A Topic-Triggered Language Model for Statistical Machine Translation

Give me a shot

- **Film topic**: 照相 (photo)
- **Oral topic**: 机会 (chance)
- **Military topic**: 射击 (gun shoot)
- **Sports topic**: 进球 (goal)

Heng Yu  Jinsong Su  Yajuan Lv  Qun Liu

Chinese Acad. of Sciences  xmu  CAS.  CAS. & DCU
Language model for SMT

• N-gram model has been dominant in SMT
  • Markov Assumption
  • Use only N-1 words of prior context
• Shall we add more contextual information?
  • syntax-based language model (charniak+ 03, shen+ 08)
  • semantic-based language model (Tan+ 11)
  • MI-based language model (Xiong+ 11)
• Our focus: add global semantic information into LM
N-gram Model
Big Picture

N-gram Model

adding more context
Big Picture

N-gram Model

adding more context

Syntax-based LM

Semantic-based LM

MI-based LM
Big Picture

N-gram Model

adding more context

Syntax-based LM

Semantic-based LM

MI-based LM

This work: Topic-triggered LM for SMT
Why we need topic information

- Different topic leads to different translations
  - N-gram context can’t capture global topic
  - Sentence-level syntax can’t help.
- The intuition is incorporate topic information into LM
  - Simplicity: compatible with N-gram model
  - Tailored for SMT: online estimation
Our approach

- use a variant of topic model: HTMM (Gruber et al 07)
  - estimate sentence level topic distribution
- estimate N-gram probability in different topic context
- Online adapt to different topic distribution
- result: +0.76 Bleu improvement on CH-En task
Outline

- Motivations

- Topic-specific language model
  - Estimate Topic distribution
  - Calculate topic-based language model probability

- Online adapting to different topics
  - Topic Projection
  - Topic-triggered estimation

- Experiments
Topic Distribution

- Topic distribution: the probabilities of a sentence appear in a certain topic
- document level context
Bush held a meeting with sharon
politics or military
• Topic distribution: the probabilities of a sentence appear in a certain topic
• document level context

Bush held a meeting with Sharon politics or military
Give me a shot various topics
Estimating topic distribution

- Conventional method
  - LDA: Bag of words assumption (Blei et al. 2003)
  - PLSA (Hofmann, 1999)
  - Disadvantage:
    - Ignore the order of words
    - Only suit for 1-gram probability estimation
- HTMM (Gruber et al, 07)
  - All words in one sentence share the same topic
  - The topics of sentences are modeled as Markov Chain
Calculating topic-based LM probability

• add topic of current sentence($t$) as Latent variable

$$P(e) = \sum_t P(e, t) = \sum_t P(e|t) \cdot P(t)$$

• Apply to N-gram model

$$P(e|t) = P(w_1|t) \cdot P(w_2|w_1, t) \cdots P(w_i|w_{i-n+1}, t)$$

• use Maximum Likelihood Estimation and WB smoothing
Language Model training

- Train language model for each topic

Corpus → Topic Distribution → Topic specific LMs

Friday, November 15, 13
Outline

• Motivations

• Topic-specific language model
  • Estimate Topic distribution
  • Calculate topic-based language model probability

• Online adapting to different topics
  • Topic Projection
  • Topic-triggered estimation

• Experiments
Online Adaptation for SMT
Online Adaptation for SMT

- Challenge: No target side before translation
  - impossible to estimate target side topic distribution
Online Adaptation for SMT

- Challenge: **No target side** before translation
  - impossible to estimate target side topic distribution

- Solution: **Topic Projection**
  - estimate source side topic distribution
  - use projection matrix to project to target side topic space
Online Adaptation for SMT

• Challenge: No target side before translation
  • impossible to estimate target side topic distribution

• Solution: Topic Projection
  • estimate source side topic distribution
  • use projection matrix to project to target side topic space

\[
P(e) = \sum_{t_e} P(e|t_e) \cdot \sum_{t_f} P(t_e|t_f) \cdot P(t_f)
\]
Online Adaptation for SMT

- Challenge: **No target side** before translation
  - impossible to estimate target side topic distribution

- Solution: **Topic Projection**
  - estimate source side topic distribution
  - use projection matrix to project to target side topic space

\[
P(e) = \sum_{t_e} P(e|t_e) \cdot \sum_{t_f} P(t_e|t_f) \cdot P(t_f)
\]
Online Adaptation for SMT

- **Challenge:** No target side before translation
  - impossible to estimate target side topic distribution
- **Solution:** Topic Projection
  - estimate source side topic distribution
  - use projection matrix to project to target side topic space

\[
P(e) = \sum_{t_e} P(e|t_e) \cdot \sum_{t_f} P(t_e|t_f) \cdot P(t_f)
\]
Online Adaptation for SMT

- **Challenge:** No target side before translation
  - impossible to estimate target side topic distribution
- **Solution:** Topic Projection
  - estimate source side topic distribution
  - use projection matrix to project to target side topic space

\[
P(e) = \sum_{t_e} P(e|t_e) \cdot \sum_{t_f} P(t_e|t_f) \cdot P(t_f)
\]
Topic Projection

- Build Project Matrix from parallel corpus
  - Learn topic model on both side
  - collect topic co-occurrence via alignment
  - estimate project probability using MLE
Topic Projection

- Build Project Matrix from parallel corpus
  - Learn topic model on both side
  - collect topic co-occurrence via alignment
  - estimate project probability using MLE
Topic Projection

- Build Project Matrix from parallel corpus
  - Learn topic model on both side
  - collect topic co-occurrence via alignment
  - estimate project probability using MLE
Topic Projection

- Build Project Matrix from parallel corpus
  - Learn topic model on both side
  - collect topic co-occurrence via alignment
  - estimate project probability using MLE

source → topic model ← word alignment → topic model ← target
Topic Projection

- Build Project Matrix from parallel corpus
  - Learn topic model on both side
  - collect topic co-occurrence via alignment
  - estimate project probability using MLE
Topic Projection

Source side
Topic distribution
Topic Projection

Source side
Topic distribution
Topic Projection

Source side
Topic distribution

Topic Assignment
Alignment
Topic Projection

Source side
Topic distribution

Topic Assignment
Alignment

Topic-to-Topic
Projection Matrix
Topic Projection

Source side

Topic distribution

Topic Assignment

Alignment

Topic-to-Topic Projection Matrix

Projected Target

Distribution
One to Many Projection

- topic alignment are not essentially one-to-one.

<table>
<thead>
<tr>
<th>e-topic</th>
<th>f-topic 1</th>
<th>f-topic 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>enterprises</td>
<td>农业 (agricultural)</td>
<td>企业 (enterprise)</td>
</tr>
<tr>
<td>rural</td>
<td>农村 (rural)</td>
<td>市场 (market)</td>
</tr>
<tr>
<td>state</td>
<td>农民 (peasant)</td>
<td>国有 (state)</td>
</tr>
<tr>
<td>agricultural</td>
<td>改革 (reform)</td>
<td>公司 (company)</td>
</tr>
<tr>
<td>market</td>
<td>财政 (finance)</td>
<td>金融 (finance)</td>
</tr>
<tr>
<td>reform</td>
<td>社会 (social)</td>
<td>银行 (bank)</td>
</tr>
<tr>
<td>$P(z_f</td>
<td>z_e)$</td>
<td>0.38</td>
</tr>
</tbody>
</table>
Online adaption Framework

Source side

Target side
Online adaptation Framework

input → Topic Distribution

Source side

Target side
Online adaptation Framework

Source side

Target side
Online adaption Framework

Source side:
- Topic Distribution
- Projection Matrix
- Topic Distribution

Target side:
- input
- Topic Distribution
Online adaption Framework

Source side

Target side

input

Topic Distribution

Projection Matrix

Topic Distribution

Topic-based language model
Online adaption Framework

Source side

Target side

input

Topic Distribution

Projection Matrix

Topic Distribution

Topic-based language model

LM pro
Experiments

- Data-sets:
  - Training set (translation model): 1.5M sent.
  - Training set (language model): GIGA-Xinhua 10M sent.
  - dev-set: Nist06
  - test-set: Nist04, 05, 08
  - All with document boundary

<table>
<thead>
<tr>
<th>Data</th>
<th>Sentence</th>
<th>documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language model training</td>
<td>10M</td>
<td>980K</td>
</tr>
<tr>
<td>Translation model training</td>
<td>1.5M</td>
<td>99.4K</td>
</tr>
<tr>
<td>Tuning</td>
<td>616</td>
<td>52</td>
</tr>
<tr>
<td>Testing(04)</td>
<td>1788</td>
<td>200</td>
</tr>
<tr>
<td>Testing(05)</td>
<td>1082</td>
<td>100</td>
</tr>
<tr>
<td>Testing(08)</td>
<td>1357</td>
<td>109</td>
</tr>
</tbody>
</table>
Experimental setting

- Decoder: Hierarchical phrase-based system (Chiang+ '07)
  - add topic-based lm as additional feature to the system
- Language Model training tool: SRILM toolkit (Stolcke+ '02)
  - use wb-smoothing
  - 5-gram
- Topic Model tool: Open HTMM (Gruber+ '07)
  - parameters: $a=1.5$, $\beta=1.01$, iters=100
  - topic number=30
Internal Comparison

- Our method can be viewed as a soft clustering
  - a sent. belong to a topic cluster with a certain probability
- Compare with hard clustering
  - one sent. can only be clustered to one topic

<table>
<thead>
<tr>
<th></th>
<th>Hard</th>
<th>soft</th>
</tr>
</thead>
<tbody>
<tr>
<td>avg_bleu</td>
<td>33.1</td>
<td>33.325</td>
</tr>
<tr>
<td></td>
<td>33.325</td>
<td>33.55</td>
</tr>
<tr>
<td></td>
<td>33.55</td>
<td>33.775</td>
</tr>
<tr>
<td>avg_TER</td>
<td>42.2</td>
<td>43.2</td>
</tr>
<tr>
<td></td>
<td>42.45</td>
<td>42.7</td>
</tr>
<tr>
<td></td>
<td>42.95</td>
<td>43.2</td>
</tr>
</tbody>
</table>

Friday, November 15, 13
External Comparison

- Baseline: HPB model with N-gram language model
- Comparison method:
  - Tam ’07: use BLSA for language model adaption

![Average BLEU comparison](image1)

![Average TER comparison](image2)
Topic Number affection

- Topic number for HTMM is important to the performance
  - too small: insufficient clustering
  - too big: data sparse problem

- topic number 30 gets the best result
Conclusion

- a simple yet effective topic-triggered LM for SMT
  - add global topic context into N-gram model
  - online adapt to different topic distribution
  - directly integrate into SMT system
  - results in +0.76 significant improvement on various data-sets
- Future work
  - apply to more language pairs
  - combine with topic-based translation model
Thank you!