Improve SMT with Source-Side "Topic-Document" Distributions

Zhengxian Gong Guodong Zhou Liangyou Li

School of Computer Science and Technology
Soochow University, Suzhou, China 215006
{zhxgong, gdzhou, 20104227013}@suda.edu.cn

Abstract

Topic modeling is a popular framework to analyze large text collections. In the previous work, employing topic modeling into statistic machine translation mainly depends on one major topic of the test document. Different from the previous work, the proposed approaches will coverage not only major topic but also sub-topics. The basic idea of this paper is assumed that better translation quality, closer similarity of "topic-document" distributions between the target-side and the source-side documents. We first give some initial experimental results to support this assumption. Then we transfer generating such a target document into selecting target-side sentences by an effective algorithm. A preliminary study showed that enforcing "topic-document" distributions to be consistent between target-side and source-side in SMT can potentially improve translation quality.

1 Introduction

Topic modeling is a popular framework to analyze large text collections, which softly cluster documents and terms into a fixed number of topics. There are a few studies on employing topic modeling into Statistical Machine Translation (SMT).

Zhao et al. (2006) assumed that the training corpus are composed of documents, and proposed a model called "BiTAM" to improve the performance of word alignment, which consists of topic-dependent translation lexicons modeling p(c|e, k), where c, e and k denote the source word, target word and the topic index respectively.

Tam et al. (2007) first built the source-side and target-side topic models respectively, ie. p(c|k), p(e|k). Then, they proposed a bilingual-LSA model to automatically build the one-to-one correspondence between the source and target topic models.

Foster et al.(2007) described a mixture-model approach to implement a statistical machine translation system for new domains. Such approaches normally first partition the training data into different specific domains, then train a sub-model on each specific domain and finally combine a specific domain translation model with a general domain translation model depending on various text distances, one of which is using Latent Semantic Analysis (LSA, one of topic modeling methods).

In the previous work, they dynamically choose translation model or language model according to the major topic of the test document, in some sense, they regarded "topic" as "domain". However, our proposed approaches will simultaneously coverage multiple topics of a test document, including not only the major topic but also some sub-topics.

Due to use parallel document pairs, we follow the following assumption: better translation quality, closer similarity of "topic-document" distributions between the target-side and the source-side documents. First, some initial experimental results are given to support this assumption. For each source-side sentence, we obtain a ranked N-best list of candidate translations in the target language based on a baseline system. It notes these sentences all belong to one document. After that, an effective algorithm for selecting target sentences to compose a target document with the minimal deviation to the "topic-document" distributions passed from the source-side.

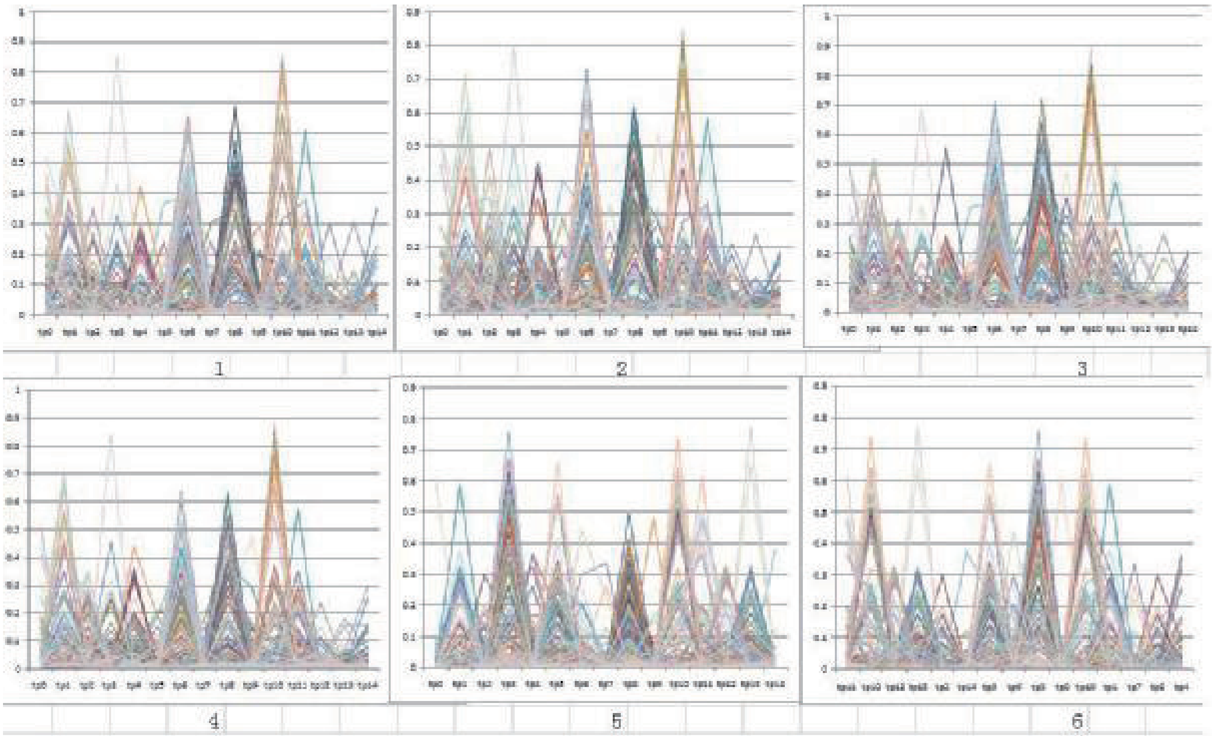

Figure 1. initial experimental results

The rest of this paper is as follows. Section 2 gives initial experimental results. Section 3 describes our document-level system framework and introduces a baseline system. Section 4 presents an effective algorithm for target sentences selection. Section 5 presents experimental results. Finally, Section 6 draws a conclusion.

2 Initial Experiment

2.1 Corpus

In this paper, we use FBIS as the training data, the 2003 NIST MT evaluation test data as the development data, and the 2005 NIST MT test data as the test data. Table 1 shows the statistics of these data sets (all these data with document boundaries annotation).

Corpus		Sentences	Documents
Role	Name		
Train	FBIS	239413	10353
Dev	NIST2003	919	100
Test	NIST2005	1082	100

Table 1: Corpus statistics

2.2 Monolingual Topic Modeling

Among various topic models, Latent Dirichlet Allocation (LDA, Blei et al., 2003) has drawn most attention recently in the NLP community and has been applied successfully in topic detection. In principle, LDA is a generative three-level hierarchical Bayesian probabilistic model for analyzing the content of documents and the meaning of words. Similar to other topic models, such as LSA, PLSA and PLSI, LDA assumes that documents are mixtures of topics and a topic can be represented as a probability distribution over words. In this paper, we use LDA to capture the topics in a document.

We first use a LDA tool[1] to train topic models for the source-side(Chinese) and target-side(English) documents respectively in our training parallel corpus, FBIS, with a fixed number K(tuned to 15) topics.

Using LDA, for each word w, we can obtain the "topic-word" distribution $p(w|z_i)$ (topic $z_i \in K$), and the "topic-document" distribution $P(z_i|d)$ for each document d. Moreover, using the obtained LDA models, we can infer the topic distributions of a new test document, namely $p(z_i \,|d_{new})$ for each topic $z_i \in K$.

[1] http://www.arbylon.net/projects/

2.3 Initial Experiments Results

In this section, we test the assumption that better translation quality, closer similarity of "topic-document" distributions between the target-side and the source-side documents.

In our test corpus, there are 100 documents (showed in table 1) and each source-side document has four reference texts (target-side documents) according to different system ID, we inference "topic-document" distributions for each reference documents (the 1-4 part of Figure 1) . We also inference "topic-document" distributions for 100 source-side documents based on trained source-side LDA modeling (the 5 part of Figure 1).

At the first sight, there is no obvious relation between 1 (or 2, 3, 4) and 5. But if re-arrange the topic sequence according to the "correspondence" (described in the next section 2.4) between the source-side and target-side topic modeling, we can observe that there are very close similarity distributions between source-side and target-side(the 6 part of Figure 1).

From the above experimental results, we can slightly testify the correctness of our assumption in an empirical way. Now, we should address the problem on how to build the "correspondence" between the source-side and target-side topic modeling.

2.4 Correspondence for monolingual topic modeling

As described in Section 2.2, we trained monolingual LDA models for source-side language and target-side language respectively in advance. In order to supervise decoder to generate final target-side translation text according to the inferred topic distributions from the source-side document, we need a bridge, which is similar to Tam's work (Tam et al.,2007) by enforcing a one-to-one topic correspondence between the source and the target LDA-style models.

However, we found the rigorous "one-to-one" topic correspondence cannot be obtained in our experiment. The phenomenon of mismatch exists in our experiment, for example, there are two topics in the source-side which can't match any topic in the target side, and in the same time, there are two topics in the source-side can be matched with one target-side topic twice.

The reasons for why we cannot build the rigorous "one-to-one" correspondence maybe have two aspects: (1) the scale of corpus is not enough large; (2) the effect of "polysemy" and "synonymy" between different languages is different.

So here we propose a simpler approach for small number topics like this:

1) Using GIZA++(Och and Ney,2000) in two directions to perform word alignment on parallel corpus, and augmented to improve recall using the grow-diagonal-final heuristic.
2) Choose the top-n (n is fine-tuned to 200 in this paper) word-topic distribution of each topic in both languages.
3) With the help of lexical mapping (obtained from Step 1), pairwise comparison is performed based on Step 2. We count the mapping words between two topics in both languages and determine the mapping relation according to the maximum numbers of mapping.

3 Document-level SMT

3.1 A Phrase-based Baseline

It is well known that the translation process of SMT can be modeled as obtaining the best translation **e** of the source sentence **f** by maximizing following posterior probability (Brown et al., 1993):

$$e_{best} = \arg\max_{e} P(e \mid f) = \arg\max_{e} P(f \mid e) P_{lm}(e) \quad (1)$$

where P(e|f) is a translation model and P_{lm} is a language model. For each sentence in the source language(f), we can obtain a ranked n-best list of candidate translations in the target language based on a baseline system. Usually, we say the top 1 of the N-best translations is a best translation.

Our system adopted Moses (a state-of-art phrase-based SMT system) as a baseline, which follows Koehn et al. (2003).

3.2 The workflow of document-level SMT

Given a test document, our system works as follows:

1) Inference a new test document d_{new} based on obtained LDA models and get "topic-document" distributions P(z_{si}| d_{new_s}) for the source-side document, here i <=K and the footnote of *s* means source-side;

2) Based on initial experimental results, the value of P(z$_{ti}$| d_{new_t})(here d_{new_t} is unknown) is approximate to P(z$_{si}$| d_{new_s}) according to the correspondence for monolingual topic modeling, here *t* means target-side. And thus get P(*w* |z$_{ti}$) based on trained target-side LDA models;
3) For each source-side sentence in one document, we obtain a ranked N-best list of candidate translations in the target language from Moses;
4) Selecting new 1-best candidate translation for each source-side sentence to compose a target-side document d_{new_t} which has maximum similarity topic-distributions to previous P(z$_{ti}$| d_{new_t}).

Obviously, the fourth step is a key step. Given the target document, it will be easy to gain the "topic-document" distributions based on previous trained target-side topic model. However, such target document does not exist because our decoder needs to translate sentence by sentence until to the last sentence. For a test document, with M source-side sentences, and each source-side sentence corresponds to N-best list of candidate translations, there will be N^M target documents waiting for determining. With the growth of N and M, the computational complexity is too high.

4 Generate target-side document

In this section, we propose to transfer the task of generating optimum target document into selecting better sentences.

4.1 Transfer generating optimum target document into selecting sentences

Assumed *H* represents the faithful target document's probability distributions over topics, i.e.

$$H = (P(T_1 \mid D_i), P(T_2 \mid D_i), \ldots, P(T_k \mid D_i)),$$

P(T$_j$|D$_i$) stands for the probability of being topic T$_j$ given document D$_i$. Due to the parallel corpus, we also think the target-side "document-topic" distributions are similar to the source-side ones (the initial experiment results showed in Section 2.3). So we assume H is constant here.

Q represents one target document probability distribution over topics, i.e.

$$Q = (P(T_1 \mid D_x), P(T_2 \mid D_x), \ldots, P(T_k \mid D_x)),$$

where *P(T$_j$|D$_x$)* stands for the probability of topic *T$_j$* given one target document *D$_x$*. Now we mainly manage how to construct *Q*.

Using the Bayes rule, we have

$$P(T_j \mid D_x) = \frac{P(D_x \mid T_j) * P(T_j)}{P(D_x)} \qquad (2)$$

where

- P(D$_x$|T$_j$) stands for the probability that topic T$_j$ generates document D$_x$.
- P(T$_j$) stands for the probability of Topic T$_j$.
- P(D$_x$) stands for the probability of document D$_x$.

Let's assume that a sentence S$_r$ of a document D$_x$ represents a topic T$_j$ if the topic T$_j$ generates all the words of the sentence S$_r$ with some probability and that the document D$_x$ generates Topic T$_j$. Under this assumption, we have:

$$P(D_x \mid T_j) = \frac{1}{\| Dx \|} \sum_{Sr \in Dx} P(Sr \mid Tj) \qquad (3)$$

where ||D$_x$|| stands for the number of sentence in document D$_x$.

For the same reason, we extend one sentence into some words by the Equation (4):

$$P(Sr \mid Tj) = \frac{1}{\| Sr \|} \sum_{Wi \in Sr} P(Wi \mid Tj) \qquad (4)$$

where ||S$_r$|| stands for the number of words in sentence S$_r$.

It notes that P(W$_i$|T$_j$) stands for the probability that topic T$_j$ generates word W$_i$, which has been obtained by the previous trained target-side topic models.

Furthermore, by applying Equation 3 and 4 to Equation 2, we can get:

$$P(T_j \mid D_x) = \frac{P(T_j)}{P(D_x)} * \frac{1}{\| Dx \| * \| Sr \|} \sum_{Sr \in Dx} \sum_{W_i \in S_r} P(W_i \mid T_j) \qquad (5)$$

where P(T$_j$) and P(D$_x$) are constant. In our case, for the sake of simplicity, we set P(T$_j$) /P(D$_x$) as 1.

In this paper, we mainly investigate Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951) to measure the distance between two probability distributions as follows:

$$D_{KL}(H \| Q) = \sum_i H(i) \log \frac{H(i)}{Q(i)} \qquad (6)$$

4.2 Algorithm

We implement our system by the Algorithm 1 based on the above transformation.

The basic idea of Algorithm 1 is: assume the “topic-document” distributions are constant, update them with the distributions of continuous sentences, if the candidate leads to the minimal change among these distributions, we will choose it.

It need to point out the score(h) should be constrained(we set ξ=1), or it is possible to choose the translation candidate which contains plenty of topic words however the fluency of the whole sentence is very poor. The value of score(h) is an absolute value using the score of language model of original top 1 translation candidate in the N-best lists to minus the corresponding score of current candidate.

For the two topics in the source-side which can’t match any topic in the target side (showed in Section 2.4), we will set its corresponding $P(W_i|T_j)$ as $P(W_{si}|T_{sj1})$ and $P(W_{si}|T_{sj1})$, W_{si} is a source-side word ;T_{sj1} and T_{sj2} is the two mismatched source-side topics. For the topic in the source-side matched with multiple target-side topics, we determine it by looking through the top 1 distribution of “word-topic” for each topic. The highest value will be adopted.

```
Input: The inferred latent “topic-document” distributions H from source-side;
       The N-best list of translation candidate
Output: The new 1-best list of translation candidate

$Q=H; Q_{max}=Q; \Phi=\{\};$
For each source-side sentence
    $Q' = Q_{max};$
For each target-side translation candidate $h$
    Obtain Q’ by updating Q with Equation (5)
    If minimize $D_{KL}(H||Q')$ and Score(h)>$\xi$ then
        $Q_{max}=Q'; \Phi=\Phi\cup\{h\}$
End For
End For
Output $\Phi$
```

Algorithm 1

5 Experimentation and Discussion

5.1 Experimental Setting

Here, we use SRI language modeling toolkit to train a trigram general language model on English newswire text, mostly from the Xinhua portion of the Gigaword corpus (2007) and performed word alignment on the training parallel corpus using GIZA++(Och and Ney,2000) in two directions. For evaluation, the NIST BLEU script (version 13) with the default setting is used to calculate the Bleu score (Papineni et al. 2002), which measures case-insensitive matching of n-grams with n up to 4.

5.2 Experimental Results

The column of “BLEU_W” in Table 2 means the BLEU score computed over the whole test set and “BLEU_D” corresponds to the average BLEU score over separated documents. As reported in Table 2, small increases in BLEU (0.45 in “BLEU_W”) can be obtained by our approach.

System	BLEU on Test(%)	
	BLEU_W	BLEU_D
Moses	25.76	25.08
Ours	26.21	25.36

Table 2 experiment results

5.3 Discussion

There are 1082 sentences in our test data in sum (Table 1 showed in Section 2.1), and only 407 sentences selection make a difference, which only refers to 31 documents. For each sentence pair, we determined whether our additional processing improved or degraded performance compared to Moses output. Among the sentence pairs, 82 sentences do not change in fact, 304 sentences obtain positive change (examples showed in Table 3) and others 21 sentences degrades the performance.

We performed some manual analysis of the output. We observe that such phenomenon can broadly be attributed to two reasons:

1) The performance of baseline;

2) The negotiation between the score of language model and topic model;

Our proposed method based on the N-best list which produced by the baseline. If there is no good candidate waiting for choice, our method will not work effectively. The 82 sentences without change majorly belong to the reason 1 because there is no real change between the candidates in the N-best list. In the future work, we should integrate source-side “topic-document” distributions into our decoder.

For the reason 2, we did intervention by considering the deviation of language model score (see

1	他们 说 ， 欧元 区 最 大 国家 的 经济 产出 数据 将 持续 显示 每个 月 波动 极大 ，……。
	Moses: they said that the euro zone 's biggest country 's economic output **figures will show every month** , ...
	Ours:they said that the euro zone 's biggest country 's economic output **data will continue to demonstrate every month** ,...
	Reference: They said economic output data from the largest Eurozone countries would continue to indicate great volatility each month , …
2	乌克兰 危机 引发 东西 紧张 将 成 欧安组织 议题 。
	Moses: ukraine crisis **triggered** tension will be a topic **that** .
	Ours: ukraine crisis **triggering** tension will be a topic of **the things** .
	Reference:Triggering Tensions between East and West , Crisis in Ukraine Will Become an OSCE Topic .
3	俄罗斯 多次 指控 西方 插手 东欧 事务 ， 此事 因为 乌克兰 的 政治 危机 益 成 关注 焦点 ，……。
	Moses: russia has repeatedly accused **of meddling in west europe affairs** , the political crisis that ukraine will become the focus of attention ,…
	Ours: russia has repeatedly accused **western intervention in eastern europe affairs** , the political crisis that ukraine will become the focus of attention ,...
	Reference: Repeated accusations by Russia of Western intervention in Eastern European affairs have increasingly been a focus because of the political crisis in Ukraine ,...

Table 3: Positive examples

score(h) described in 4.2), however, it is not easy to set the reasonable threshold value(ξ) . Maybe some re-rank algorithm can be introduced here.

6 Conclusion

Based on the assumption that better translation quality, closer similarity of “topic-document” distributions between the target-side and the source-side documents, we obtained some small improvement results for statistical machine translation system. In this paper, we only implemented this assumption during the post-edit procedure. So if the quality of N-best translation is poor, our proposed method will lose effectiveness. In our feature work, we will implement this during decoding and design the corresponding MERT algorithm.

Acknowledgments

This research was supported by Projects 90920004, 60873150, and 61003155 under the National Natural Science Foundation of China, Project 20093201110006 under the Specialized Research Fund for the Doctoral Program of Higher Education of China.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. *Latent Dirichlet Allocation*. Journal of machine Learning Research, pages 993–1022.

PF Brown, SA Della Pietra, VJ Della Pietra, RL Mercer.1992.*The Mathematics of Statistical Machine Translation: Parameter Estimation*. Computational Linguistics. 19(2):263-309.

John DeNero, Alexandre Buchard-Cˆot´e, and Dan Klein. 2008. *Sampling Alignment Structure under a Bayesian Translation Model*. In Proc. of EMNLP 2008, pages 314–323, Honolulu, October.

Rukmini M. Iyer and Mari Ostendorf. 1999. *Modeling Long Distance Dependence in Language:Topic Mixtures Versus Dynamic Cache Models*. IEEE Transactions on speech and audio processing, 7(1).

Philipp Koehn, Franz Josef Och ,and DanielMarcu.2003. *Statistical Phrase-Based Translation*. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pages 48-54.

Philipp Koehn.2004.*Statistical Significance Tests for Machine Translation Evaluation*. In Proc. of EMNLP 2004, pages 388–395.

Daniel Marcu and William Wong. 2002. *A phrase-based Joint Probability Model for Statistical Machine Translation*. In Proc. of EMNLP 2002, July.

Franz Josef Och. 2003. *Minimum Error Rate Training in Statistical Machine Translation*. In Proc. of ACL2003, pages 160–167.

Franz Josef Och and Hermann Ney. 2000. *Improved Statistical Alignment Models*. In Proc. of ACL2000, pages 440–447.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-jing Zhu. 2002. *BLEU: A Method for Automatic Evaluation of Machine Translation*. In Proc. of ACL2002, pages 311–318.

Yik-Cheung Tam, Ian Lane and Tanja Schultz. 2007. *Bilingual lSA-based Adaptation for Statistical Machine Translation*. Machine Translation, 28:187-207.

Bing Zhao and Eric P. Xing .2006. *BiTAM:Bilingual Topic Ad-Mixture Models for Word Alignment*. In Proc. of ACL2006.