Institute of Automation
Chinese Academy of Sciences
中国科学院自动化研究所

Learning Latent Topic Information for Language Model Adaptation in Statistical Machine Translation

Shixiang Lu, Wei Wei, Xiaoyin Fu, Lichun Fan, Bo Xu
Interactive Digital Media Technology Research Center
Nov. 4th, 2012

Outline

- Task
- Problems
- Our Approach
- Experiments
- Conclusion

Outline

- Task
- Problems
- Our Approach
- Experiments
- Conclusion

Language Model Adaptation in SMT

- Two category
  - **Selecting similar training data for current translation task**
  - **Modify the LM parameter itself**

Language Model Adaptation in SMT

- Two category
  - **Selecting similar training data for current translation task；**
  - **Modify the LM parameter itself**

Outline

- Task
- Problems
- Our Approach
- Experiments
- Conclusion

Data selection based LM Adaptation

- Framework
  - **The generic LM** $P_g(w_i|h)$ **;**
    - Estimated by the whole LM training corpus
  - **The bias LM** $P_b(w_i|h)$ **;**
    - Estimated by the selected similar training data from LM training corpus
  - **The adapted LM** $P_a(w_i|h)$ **;**
    - used for SMT system

$$P_a(w_i|h) = \gamma P_g(w_i|h) + (1-\gamma) P_b(w_i|h)$$

How to select the similar training data?

- Monolingual approach
  - **First pass translation hypotheses s**
- Cross-lingual approach
  - **Candidate translation sentence s**
- Framework

  The sentence in the LM training corpus S

  Similarity(s,S): The similarity between s and S

Data selection based LM Adaptation

- Traditional approach
  - **TF-IDF;**
  - **Centroid similarity;**
  - **Cross-entropy difference;**
  - **Cross-lingual information retrieval;**
  - **Cross-lingual similarity**

Data selection based LM Adaptation

- Traditional approach
  - **TF-IDF;**
  - **Centroid similarity;**
  - **Cross-entropy difference;**
  - **Cross-lingual information retrieval;**
  - **Cross-lingual similarity**
- General Characteristics
  - **Perform at the word level**
  - **Exact only term matching schemes**

Data selection based LM Adaptation

- Traditional approach
  - **TF-IDF;**
  - **Centroid similarity;**
  - **Cross-entropy difference;**
  - **Cross-lingual information retrieval;**
  - **Cross-lingual similarity**
- General Characteristics
  - **Perform at the word level**
  - **Exact only term matching schemes**
- Main problem
  - **Do not consider word-topic and word-distribution information**

Outline

- Task
- Problems
- Our Approach
- Experiments
- Conclusion

Latent Topic Based Data Selection Model

- Learning word-topic and word-distribution information by Latent Dirichlet Allocation (LDA) Model
  - **Step 1, estimate word-topic information of the whole LM training corpus;**
  - **Step 2, infer the word-topic information first pass translation hypotheses ;**
  - **Step 3, compute the similarity and select the similar sentences.**

Latent Topic Based Data Selection Model

- Learning word-topic and word-distribution information by Latent Dirichlet Allocation (LDA) Model
  - **Step 1, estimate word-topic information of the whole LM training corpus;**
  - **Step 2, infer the word-topic information first pass translation hypotheses ;**
  - **Step 3, compute the similarity and select the similar sentences.**
- Similarity Computation
  - **Step 1, compute the sentence-topic** $P_{LT}(z|s) = \frac{1}{|s|}\sum_{w \in s} P(z|w)$
  - **Step 2, compute topic based sentence similarity**

$$
\begin{aligned}
P_{LT}(s|S) &= \sum_{z} P_{LT}(s|z)P_{LT}(z|S) \\
&= \sum_{z \in K} \frac{P_{LT}(z|s)P(s)}{P(z)} P_{LT}(z|S) \\
&= \frac{K}{N} \sum_{z \in K} P_{LT}(z|s)P_{LT}(z|S)
\end{aligned}
$$

Parameter Estimation (LDA model)

- How to select proper topic number?

$$Perplexity(C) = exp\{-\frac{\sum_{(s,w)\in C} lnP(w|s)}{|C|}\}$$

$$P(w|s) = \sum_{z=1}^{K} P(w|z)P(z|s)$$

Latent Topic based Cross-Lingual Data Selection Model

- Sentence projection
  - **u: source sentence; v: target sentence;**
  - Σ **:bilingual word co-occurrence matrix**
  - **Computing projection:**

$$\hat{v} = u\Sigma$$

$\hat{v}$ is emphasized that most frequently co-occur with the source term in u

Latent Topic based Cross-Lingual Data Selection Model

- Sentence projection
  - **u: source sentence; v: target sentence;**
  - Σ **:bilingual word co-occurrence matrix**
  - **Computing projection:**

$$\hat{v} = u\Sigma$$

$\hat{v}$ s emphasized that most frequently co-occur with the source term in u

- Cross-lingual sentence similarity

$\hat{v}$ **as the first pass translation hypotheses** $\hat{s}$,

$$P_{CLLT}(\hat{s}|S) = \frac{K}{N}\sum_{z\in K} P_{CLLT}(z|\hat{s})P_{CLLT}(z|S)$$

Combining Latent Topic with TF-IDF for Data Selection

- Combine these two approach together

$$P_{LT\_TF-IDF}(s|S) = \mu P_{LT}(s|S) + (1-\mu)P_{TF-IDF}(s|S)$$

$$P_{CLLT\_CLS_s}(\hat{s}|S) = \lambda P_{CLLT}(\hat{s}|S) + (1-\lambda)P_{CLS_s}(\hat{s}|S)$$

Outline

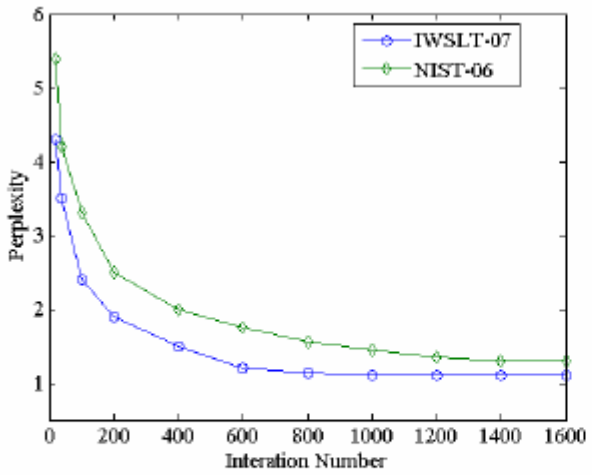
- Task
- Problems
- Our Approach
- Experiments
- Conclusion

Framework

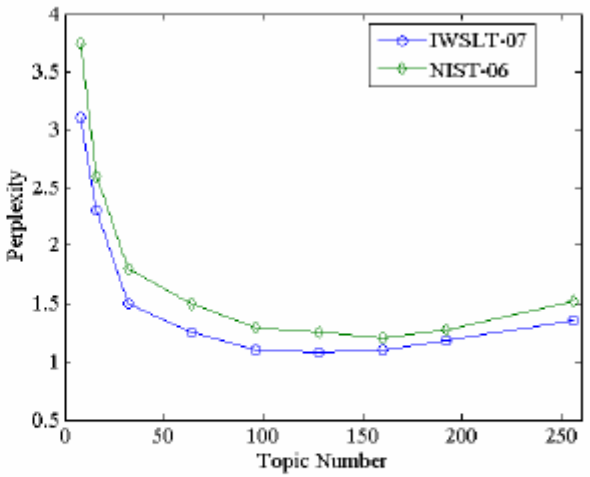
- Comparing adapted LMs with the generic LM
  - **reference translations based perplexity**
  - **SMT performance**
- Data sets
  - **IWSLT-07 (dialogue domain)**
  - **NIST-06 (news domain)**

Iteration and Topic Number Selection

- IWSLT-07 96 topics and 1000 iterations
- NIST-06 168 topics and 1400 iterations



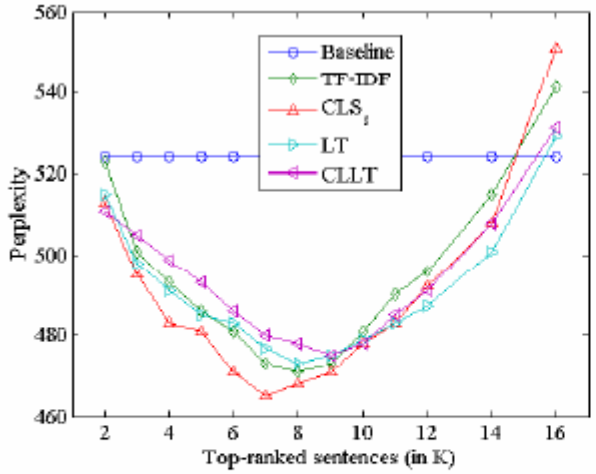
(a) interation number



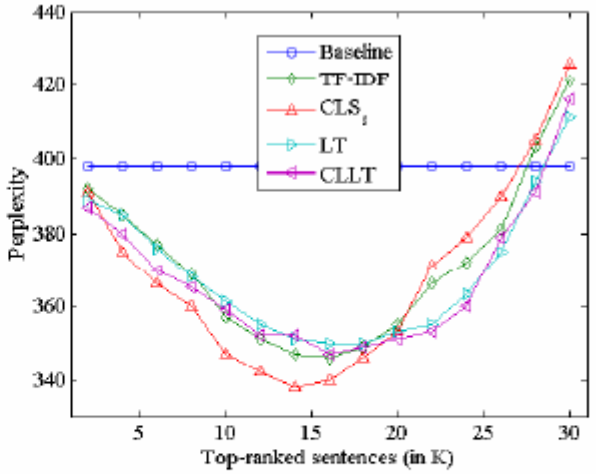
(b) topic number

Fig. 1: Perplexity vs. the number of different iterations and topics on two LM training corpus.

Perplexity Analysis



(a) IWSLT-07



(b) NIST-06

Fig. 2: English reference translation based perplexity of adapted LMs vs. the size of selected data on two test sets.

Perplexity Analysis

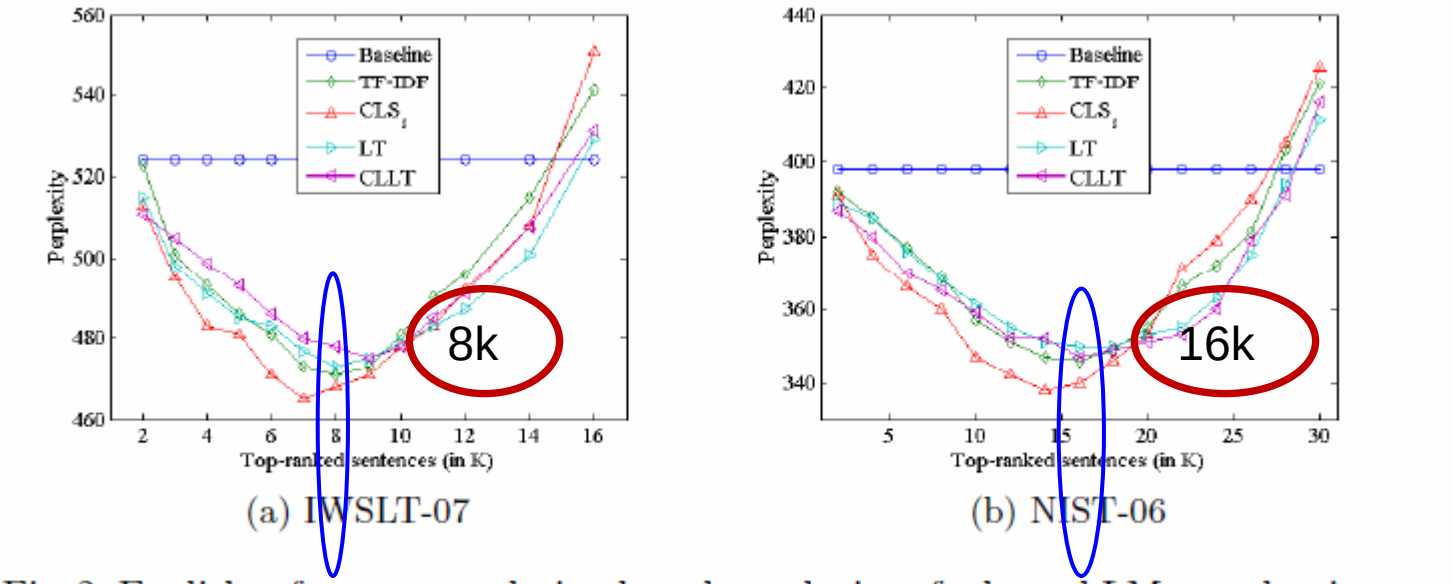


Fig. 2: English reference translation based perplexity of adapted LMs vs. the size of selected data on two test sets.

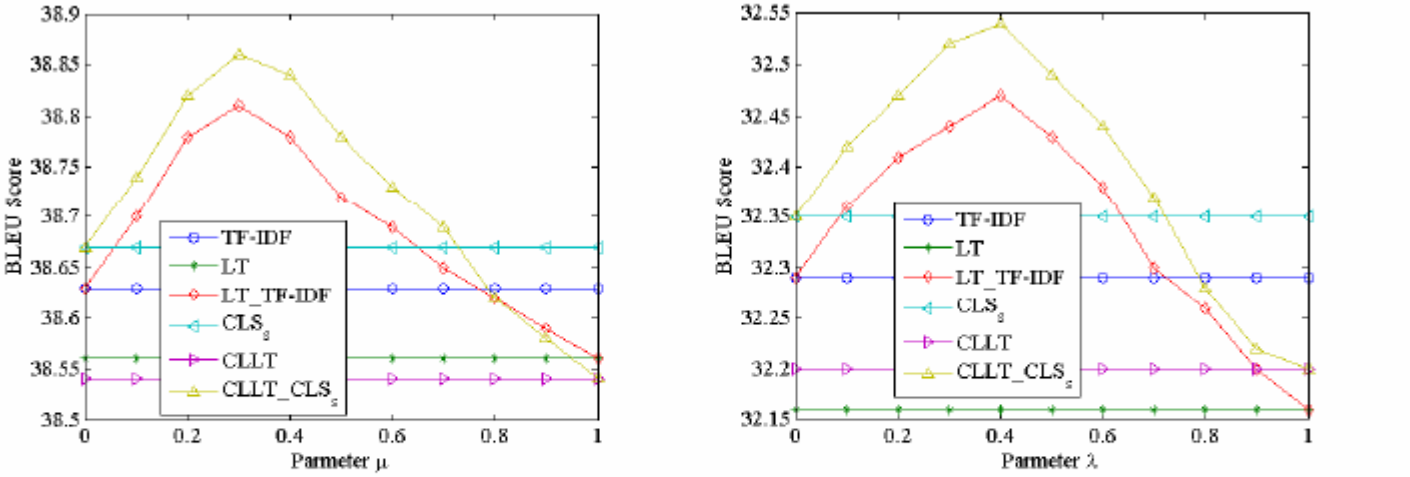
- IWSLT-07 : top 8K sentences
- NIST-06: top 16K sentences

Translation Experiments

- Tune parameter

$$P_{LT\_TF-IDF}(s|S) = \mu P_{LT}(s|S) + (1-\mu) P_{TF-IDF}(s|S)$$

$$P_{CLLT\_CLS_s}(\hat{s}|S) = \lambda P_{CLLT}(\hat{s}|S) + (1-\lambda) P_{CLS_s}(\hat{s}|S)$$



(a) IWSLT-07 (b) NIST-06

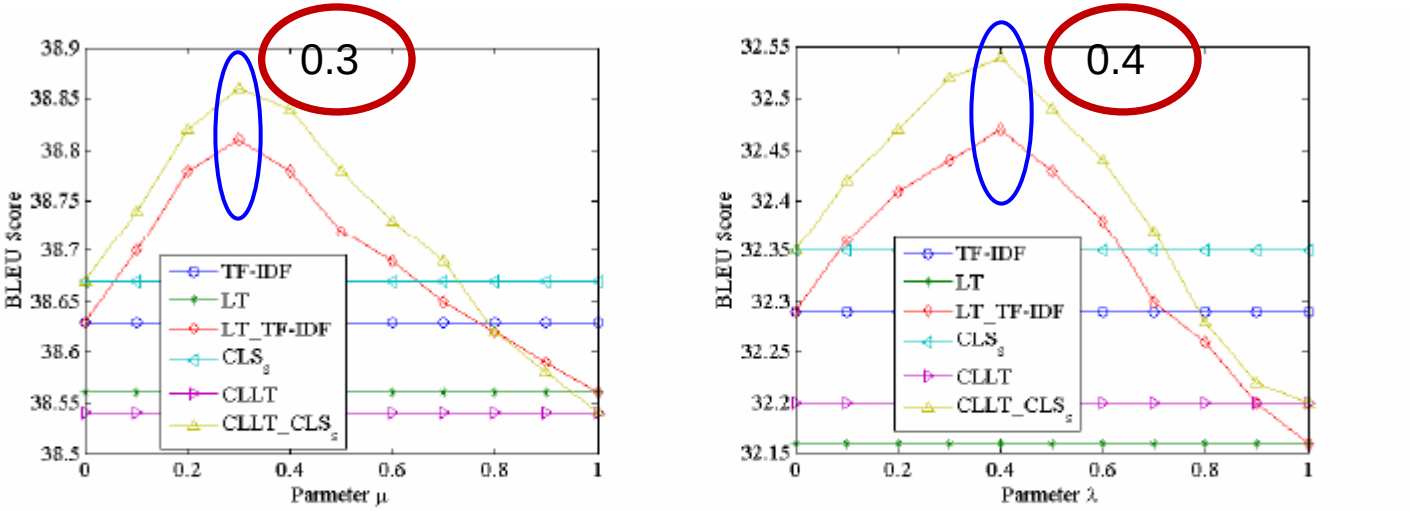
Fig. 3: The impact of parameters μ and λ to SMT performance on two development sets.

Translation Experiments

- Tune parameter

$$P_{LT\_TF-IDF}(s|S) = \mu P_{LT}(s|S) + (1-\mu)P_{TF-IDF}(s|S)$$

$$P_{CLLT\_CLS_s}(\hat{s}|S) = \lambda P_{CLLT}(\hat{s}|S) + (1-\lambda)P_{CLS_s}(\hat{s}|S)$$



(a) IWSLT-07 (b) NIST-06

Fig. 3: The impact of parameters μ and λ to SMT performance on two development sets.

Translation Experiments

Table 1: SMT performance with different data selection models for LM adaptation on two test sets.

Method	#	BLEU	
		IWSLT-07	NIST-06
Baseline	1	33.60	29.15
TF-IDF	2	34.14	29.78
CLS	3	34.08	29.73
CLS_s	4	34.18	29.84
LT	5	34.07	29.65
CLLT	6	34.05	29.69
LT_TF-IDF	7	**34.32**	**29.96**
CLLT_CLS_s	8	**34.37**	**30.03**

- CLSs performs better than CLS
  - **The added smoothing measure which makes similarity computation more accurate**

Translation Experiments

Table 1: SMT performance with different data selection models for LM adaptation on two test sets.

Method	#	BLEU	
		IWSLT-07	NIST-06
Baseline	1	33.60	29.15
TF-IDF	2	34.14	29.78
CLS	3	34.08	29.73
CLS_s	4	34.18	29.84
LT	5	34.07	29.65
CLLT	6	34.05	29.69
LT_TF-IDF	7	**34.32**	**29.96**
CLLT_CLS_s	8	**34.37**	**30.03**

- LT and CLLT do not outperform the baseline method TF-IDF

Translation Experiments

Table 1: SMT performance with different data selection models for LM adaptation on two test sets.

Method	#	BLEU	
		IWSLT-07	NIST-06
Baseline	1	33.60	29.15
TF-IDF	2	34.14	29.78
CLS	3	34.08	29.73
CLS_s	4	34.18	29.84
LT	5	34.07	29.65
CLLT	6	34.05	29.69
LT_TF-IDF	7	**34.32**	**29.96**
CLLT_CLS_s	8	**34.37**	**30.03**

- LT_TF-IDF significantly outperforms LT and TF-IDF
  - **The word-topic and word-distribution information is complementary to TF-IDF**

Translation Experiments

Table 1: SMT performance with different data selection models for LM adaptation on two test sets.

Method	#	BLEU	
		IWSLT-07	NIST-06
Baseline	1	33.60	29.15
TF-IDF	2	34.14	29.78
CLS	3	34.08	29.73
CLS_s	4	34.18	29.84
LT	5	34.07	29.65
CLLT	6	34.05	29.69
LT_TF-IDF	7	**34.32**	**29.96**
CLLT_CLS_s	8	**34.37**	**30.03**

- CLLT_CLSs outperforms LT TF-IDF, and CLSs outperforms TF-IDF
  - **first pass translation hypotheses have lots of noisy data**
  - **cross-lingual data selection model can avoid this problem**

Outline

- Task
- Problems
- Our Approach
- Experiments
- Conclusion

Conclusion

- **Word-topic and word-distribution information is very useful for data selection based LM adaptation**
- **Word-topic information is complementary to traditional approach**

Thanks !

Any questions?