work is Bilingual Topic Admixture Model for word alignment [4] where topic-dependent translation lexicon are modeled. However, their approach requires sentence-aligned parallel documents having the same requirement as phrase extraction. Using comparable corpora for parallel fragment extraction has been advocated in [5].

We organize the paper as follows: In Section 2, we review the bLSA-based LM adaptation framework and describe the proposed cluster-based bLSA training approach followed by crosslingual LM adaptation. In Section 3, we evaluate our approach on different training scenarios. In Section 4, we conclude our work.

2. BILINGUAL LATENT SEMANTIC ANALYSIS

As a review, the goal of bLSA [1] is to enforce a one-to-one topic correspondence between the source and target LDA-style models [6]. For instance, say topic 10 of the Chinese LSA model is about politics. Then topic 10 of the English LSA model corresponds to politics and so forth. Figure 1 illustrates the idea of topic transfer between monolingual LSA models followed by bLSA-based LM adaptation for SMT. With a one-to-one topic correspondence, the topic distribution inferred from the source language can be transferred to...
adapt a target LM before SMT via transfer of topic distribution. However, the major limitation of bLSA requires parallel training documents. In the following section, we shed light on using the monolingual corpora for potential coverage of crosslingual word triggers which are not captured in the parallel corpora. Our approach involves obtaining the parallel document clusters from the monolingual documents which act as inputs for the cluster-based bLSA training.

### 2.1. Parallel document clusters

We propose a platform for integrating monolingual documents via parallel document clusters. The concept of parallel document clusters is depicted in Figure 2. The idea is to use the given parallel documents to form the initial document clusters containing only a single document. Then a parallel document cluster is populated by associating each monolingual source and target documents to the corresponding closest document based on a similarity measure. We represent each document \(d\) as a \(K\)-dimensional topic posterior vector \(p(k|d)\) inferred by monolingual LSA and the distance between two documents is computed as the dot product between their posterior vectors. Consequently, partial document alignment between the monolingual documents is recovered at the cluster level. Intuitively, monolingual documents within a cluster are expected to come from similar topics. We prevent “noisy” monolingual documents from folding into a cluster by setting a threshold \(\tau\) so that any monolingual document with distance larger than \(\tau\) from all cluster centroids is removed.

### 2.2. Cluster-based bilingual LSA training

The development of cluster-based bLSA training assumes that the average topic distribution of a source and target document cluster is identical. In other words, a one-to-one topic correspondence among a pair of parallel document cluster is assumed. Given a pair of parallel document cluster \(C = \{C^{(i)}, C^{(j)}\}\) where \(i\) and \(j\) represent the index of the source and target language respectively, this assumption can be encoded as follows:

\[ \forall k : E[p^{(i)}(k|C^{(i)})] = E[p^{(j)}(k|C^{(j)})] \]
\[ \implies \sum_{d \in C^{(i)}} p^{(i)}(k|d) = \frac{\sum_{d \in C^{(j)}} p^{(j)}(k|d)}{|C^{(j)}|} \]

where \(d\) and \(k\) denote a document and topic index respectively. For monolingual LSA training using Latent Dirichlet Allocation (LDA) [6], the lower bound of the log likelihood of a document \(W_d = w_1 \ldots w_{N_d}\) denoted as \(L(W_d)\), is:

\[ \log \sum_{Z} p(W_d, Z, \theta) = \log \sum_{\theta} p(\theta) \prod_{n=1}^{N_d} \theta_{z_n} \cdot \beta_{w_n z_n} \geq E_q[\log \frac{p(\theta)}{q(\theta)}] + \sum_{n=1}^{N_d} E_q[\log \theta_{z_n}] + E_q[\log \beta_{w_n z_n}] = L(W_d) \]

where \(Z = z_1 \ldots z_{N_d}\) and \(\theta\) denote the latent topic sequence and the topic distribution vector sampled from a Dirichlet prior respectively. The lower-bound value is achieved via the Jensen’s inequality using the factorizable variational posterior distribution over the latent variables \(q(Z, \theta|d) = q(\theta) \prod_{n=1}^{N_d} q(z_n)\). Therefore, the objective function for bilingual LSA training with a pair of document cluster is the sum of the lower-bound log likelihood of the documents in the source and target cluster subject to the topic correspondence constraint in Eqn 2. With the Lagrange multipliers \(\lambda_{C,k}\), the objective function is shown as follows:

\[ L(W; \Lambda, \Gamma) = \sum_{d \in C^{(i)}} L^{(i)}(W_d; \Lambda) + \sum_{d \in C^{(j)}} L^{(j)}(W_d) + \sum_{C,k} \lambda_{C,k} \left( \sum_{d \in C^{(i)}} p^{(i)}(k|d) - \frac{\sum_{d \in C^{(j)}} p^{(j)}(k|d)}{|C^{(j)}|} \right) \]

To derive the E-step formula, we compute the partial derivative of \(L(W; \Lambda, \Gamma)\) with respect to \(q^{(i)}(z_n = k|d)\) subject to \(\sum_{k=1}^{K} q^{(i)}(z_n = k|d) = 1\) and set it to zero:

\[ q^{(i)}(z_n = k|d) = \beta^{(i)}_{w_n k} \cdot e^{E_q[\log \theta^{(i)}] + \mu^{(i)}_{w_n k}} \cdot e^{\frac{\lambda_{C,k}}{N_d}} \]
where \( \mu_{dn}^{(i)} \) is the Lagrange multiplier for probability normalization in \( q^{(i)}(z_n = k|d) \). If we assume that each document has the same number of words so that \( N_d \approx N \), we can use Eqn 3 to construct the estimated \( \hat{p}(k|d) \) which are put back to the left hand side of Eqn 2. After rearranging terms of the resulting equation, we obtain the following result:

\[
\frac{r_{j/i}(k|C)}{E[p^{(i)}(k|C^{(i)})]} e^{E_k[\log \hat{\theta}_{j/i}^{(i)}]} \approx \frac{E[p^{(i)}(k|C^{(i)})]}{E[p^{(i)}(k|C^{(i)})]} = r_{j/i}(k|C)
\]

where \( r_{j/i}(k|C) \) is the topic ratio between the target and source document cluster in \( C \). Substituting \( r_{j/i}(k|C) \) into Eqn 3 and using the standard result of LDA for \( \theta^{(i)} \) yield the following variational E-steps for a document in \( C^{(i)} \):

\[
q^{(i)}(z_n = k|d) \propto \beta_{w_{dk}} e^{E_k[\log \hat{\theta}_{j/i}^{(i)}]} \cdot r_{j/i}(k|C)
\]

\[
\gamma_{dk} = \alpha_k + \sum_n \phi_{dnk}^{(i)}
\]

Our E-steps resemble those in LDA except the extra term \( r_{j/i}(k) \) to enforce a one-to-one topic correspondence between \( C^{(i)} \) and \( C^{(j)} \) in Eqn 1. By symmetry, the E-steps for documents on the target language \( j \) can be performed in a similar fashion. After performing the E-steps on all monolingual documents, \( r_{j/i}(k|C) \) is updated using Eqn 5 which are then substituted back to the E-steps iteratively until convergence is reached. The M-step follows the standard derivation of LDA [6] which is not shown here due to limited space.

### 2.3. Crosslingual LM adaptation

Given a source document, we apply the E-steps to estimate the variational topic posterior \( q(\theta) \) on the source language. We use the MAP estimate \( \hat{\theta}_{j/i}^{(i)} \) as the topic weights on the source language for the target LSA to obtain an in-domain LSA marginal:

\[
p_{lsa}^{(j)}(w) = \sum_{k=1}^{K} \beta_{w_{dk}} \hat{\theta}_{j/i}^{(i)}
\]

where \( \hat{\theta}_{j/i}^{(i)} = \sum_{k=1}^{K} \gamma_{dk} \) for LDA. We integrate the LSA marginal into the target back-translation using marginal adaptation [7] which minimizes the Kullback-Leibler divergence between the adapted LM and the background LM:

\[
p_{bg}(w|h) \propto \left( \frac{p_{lsa}(w)}{p_{bg}(w)} \right)^{\beta} \cdot p_{bg}(w|h).
\]

When \( w \) is a stopword such as “is” and “the” or punctuations, the N-gram probability is not adapted because predicting stopwords mostly relies on the syntactic context only.

### 3. EXPERIMENTAL SETUP

Our bLSA training setup employed parallel Chinese–English corpora from the Donga news websites 1 containing 28k parallel documents with 13M Chinese characters and 9M English words. We applied Latent Dirichlet-Tree Allocation [8]

1http://chinaenglish.donga.com

| Topics | Top new words sorted by \( P(w|k) \) |
|--------|------------------|
| “CH (Art)” | film reward, ballet, art festival, ballet club |
| “EN (Art)” | ballet, ballads, edinburgh, pianist |
| “CH (Economy)” | export rate, life condition, 2nd season |
| “EN (Economy)” | diesel, greenspan, durable, dived |
| “CH (Electronics)” | router, broadband service, album |
| “EN (Electronics)” | 3g, bro, pixel, copying, piracy, sw |

Table 1. New topical words discovered by bLSA from pseudo-monolingual corpora. Words on the Chinese side are translated into English for illustration purpose.

which is a LDA-style model for modeling topic correlation. Number of latent topics \( K \) was set to 50. Sentence-aligned parallel corpora containing 1M sentence pairs were used for phrase extraction via the “PESA” (Phrase Pair Extraction as Sentence Splitting) approach [9]. SMT translation was performed using an approach similar to that in [10] with decoding parameters optimized on BLEU via minimum error rate training on the RT04 development set for spoken language translation. Our background 4-gram LM was trained on a combination of Donga corpora for bLSA, parallel sentences for SMT, and the 2004 Xinhua News corpora using modified Kneser-Ney smoothing.

To show the progress of incorporating monolingual corpora, we randomly split the corpora into two parts: 10% of the documents (2.8k) as parallel seed documents and the remaining 90% as pseudo-monolingual (p-mono) documents (25k) where one-to-one document correspondence were omitted. We compare different bLSA training scenarios from \( A \) to \( F \): Scenario \( A \) uses only 10% of the parallel corpora as a baseline. Scenario \( B \) incorporates the remaining 90% of pseudo-monolingual portion on top of scenario \( A \) without constraint, i.e. the topic ratios \( r_{*/*}(k|C) \) in Eqn 6 are set to 1 meaning that parallel document clusters are not used. Scenario \( C \) has similar settings as scenario \( B \) except that parallel document clusters are formed using the proposed approach. Scenario \( D \) resembles scenario \( B \) except that real monolingual corpora from the Chinese and English 2004 Xinhua news. Scenario \( E \) shares the same rationale as scenario \( C \) but using real monolingual corpora.

### 3.1. Results

Table 1 shows the top new words discovered by bLSA from the pseudo-monolingual corpora by filtering out words which are already covered in the initial parallel seed documents. The new words tend to be crosslingual word triggers suggesting that our approach works well in the pseudo-monolingual
We proposed an approach to incorporate monolingual corpora for bilingual LSA training. The key of our approach is to enforce a one-to-one topic correspondence between parallel document clusters populated by monolingual documents. The proposed bLSA training is based on variational EM and Lagrangian theory. Results show that our approach successfully incorporates monolingual corpora and produces slightly better cross-lingual LM adaptation results than the baseline without monolingual corpora. Incorporating monolingual corpora without the parallel document clusters can lead to severe performance degradation implying that a one-to-one topic correspondence constraints between parallel document clusters is crucial.

5. REFERENCES


