INCORPORATING MONOLINGUAL CORPORA INTO BILINGUAL LATENT SEMANTIC ANALYSIS FOR CROSSLINGUAL LM ADAPTATION

Yik-Cheung Tam and Tanja Schultz

InterACT, Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA

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

This work is partly supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-2-0001. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

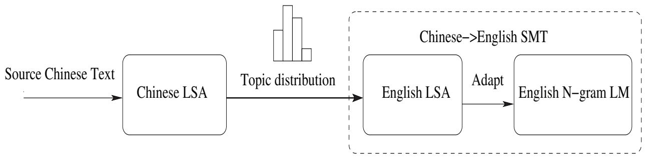

Fig. 1. Bilingual LSA framework for crosslingual LM adaptation in SMT via transfer of topic distribution.

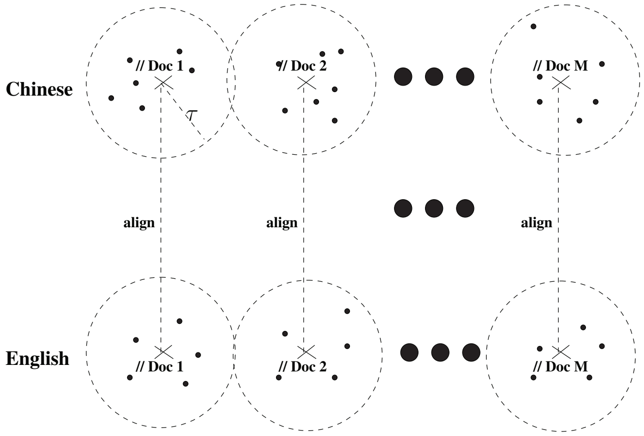

Fig. 2. Parallel document clusters formed by monolingual documents (black dots) using M parallel seed documents.

the target language to adapt a target LM *before* SMT decoding. 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 parallel 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 τ so that any monolingual document with distance larger than τ 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:

$$\forall k : E[p^{(i)}(k|C^{(i)})] = E[p^{(j)}(k|C^{(j)})] \qquad (1)$$

$$\implies \frac{\sum_{d \in C^{(i)}} p^{(i)}(k|d)}{|C^{(i)}|} = \frac{\sum_{d \in C^{(j)}} p^{(j)}(k|d)}{|C^{(j)}|} \qquad (2)$$

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 $W_d = w_1...w_{N_d}$, denoted as $L(W_d)$, is:

$$\log \int_\theta \sum_Z p(W_d, Z, \theta) = \log \int_\theta p(\theta) \prod_{n=1}^{N_d} \theta_{z_n} \cdot \beta_{w_n z_n} \geq$$

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

where $Z = z_1...z_{N_d}$ and θ 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 λ_{Ck}, the objective function is shown as follows:

$$L(W; \Lambda, \Gamma) = \sum_{d \in C^{(i)}} L^{(i)}(W_d; \Lambda) + \sum_{d \in C^{(j)}} L^{(j)}(W_d)$$

$$+ \sum_{C,k} \lambda_{Ck} \left(\frac{\sum_{d \in C^{(i)}} p^{(i)}(k|d)}{|C^{(i)}|} - \frac{\sum_{d \in C^{(j)}} p^{(j)}(k|d)}{|C^{(j)}|} \right)$$

$$\text{where } p(k|d) \approx \frac{\sum_{n=1}^{N_d} q(z_n = k|d)}{N_d} \text{ for large } N_d$$

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:

$$q^{(i)}(z_n = k|d) = \beta^{(i)}_{w_{dn}k} \cdot e^{E_q[\log \theta^{(i)}_k] + \mu^{(i)}_{dn}} \cdot e^{\frac{\lambda_{Ck}}{|C^{(i)}| \cdot N^{(i)}_d}} \qquad (3)$$

where $\mu^{(i)}_{dn}$ 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 $p(k|d)$ which are put back to the left hand side of Eqn 2. After rearranging terms of the resulting equation, we obtain the following result:

$$e^{\frac{\lambda_{Ck}}{|C^{(i)}|\cdot N}} = \frac{E[p^{(j)}(k|C^{(j)})]}{\frac{1}{|C^{(i)}|\cdot N}\sum_{d\in C^{(i)}}\sum_n \beta^{(i)}_{w_{dn}k}\cdot e^{E_q[\log \theta^{(i)}_k]+\mu^{(i)}_{dn}}} \quad (4)$$

$$\approx \frac{E[p^{(j)}(k|C^{(j)})]}{E[p^{(i)}(k|C^{(i)})]} = r_{j/i}(k|C) \quad (5)$$

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 $q(\theta; \{\gamma_{dk}\})$ yield the following variational E-steps for a document in $C^{(i)}$:

$$q^{(i)}(z_n = k|d) \propto \beta^{(i)}_{w_{dn}k} \cdot e^{E_q[\log \theta^{(i)}_k]} \cdot \mathbf{r_{j/i}(k|C)} \quad (6)$$

$$\gamma^{(i)}_{dk} = \alpha^{(i)}_k + \sum_n^{N^{(i)}_d} \phi^{(i)}_{dnk} \quad (7)$$

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 proceeded 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 $\hat{\theta}^{(i)}_k$ as the topic weights on the source language for the target LSA to obtain an in-domain LSA marginal: $p^{(j)}_{lsa}(w) = \sum_{k=1}^{K} \beta^{(j)}_{wk} \cdot \hat{\theta}^{(i)}_k$ where $\hat{\theta}^{(i)}_k = \frac{\gamma_k}{\sum_{k'=1}^{K}\gamma_{k'}}$ for LDA. We integrate the LSA marginal into the target background LM using marginal adaptation [7] which minimizes the Kullback-Leibler divergence between the adapted LM and the background LM: $p_a(w|h) \propto \left(\frac{p_{lsa}(w)}{p_{bg}(w)}\right)^{\beta} \cdot p_{bg}(w|h)$. 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] 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.

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.

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 using 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. Scenario F serves as an ideal case where 100% parallel corpora are available. We applied different bLSA models for crosslingual LM adaptation at the story level using the source manual transcription of the RT04 test set comprising CCTV, RFA and NTDTV shows. Performance metrics are 4-gram perplexity and BLEU.

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

[1] http://{china,english}.donga.com

Scenario	CCTV	RFA	NTDTV	Overall
BG EN 4-gram	16.12	8.83	14.04	13.22
	(85)	(189)	(127)	126)
A. 10% //	16.26	8.90	14.09	13.28
(baseline)	(78)	(181)	(115)	(117)
B. + p-mono	16.46	8.68	14.29	13.36
(blind)	(81)	(189)	(116)	(121)
C. + p-mono	**16.52**	**8.95**	**14.31**	**13.47**
(// doc cluster)	**(75)**	**(178)**	**(109)**	**(113)**
D. + real mono	15.66	8.87	14.28	13.12
(blind)	(91)	(192)	(135)	(133)
E. + real mono	**16.30**	**9.04**	**14.40**	**13.44**
(// doc cluster)	**(76)**	**(178)**	**(114)**	**(115)**
F. 100% //	16.44	9.06	14.38	13.49
(golden line)	(74)	(172)	(107)	(111)

Table 2. bLSA based LM adaptation performance on BLEU (target perplexity) on different training scenarios.

case. Table 2 shows the results on target word perplexity and BLEU after bLSA-based LM adaptation. The baseline bLSA of scenario A shows reduction in perplexity compared to the unadapted LM which is surprisingly decent given the small amount of parallel training data. Incorporating pseudo-monolingual documents further reduces perplexity in scenario C compared to scenario A, and approaches to the ideal case in scenario F using full parallel corpora. Given that scenario A and F set the overall upper-bound and lower-bound perplexity of 117 and 111 respectively, our proposed approach works well with the overall perplexity of 113. On the other hand, folding in monolingual corpora without parallel document clusters as constraints in scenario B degrades perplexity compared to scenario A. This indicates that using parallel document clusters as constraints are crucial in incorporating monolingual documents. We observed similar trend in perplexity performance when real monolingual corpora were employed, reducing perplexity in scenario E, but deteriorating perplexity in scenario D even further compared to scenario B. This implies that our approach becomes more crucial when incorporating real monolingual documents. Regarding SMT performance, the improvement on BLEU is consistently shown in scenario C and E like their perplexity performance although the gain is not significant. But since the difference in BLEU between the best scenario F and the baseline scenario A is only 0.21%, the gain after incorporating monolingual corpora using our approach is reasonable with 0.19% and 0.15% improvement in scenario C and E respectively using a single target reference for scoring.

4. CONCLUSIONS

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 crosslingual 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

[1] Y. C. Tam, I. Lane, and T. Schultz, “Bilingual LSA-based LM adaptation for spoken language translation,” in *Proc. of ACL*, 2007.

[2] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning*, MIT Press, Cambridge, MA, 2006.

[3] W. Kim and S. Khudanpur, “Lexical triggers and latent semantic analysis for cross-lingual language model adaptation,” *ACM Transactions on Asian Language Information Processing*, vol. 3, no. 2, pp. 94–112, June 2004.

[4] B. Zhao and E. P. Xing, “BiTAM: Bilingual topic admixture models for word alignment,” in *Proc. of ACL*, 2006.

[5] C. Quirk, R. U., and A. Menezes, “Generative models of noisy translations with applications to parallel fragment extraction,” in *Proc. of MT Summit*, 2007.

[6] D. Blei, A. Ng, and M. Jordan, “Latent Dirichlet Allocation,” in *Journal of Machine Learning Research*, 2003, pp. 1107–1135.

[7] R. Kneser, J. Peters, and D. Klakow, “Language model adaptation using dynamic marginals,” in *Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH)*, 1997, pp. 1971–1974.

[8] Y. C. Tam and T. Schultz, “Correlated latent semantic model for unsupervised language model adaptation,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

[9] S. Vogel, “PESA: Phrase pair extraction as sentence splitting,” in *Proc. of MT Summit*, 2005.

[10] S. Vogel, “SMT decoder dissected: Word reordering,” in *Proc. of ICNLPKE*, 2003.