Improving SMT by Model Filtering and Phrase Embedding

Chengqing Zong

National Laboratory of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences

cqzong@nlpr.ia.ac.cn
Outline

1. Introduction
2. Diverse Features for SMT Model Pruning
3. Bilingually-constrained Phrase Embedding
4. Conclusion
1. Introduction

- The Chinese-English SLT system jointly developed by Alex’s team and my group was successfully demonstrated in Beijing Inter. Exhibition of Sci. & Tech., May 22-26, 2004, and in Barcelona Inter. Culture Forum, July 16-18, 2004.

- Intelligent terminals have become increasingly popular
1. Introduction

The problem of pruning the translation model for a statistical machine translation (SMT) system

- Phrase-based MT model (PBTM)
- Hierarchical phrase-based translation model (HPBSTM)

How to choose the most promising translation candidates from a large-scale translation table with several conditional probabilities?
1. Introduction

- How to distinguish the phrases with different semantic meanings?
  - Proposed a Bilingually-constrained Recursive Auto-encoders (BRAE) to learn semantic phrase embeddings (compact vector representations for phrases).
Outline

1. Introduction

2. Diverse Features for SMT Model Pruning

3. Bilingually-constrained Phrase Embedding

4. Conclusion
2. Diverse Features for SMT Model Pruning

- The translation pairs are extracted by a heuristic method based on the word alignments learned from parallel corpora.

Fig. 1 The translation pairs [Och and Ney, 2004]
2. Diverse Features for SMT Model Pruning

- Generally, two typical sorts of phrase-pairs cause the phrase table to be redundant and much larger than expected.

- One is that a distinct source phrase corresponds to many translation options (1-to-many).

| 大量 || a big amount of |
| 大量 || a big amount |
| 大量 || a big way |
| 大量 || a big |
| 大量 || a charter flight |
| 大量 || a chernobyl |
| 大量 || a large number |
| 大量 || a considerable amount of |
| 大量 || a considerable quantity of |
| 大量 || a deeper level |
| 大量 || a decrease in |
| 大量 || a drink |
| 大量 || a large sum of |
| 大量 || a survival |
2. Diverse Features for SMT Model Pruning

- The other case is that a distinct source phrase has only one or two translation options (1-to-few).
2. Diverse Features for SMT Model Pruning

- Some existing pruning methods
  - Use heuristic rules to cut off overloaded translation options

Those methods always do not work for the second case (1-to-few).

- Some methods adopt Fisher’s significance test to prune the weakly associated pairs
Fisher’s significance test

Given a training set that contains $N$ parallel sentences and a phrase-pair $(s, t)$ to be valued, we need to collect $C(s)$, $C(t)$ and $C(s, t)$, which represent the counts of sentences that contain $s$, $t$ and the pair $(s, t)$, respectively.

The probability that the pair $(s, t)$ occurs times by chance is given by the hyper-geometric distribution, as follows

$$P_h(c(s), c(t), c(s, t)) = \binom{c(s)}{c(s,t)} \times \binom{N-c(s)}{c(t)-c(s,t)} \times \binom{N}{c(t)}$$
The corresponding $p$-value is the sum of $P_v(C(s', t'))$ under the circumstance that $C(s', t')$ is not less than $C(s, t)$, while $C(s')$ equals $C(s)$ and $C(t')$ equals $C(t')$,

$$P_v(c(s), c(t), c(s, t)) = \sum_{c(s', t') \geq c(s, t)} \sum_{c(s') = c(s)} \sum_{c(t') = c(t)} P_h(c(s'), c(t'), c(s', t'))$$

If two phrase pairs occur the same times, and their source sides and target sides have the same times, their $p$-value is same.

If the $p$-value of a phrase pair is less than a threshold value, it will be ignored.
2. Diverse Features for SMT Model Pruning

• Some methods adopt Fisher’s significance test to prune the weakly associated pairs

However, they may discard many useful pairs unexpectedly, such as named entities that occur rarely in a parallel training corpus, which could damage the translation quality.

• Other methods using relative entropy are good at pruning redundant phrases, especially for those long phrases that can be replaced by combining shorter phrases

But the score of combined phrases often differs from the original one.
2. Diverse Features for SMT Model Pruning

- **Our motivations**
  - We consider the pruning task as a classification problem
  - Many effective heuristic measures mentioned above are encoded as strong representative features under the classification framework

We have explored rich statistical and syntactic features that are applied to the classification framework to compact the translation tables of PBTM and HPBPTM as much as possible.
2. Diverse Features for SMT Model Pruning

Fig. 2 Our Motivations
2. Diverse Features for SMT Model Pruning

● Some details of the methods

(1) Training set for the classifier

\[ P = \{ \text{Phrase pairs or rules truly used in the final translation} \} \]

\[ P \] is a sub-set of \[ C \].

\[ C = \{ \text{Phrase pairs or rules that have been involved into the translation lattice once or more} \} \]

\[ W = \{ \text{Dev-phrase or Dev-rule} \} \]

Fig. 3 Different Sets
2. Diverse Features for SMT Model Pruning

(2) Four types of features for pruning the phrase-based models

- Bi-directional phrase translation probabilities and bi-directional lexicalized translation probabilities
- Reordering probabilities
- Syntactic constraints on source side

Fig. 4
2. Diverse Features for SMT Model Pruning

The source syntactic constraint features are listed as follows:

- **SHS**: if the source phrase starts with a syntactic sub-phrase (more than one word), then $SHS=1$, otherwise, $SHS=0$;

![Diagram showing SHS feature](image)

**Fig. 5**
2. Diverse Features for SMT Model Pruning

- **STS**: if the source phrase ends with a syntactic sub-phrase (more than one word), then $STS=1$; otherwise, $STS=0$;
2. Diverse Features for SMT Model Pruning

✓ **SHA**: if the first word of the source phrase is aligned, then \( SHA = 1 \); otherwise, \( SHA = 0 \);

![Tree diagrams showing alignment and non-alignment](image)
2. Diverse Features for SMT Model Pruning

- $STA$: if the last word of the source phrase is aligned, then $STA=1$; otherwise, $STA=0$;

![Diagram](image)

- $STA = 1$

- $STA = 0$

Fig. 8
2. Diverse Features for SMT Model Pruning

✓ **SSW**: if the source phrase is a single word, then $SSW=1$; otherwise, $SSW=0$;

| Phrase Pair | 不 是 什 麼 ||| is not a |
|-------------|----------------|
| **SHS**     | 0              |
| **STS**     | 0              |
| **SHA**     | 1              |
| **STA**     | 0              |
| **SSW**     | 0              |

The syntactic constraints on target side are the same as those of the source syntactic constraints.
2. Diverse Features for SMT Model Pruning

- $p$-value feature

We add the length ratio as a feature. Let $L_s$ be the length (the word count) of the source phrase, and let $L_t$ be the length of the target phrase. Then

\[
\text{LenRatio} = \frac{\min(L_s, L_t)}{\max(L_s, L_t)}
\]
2. Diverse Features for SMT Model Pruning

(3) Features for pruning the hierarchical phrase-based models

In terms of the features for the rules with non-terminals, there are a total of three types:

- The rule translation probabilities and lexicalized translation probabilities
- LenRatio:
  \[ LenRatio = \frac{\min(L_{s\_terminal}, L_{t\_terminal})}{\max(L_{s\_terminal}, L_{t\_terminal})} \]
- Dependency syntactic features
  Relax-Well-Formed (RWF) structure (Z. Wang et al., 2010)
2. Diverse Features for SMT Model Pruning

Given a sentence $S = w_1w_2...w_n$, and let $d_1d_2...d_n$ represent the position of parent word for each word. If $w_i$ is a root, then $d_i=-1$.

Given a dependency structure $w_i ... w_j$, it will be called a Relax-Well-Formed structure if and only if it satisfies the following conditions:

- $d_h \not\in [i,j] \mathbb{Z}$, where $h \not\in [i,j]
- \forall k \in [i,j], d_k \in [i,j]$ or $d_k = h$, where $h \not\in [i,j]
2. Diverse Features for SMT Model Pruning

We say that $\text{RWF}(\text{rule}) = \text{true}$ if a rule is Relax-Well-Formed.

Given a rule $(r_l, r_r)$, we obtain the values in the triple vector $(D_l, D_r, D_b)$ for $r_l$ and $r_r$ as follows,

$$D_l = \frac{\sum_{i \in S_s} I(\text{RWF}(r_{li}) = \text{true})}{\sum_{i \in S_s} I(r_{li})}$$  
$D_l$ is the dependency feature value for the source side of a rule.

$$D_r = \frac{\sum_{j \in S_t} I(\text{RWF}(r_{rj}) = \text{true})}{\sum_{j \in S_t} I(r_{rj})}$$  
$D_r$ is for the target side.

$$D_b = \frac{\sum_{i \in S_s} \sum_{j \in S_t} I(\text{RWF}(r_{li}) = \text{true}, \text{RWF}(r_{rj}) = \text{true})}{\sum_{i \in S_s} \sum_{j \in S_t} I(r_{li}, r_{rj})}$$  
$D_b$ is for both sides.
2. Diverse Features for SMT Model Pruning

Summarization on our motivations

● Use the following diverse features
  • Bi-directional phrase translation probabilities and bi-directional lexicalized translation probabilities
  • Reordering probabilities
  • Syntactic constraints on source side
  • p-value feature

● Build an SVM-based classifier
  • Trained by the open toolkit LIBSVM
  • Chose Radial Basis Function (RBF) as our kernel function
2. Diverse Features for SMT Model Pruning

- Experiments
  - Experimental Setup
    - Training data: FBIS corpus, 7.1 million Chinese words and 9.2 million English words
    - Development set: NIST03, 919 sentences
      - NIST04, 1,788 sentences
    - Test set: NIST05 and NIST06 are only used as testing data in the translation task and never used for training the classifier
2. Diverse Features for SMT Model Pruning

- To obtain the original phrase-based and hierarchical phrase-based translation model, we train a 5-gram language model with SRILM on the FBIS English part.
- We obtain the source-to-target and target-to-source word alignments by GIZA++.
- These alignments are then symmetrized with grow-diag-final-and strategy.
- The translation model is generated by Moses (2010-8-13 Version), using the default parameter settings.
- For the phrase-based translation model, the maximum length of the phrases in the phrase table is 7.
- For pruning settings, the beam size is 200, and 20 translation options are retrieved for each input phrase.
2. Diverse Features for SMT Model Pruning

The features contained in the baseline translation system:

- bi-directional phrase translation probabilities
- bi-directional lexicalized translation probabilities
- bidirectional standard lexicalized reordering probabilities
- phrase penalty, word penalty
- distance-based reordering model score
- language model

For the hierarchal phrase-based translation model, we use the same setup as the phrase-based model, except we limit the number of symbols on each side of a rule to 5.
2. Diverse Features for SMT Model Pruning

For the training data for the SVM-based classifier,

- 10,000 positive training data instances
- 10,000 negative training data instances.

Dependency syntactic features

- Berkeley parser
- Trained on the Penn Chinese Treebank 6.0
2. Diverse Features for SMT Model Pruning

- Results on the Phrase-Based Model

Table 2: Performance of the SVM-based Classifier

<table>
<thead>
<tr>
<th>Combination of Features</th>
<th>Accuracy(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLTP+LenRatio+PV</td>
<td>94.90</td>
</tr>
<tr>
<td>SS+TS</td>
<td>84.92</td>
</tr>
<tr>
<td>PLTP+LRP+LenRatio+PV</td>
<td>95.81</td>
</tr>
<tr>
<td>PLTP+LRP+LenRatio+TS+PV</td>
<td>94.98</td>
</tr>
<tr>
<td>PLTP+LRP+LenRatio+SS+TS+PV</td>
<td>93.26</td>
</tr>
<tr>
<td>PLTP+LRP+Lt/Ls+Ls/Lt+PV</td>
<td>95.06</td>
</tr>
<tr>
<td>PLTP+LRP+Lt/Ls+Ls/Lt+TS+PV</td>
<td>94.38</td>
</tr>
</tbody>
</table>

- PLTP: phrase and lexicalized phrase translation probabilities, bi-directional
- LRP: The standard lexicalized reordering probabilities; PV: $p$-value
- SS: Source syntactic features; TS: Target syntactic features
- Length of the target phrase (Lt); Ls: Length of the source phrase
### 2. Diverse Features for SMT Model Pruning

Table 3: Size and BLEU for different combined features compared to the baseline (POS: NEG = 1:1)

<table>
<thead>
<tr>
<th>Feature Combination</th>
<th>Table Size</th>
<th>NIST’05 BLEU</th>
<th>NIST’06 BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>15,428,040</td>
<td>25.30</td>
<td>26.86</td>
</tr>
<tr>
<td>PLTP+LenRatio+PV</td>
<td>9,355,935 (61%)</td>
<td>25.12</td>
<td>26.98</td>
</tr>
<tr>
<td>SS+TS</td>
<td>8,514,694 (55%)</td>
<td>24.65</td>
<td>25.04</td>
</tr>
<tr>
<td>PLTP+LRP+LenRatio+PV</td>
<td>6,336,658 (41%)</td>
<td>25.33</td>
<td>26.87</td>
</tr>
<tr>
<td>PLTP+LRP+LenRatio+TS+PV</td>
<td><strong>3,756,190 (24%)</strong></td>
<td>25.26</td>
<td>26.73</td>
</tr>
<tr>
<td>PLTP+LRP+LenRatio+SS+TS+PV</td>
<td>6,584,166 (43%)</td>
<td>25.25</td>
<td>26.87</td>
</tr>
<tr>
<td>PLTP+LRP+Ls/Lt+Lt/Ls+PV</td>
<td>5,569,328 (36%)</td>
<td>25.34</td>
<td>26.76</td>
</tr>
<tr>
<td>PLTP+LRP+Ls/Lt+Lt/Ls+PV+TS</td>
<td>4,962,691 (32%)</td>
<td><strong>25.40</strong></td>
<td>26.63</td>
</tr>
</tbody>
</table>
2. Diverse Features for SMT Model Pruning

Fig. 10  Changes in the table size of NIST05 and NIST06
2. Diverse Features for SMT Model Pruning

Table 4: Basic Statistics of the Original and Pruned Phrase under the Features PLTP + LRP + LenRatio + TS + PV

<table>
<thead>
<tr>
<th></th>
<th>Whole Table</th>
<th>NIST05</th>
<th>NIST06</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>original</td>
<td>pruned</td>
<td>original</td>
</tr>
<tr>
<td>DSN</td>
<td>5,691,593</td>
<td>2,924,778</td>
<td>21,319</td>
</tr>
<tr>
<td>ACPS</td>
<td>2.71</td>
<td>1.28</td>
<td>101.5</td>
</tr>
<tr>
<td>ALS</td>
<td>4.17</td>
<td>4.04</td>
<td>2.157</td>
</tr>
<tr>
<td>ALT</td>
<td>4.02</td>
<td>3.68</td>
<td>3.41</td>
</tr>
</tbody>
</table>

DSN: distinct source phrase numbers;
ACPS: average candidate-options per distinct source phrase;
ALS: average length of distinct source phrases;
ALT: average length of target phrases
2. Diverse Features for SMT Model Pruning

- Testing on a Different Scale of Negative Data

Table 5. BLEU score on the test set with an increasing rate of negative data

<table>
<thead>
<tr>
<th>POS:NEG</th>
<th>NIST’05/’06</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GROUP1</td>
</tr>
<tr>
<td>Baseline</td>
<td>25.30/26.86</td>
</tr>
<tr>
<td>1:1.00</td>
<td>25.33/26.87</td>
</tr>
<tr>
<td>1:2.00</td>
<td>25.40/26.75</td>
</tr>
<tr>
<td>1:3.00</td>
<td>25.44/26.63</td>
</tr>
<tr>
<td>1:4.00</td>
<td>25.48/26.59</td>
</tr>
<tr>
<td>1:5.00</td>
<td>25.37/26.56</td>
</tr>
</tbody>
</table>

- Group 1 of features: PLTP, LRP, LenRatio and PV
- Group 2 of features: Group 1 + TS
2. Diverse Features for SMT Model Pruning

Fig. 11 The reduction in the whole table size
2. Diverse Features for SMT Model Pruning

Fig. 12 The translation quality of the testing set with different ratios of negative and positive data
2. Diverse Features for SMT Model Pruning

• Results on the Hierarchical Phrase-Based Models

Table 6. Size and BLEU of different combined features in HPBTM

<table>
<thead>
<tr>
<th>Feature</th>
<th>Table Size</th>
<th>BLEU NIST05</th>
<th>BLEU NIST06</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>69,393,064 (100%)</td>
<td>25.58</td>
<td>28.15</td>
</tr>
<tr>
<td>RWF</td>
<td>45,338,911 (65%)</td>
<td><strong>25.91</strong></td>
<td>28.11</td>
</tr>
<tr>
<td>RLTP+PV+LenRatio</td>
<td>28,614,096 (41%)</td>
<td>25.73</td>
<td>28.00</td>
</tr>
<tr>
<td>RLTP+PV+LenRatio+DSS+DTS+DBOTH</td>
<td>24,954,654 (36%)</td>
<td>25.77</td>
<td>28.05</td>
</tr>
<tr>
<td>RLTP+PV+LenRatio+DTS+DBOTH</td>
<td>23,126,972 (33%)</td>
<td>25.69</td>
<td><strong>28.15</strong></td>
</tr>
</tbody>
</table>

• RWF: Relax-Well-Formed
• DSS: Dependency syntactic feature of source side
• DTS: Dependency syntactic feature of target side;
• DBOTH: Dependency syntactic feature of both sides
2. Diverse Features for SMT Model Pruning

Fig. 13 Details of Phrase-table and Rule-table
2. Diverse Features for SMT Model Pruning

Table 6. Accuracy of Classifier with Different feature Combinations

<table>
<thead>
<tr>
<th>Feature Combination</th>
<th>Accuracy</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RLTP+PV+LenRatio</td>
<td>98.07%</td>
<td>95.99%</td>
</tr>
<tr>
<td>RLTP+PV+LenRatio+DSS+DTS+DBOTH</td>
<td>96.97%</td>
<td>95.24%</td>
</tr>
<tr>
<td>RLTP+PV+LenRatio+DTS+DBOTH</td>
<td>97.21%</td>
<td>95.42%</td>
</tr>
</tbody>
</table>

P-classifier is for the phrase-table; R-classifier is for the rule-table.
2. Diverse Features for SMT Model Pruning

Fig. 14 The reduction in the whole table size with different scales of negative data in the hierarchical model
2. Diverse Features for SMT Model Pruning

- Summarization on SMT model pruning
  - The classifier-based approach is effective at pruning approximately 80% of the phrase-pairs and 70% of the rules without harming the translation quality.
  - For the Chinese-to-English translation task, it is better not to use the source syntactic information as a single feature.
  - Training data of the classifier are generated automatically from the decoding path with a tuned translation model and development data.
  - Our unified framework breaks the limitation of specific translation model pruning, which makes it possible to extend and transfer to other syntactic translation models.
2. Diverse Features for SMT Model Pruning

For more details about this work, please refer to:

Outline

1. Introduction
2. Diverse Features for SMT Model Pruning
3. Bilingually-constrained Phrase Embedding
4. Conclusion
3. Bilingually-constrained Phrase Embedding

Word Embedding:
Maps a word into a real-valued vector

meeting:   [0.25, 0.12, 0.36]
conference:  [0.24, 0.10, 0.35]
3. Bilingually-constrained Phrase Embedding

• View any phrase as a single unit, using the word embedding method
  – Context of a phrase is always limited
  – Cannot enumerate all the phrases

• View any phrase as bag-of-words
  – Sum of word embeddings
  – Ignore the word order information

• View any phrase as meaningful combination of words
  – Learn the way of combination
  – Apply recursive auto-encoders
3. Bilingually-constrained Phrase Embedding

- **Unsupervised Phrase Embedding with Recursive Auto-encoder (RAE)**

- No other supervision except for minimizing the reconstruction error

\[
E_{total} = \sum_{node} E_{Rec\_node}
\]

Unsupervised RAE can only learn some **syntactic information** of the phrase, but **cannot learn the correct semantics**
3. Bilingually-constrained Phrase Embedding

- Semi-supervised Phrase Embedding

$$E([x_1, x_2]) = \alpha E_{REC}([x_1, x_2]) + (1 - \alpha) E_{PE}(y)$$

Parsing: label is the **syntactic category**
Sentiment Analysis: label is the **polarity**
Phrase Reordering: label is the **swap or monotone**

Semi-supervised RAE can **group** the phrases with the **same syntactic category, the same polarity and the same reordering pattern**
3. Bilingually-constrained Phrase Embedding

- Problems of the Unsupervised and Semi-supervised Phrase Embedding
  - Unsupervised phrase embedding
    - only learn some syntactic information of the phrase, but cannot learn the correct semantics
    - More like syntactic embedding
  - Semi-supervised phrase embedding
    - Group the phrases with the similar role (polarity or reordering pattern)
    - More like role embedding
  - How to learn semantic phrase embedding?
3. Bilingually-constrained Phrase Embedding

- Extension to Semi-supervised Phrase Embedding

If some gold semantic phrase representation are given as labels

We can learn how to embed each phrase semantically!
Problem
  – We have no gold representation for any phrase

What can we make full use of?
  – If phrases are represented semantically, two phrases sharing
    the same meaning should have the same phrase embedding

What can we learn from the above fact?
  – If a model can always learn the same representation for any
    phrase pairs which have the same meaning, the
    representation must encode the semantics of the phrase, and
    the model is our desire
3. Bilingually-constrained Phrase Embedding

• Bilingually-constrained Phrase Embedding
  
  • What we need to learn the model?
    – The phrase pairs sharing the same meaning
  
  • Assumption
    – The phrase and its correct translation are in the same meaning, and the semantic representation between them should be the same (translation equivalents)
3. Bilingually-constrained Phrase Embedding

If source-side semantic phrase embedding is given, we can learn how to train the target-side embedding.

If target-side semantic phrase embedding is given, we can learn how to train the source-side embedding.

- Basic Objective Function
  - Minimizing the semantic distance between the representations of the translation equivalents.

France and Russia

France and Russia
3. Bilingually-constrained Phrase Embedding

- Basic Objective Function

\[ E(s, t; \theta) = \alpha E_{REC}(s, t; \theta) + (1 - \alpha) E_{SEM}(s, t; \theta) \]

- Reconstruction error
- Semantic distance (error)
3. Bilingually-constrained Phrase Embedding

\[ E_{REC}(s, t; \theta) = E_{REC}(s; \theta) + E_{REC}(t; \theta) \]

Source reconstruction error

Target reconstruction error
3. Bilingually-constrained Phrase Embedding

\[ E_{SEM}(s, t; \theta) = E_{SEM}(s|t, \theta) + E_{SEM}(t|s, \theta) \]

The distance in the target embedding space

The distance in the source embedding space
3. Bilingually-constrained Phrase Embedding

\[ E_{SEM}(s|t, \theta) = \frac{1}{2} \| p_t - f(W_s l p_s + b_s^l) \|^2 \]

\[ E_{SEM}(t|s, \theta) = \frac{1}{2} \| p_s - f(W_t l p_t + b_t^l) \|^2 \]
3. Bilingually-constrained Phrase Embedding

\[ E_{SEM}(s|t, \theta) = \frac{1}{2} \left\| p_t - f(W_s^l p_s + b_s^l) \right\|^2 \]  

Translation Equivalents

Ideally, we require the distance between translation equivalents is much smaller than that between non-translation pairs.
3. Bilingually-constrained Phrase Embedding

\[ E_{SEM}(s|t, \theta) = \frac{1}{2} \left\| p_t - f(W_s^l p_s + b_s^l) \right\|^2 \]

Translation Equivalents

\[ E_{SEM}^*(s|t, \theta) = \max\{0, E_{SEM}(s|t, \theta) - E_{SEM}(s|t', \theta) + 1\} \]

Max-Semantic-Margin non-translation pair
3. Bilingually-constrained Phrase Embedding

- **Parameter Training**

\[ E(s, t; \theta) = \alpha E_{REC}(s, t; \theta) + (1 - \alpha) E_{SEM}(s, t; \theta) \]
\[ = \alpha (E_{REC}(s; \theta) + E_{REC}(t; \theta)) \]
\[ + (1 - \alpha) (E_{SEM}^{*}(s|t, \theta) + E_{SEM}^{*}(t|s, \theta)) \]

source-side parameters can be tuned as long as the
representation of the target phrase is given

target-side parameters can be tuned as long as the
representation of the source phrase is given
3. Bilingually-constrained Phrase Embedding

- Co-training style Training Algorithm

**Step 1, Pre-training:** learn respectively the source and target phrase embedding with standard unsupervised RAEs
Step 2, Fine-tuning: a) regard target phrase embedding as the gold semantic representation of source phrase, and then refine the parameters for source phrase embedding process;
3. Bilingually-constrained Phrase Embedding

Step 2, Fine-tuning: b) regard source phrase embedding as the gold semantic representation of target phrase, and then refine the parameters for target phrase embedding process.
3. Bilingually-constrained Phrase Embedding

Step 3, Termination Check: If overall error reaches a minima or the iterations reach the pre-defined number, we terminate, otherwise, we update the source and target phrase representations and go to Step 2
3. Bilingually-constrained Phrase Embedding

- Some good examples

<table>
<thead>
<tr>
<th>military force</th>
<th>military power</th>
<th>armed forces</th>
<th>do not agree</th>
<th>do not favor</th>
<th>not to approve</th>
<th>each people in this nation</th>
<th>all the people in the country</th>
<th>people all over the country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similar ones:</td>
<td>Similar ones:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3. Bilingually-constrained Phrase Embedding

- **Experiments**
  
  - **Usage of Phrase Embedding in SMT**
    - For each phrase pair in phrase table, after obtaining source and target phrase embedding $p_s$ and $p_t$, map $p_s$ into target space $p'_s$ and semantic similarity $Sem(p'_s, p_t)$; similarly, we can get $Sem(p'_t, p_s)$,
    - As another two translation features, integrate in phrase-based decoding
    - Use the $Sem(p'_s, p_t)$ and $Sem(p'_t, p_s)$ to prune phrase table
3. Bilingually-constrained Phrase Embedding

- **Experimental Setup**
  - Decoder: MEBTG
  - Training: 0.96M bitext + 1.1M entity pairs, English Gigaword Xinhua News
  - Tuning set: NIST03
  - Test set: NIST04, NIST05, NIST06, NIST08 (news part in NIST06-08)
## 3. Bilingually-constrained Phrase Embedding

- As extra two features

<table>
<thead>
<tr>
<th>System</th>
<th>NIST03</th>
<th>NIST04</th>
<th>NIST05</th>
<th>NIST06</th>
<th>NIST08</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEBTG</td>
<td>35.81</td>
<td>36.91</td>
<td>34.69</td>
<td>33.83</td>
<td>27.17</td>
<td>34.82</td>
</tr>
<tr>
<td>+2feats 50-dim</td>
<td>36.43</td>
<td>37.64</td>
<td>35.35</td>
<td>35.53</td>
<td>28.59</td>
<td>35.84</td>
</tr>
<tr>
<td></td>
<td>(0.62↑)</td>
<td>(0.73↑)</td>
<td>(0.66↑)</td>
<td>(1.70↑)</td>
<td>(1.42↑)</td>
<td>(1.02↑)</td>
</tr>
<tr>
<td>+2feats 100-dim</td>
<td>36.45</td>
<td>37.44</td>
<td>35.58</td>
<td>35.42</td>
<td>28.57</td>
<td>36.03</td>
</tr>
<tr>
<td></td>
<td>(0.64↑)</td>
<td>(0.53↑)</td>
<td>(0.89↑)</td>
<td>(1.59↑)</td>
<td>(1.40↑)</td>
<td>(1.21↑)</td>
</tr>
<tr>
<td>+2feats 200-dim</td>
<td>36.34</td>
<td>37.35</td>
<td>35.78</td>
<td>34.87</td>
<td>27.84</td>
<td>35.62</td>
</tr>
<tr>
<td></td>
<td>(0.53↑)</td>
<td>(0.44↑)</td>
<td>(1.09↑)</td>
<td>(1.04↑)</td>
<td>(0.67↑)</td>
<td>(0.80↑)</td>
</tr>
</tbody>
</table>

\[\]
3. Bilingually-constrained Phrase Embedding

- Phrase table pruning

<table>
<thead>
<tr>
<th>System</th>
<th>Phrase-Table</th>
<th>NIST03</th>
<th>NIST04</th>
<th>NIST05</th>
<th>NIST06</th>
<th>NIST08</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEBTG</td>
<td>100%</td>
<td>35.81</td>
<td>36.91</td>
<td>34.69</td>
<td>33.83</td>
<td>27.17</td>
<td>34.82</td>
</tr>
<tr>
<td>0.4</td>
<td>52%</td>
<td>35.94</td>
<td>36.96</td>
<td>35.00</td>
<td>34.71</td>
<td>27.77</td>
<td><strong>35.16</strong></td>
</tr>
<tr>
<td>0.5</td>
<td>44%</td>
<td>35.67</td>
<td>36.59</td>
<td>34.86</td>
<td>33.93</td>
<td>27.25</td>
<td><strong>34.89</strong></td>
</tr>
<tr>
<td>0.6</td>
<td>35%</td>
<td>35.86</td>
<td>36.71</td>
<td>34.93</td>
<td>34.63</td>
<td>27.34</td>
<td><strong>35.05</strong></td>
</tr>
<tr>
<td>0.7</td>
<td><strong>28%</strong></td>
<td>35.55</td>
<td>36.62</td>
<td>34.57</td>
<td>33.97</td>
<td>27.10</td>
<td><strong>34.76</strong></td>
</tr>
<tr>
<td>0.8</td>
<td>20%</td>
<td>35.06</td>
<td>36.01</td>
<td>34.13</td>
<td>33.04</td>
<td>26.66</td>
<td>34.04</td>
</tr>
</tbody>
</table>
3. Bilingually-constrained Phrase Embedding

● Summary on phrase embedding

  • Bilingually-constrained Phrase Embedding
    – Can learn how to semantically embed both source and target phrases
    – Can learn how to transform the semantic embedding space in one language to the other

  • The model is proven to be effective in phrase table pruning and decoding with phrasal semantic similarity
3. Bilingually-constrained Phrase Embedding

For more details about this work, please refer to:

J. Zhang, S. Liu, M. Li, M. Zhou and C. Zong.
1. Introduction
2. Diverse Features for SMT Model Pruning
3. Bilingually-constrained Phrase Embedding
4. Conclusion
Conclusion

• Speech and translation quality
  • Pruning phrase table or translation rule for PBTM and HPBTM
    - Whether the source syntactic information has the same effect in English-to-Chinese translation?
    - What type of phrase/rule should be pruned?
  • Development of an approach to phrase embedding

• How to combine the two work?
References


References


