Translation of Patent Sentences with a Large Vocabulary of Technical Terms Using Neural Machine Translation

Zi Long, Takehito Utsuro
University of Tsukuba

Tomoharu Mitsuhashi
Japan Patent Information Organization

Mikio Yamamoto
University of Tsukuba

WAT2016, December 12, 2016 @ Osaka, Japan
Neural Machine Translation (encoder-decoder Model and attention mechanism)


A large vector that represents the entire input sentence
Neural Machine Translation (encoder-decoder Model and attention mechanism)


(input)

(Encoder)

(Decoder)

(output)
Neural Machine Translation VS Statistic Machine Translation

<table>
<thead>
<tr>
<th></th>
<th>Neural Machine Translation (NMT)</th>
<th>Statistic Machine Translation (SMT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>fluency</td>
<td>high</td>
<td>low</td>
</tr>
<tr>
<td>vocabulary</td>
<td>small</td>
<td>large</td>
</tr>
</tbody>
</table>
Neural Machine Translation VS Statistic Machine Translation

Neural Machine Translation (NMT)
- high fluency
- small vocabulary
- inappropriate for translating technical terms

Statistic Machine Translation (SMT)
- phrase-level
- large vocabulary
- store explicit phrase translation table

NMT with a Large Vocabulary of Technical Terms

- Step 1. training NMT model with technical term tokens

- Step 2. applying NMT model with technical term tokens
  - Approach 1. NMT decoding and SMT technical term translation
  - Approach 2. NMT rescoring of 1,000-best SMT translations (not as fluent as Approach 1)
NMT with a Large Vocabulary of Technical Terms

- Step 1. training NMT model with technical term tokens
- Step 2. applying NMT model with technical term tokens
  - Approach 1. NMT decoding and SMT technical term translation
  - Approach 2. NMT rescoring of 1,000-best SMT translations (not as fluent as Approach 1)
NMT training after replacing Technical Terms with Tokens

1. aligning technical term pairs using SMT phrase translation table and word alignment

Japanese sentence: cmac/ユニット/312/は/信号/を/ブリッジ/インタフェース/388/に/提供/する/。

Chinese sentence: cmac/单元/312/将/信号/提供/给/桥架/接口/388/。

(cmac unit 312 provides a signal to the bridge interface 388.)
NMT training after replacing Technical Terms with Tokens

2. replacing each aligned technical term pair with an identical technical term token “TT\textsubscript{i}” (i = 1, 2, …)
NMT training after replacing Technical Terms with Tokens

Japanese sentence: cmac/ユニット/312/は/信号/を/ブリッジ/インタフェース/388/に/提供/する/。

Chinese sentence: cmac/单元/312/将/信号/提供/给/桥架/接口/388/。

(cmac unit 312 provides a signal to the bridge interface 388.)

Japanese sentence with technical term tokens "TT₁", "TT₂":
TT₁ /312/は/信号/を/TT₂ /388/に/提供/する/。

Chinese sentence with technical term tokens "TT₁", "TT₂":
TT₁ /312/将/信号/提供/给/TT₂ /388/。

(TT₁ 312 provides a signal to the TT₂ 388.)
NMT with a Large Vocabulary of Technical Terms

- **Step 1.** training NMT model with technical term tokens

- **Step 2.** applying NMT model with technical term tokens
  - Approach 1. NMT decoding and SMT technical term translation
  - Approach 2. NMT rescoring of 1,000-best SMT translations (not as fluent as Approach 1)
NMT Decoding and SMT technical Term Translation

input Japanese sentence

extracted Japanese technical terms

phrase translation table

NMT translation model (with technical term tokens)

output Chinese translation (with technical term tokens)

output Chinese translation

replacing technical terms with technical term tokens

decoding by NMT translation model (with technical term tokens)

replacing technical term tokens with technical term translation by SMT
Training and Test Sets

- 2.8M parallel sentences extracted from Japanese-Chinese patent families
- Randomly selected 1,000 sentence pairs as the test set

pair 1:

J: ・・冷蔵庫および冷蔵庫扉閉鎖装置・・
C: ・・电冰箱及电冰箱门锁闭装置・・。

(closing appliance of Refrigerator's door)

pair 2:

J: ・・真空断熱材及びその製造方法・・.
C: ・・真空绝热材料及其制作方法・・。

(vacuum thermal insulation)

pair n:

J: ・・運動に関する3次元デカルト座標系を定める。
C: ・・以限定用于描述运动的三维笛卡尔坐标系・・。

(Cartesian coordinate system)
Experiments Settings

- **Baseline SMT (PBMT)**
  - phrase-based SMT model trained with the same training set using Moses.

- **Baseline NMT**
  - uni-direction model with attention mechanism
  - 3 layer deep LSTMs with 512 cells in each layer and a 512-dimensional word embedding.
  - limit both the Japanese vocabulary and the Chinese vocabulary to 40,000 most frequently used word
  - more training details.
Experiments Settings

- NMT with PosUnk model [Luong+ 2015]
  - same training parameters with the baseline NMT
  - training NMT model with PosUnk model

- NMT with technical term tokens
  - same training parameters with the baseline NMT
  - training NMT model after replacing technical terms with tokens.
Evaluation Results
(automatic evaluation - BLEU)

- BLEU score

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline SMT (PBMT)</td>
<td>52.5</td>
</tr>
<tr>
<td>Baseline NMT</td>
<td>53.5</td>
</tr>
<tr>
<td>NMT with PosUnk model [Luong+ 2015]</td>
<td>54.0</td>
</tr>
<tr>
<td>NMT with technical term tokens</td>
<td>55.3</td>
</tr>
</tbody>
</table>
Evaluation Results
(automatic evaluation - RIBES)

- RIBES score

<table>
<thead>
<tr>
<th>Model</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline SMT (PBMT)</td>
<td>88.5</td>
</tr>
<tr>
<td>Baseline NMT</td>
<td>90.0</td>
</tr>
<tr>
<td>NMT with PosUnk model [Luong+ 2015]</td>
<td>90.4</td>
</tr>
<tr>
<td>NMT with technical term tokens</td>
<td>90.8</td>
</tr>
</tbody>
</table>
Evaluation Results
(human evaluation – pairwise evaluation)

- Pairwise Evaluation (scores range from -100 to 100)

<table>
<thead>
<tr>
<th>Baseline NMT</th>
<th>NMT with technical term tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.0</td>
<td>36.5</td>
</tr>
</tbody>
</table>
Evaluation Results
(human evaluation – JPO adequacy evaluation)

- JPO adequacy evaluation (scores range from 1 to 5)

- Baseline SMT (PBMT): 3.5
- Baseline NMT: 3.8
- NMT with technical term tokens: 4.3
Example of correct translations produced by NMT decoding

Japanese sentence

次に、酸化膜をhf洗浄により除去した後、貼り合わせウェーハの剥離面から酸素イオンを注入した。

(Next, after removing an oxide film by hf washing, we inject oxygen ions from the peeled surface of the laminated wafer.)

NMT decoding by the PROPOSED NMT

接着，通过hf洗涤除去氧化膜后，从贴合晶片的剥离面注入氧离子。

(TRANSLATION: Subsequently, after the oxide film was removed by washing with hf, and oxygen ions were injected from the peeled surface of the laminated wafer.)

NMT decoding by the BASELINE NMT

接着，通过hf清洗除去氧化膜后，从貼合的UNK的剥离面注入氧气。

(TRANSLATION: Then, after the oxide film was removed by hf cleaning, oxygen was injected from the peeled surface of the laminated UNK.)

The Chinese word “晶片”(wafer) is out of the vocabulary

Incorrect!
Next, after removing an oxide film by hf washing, we inject oxygen ions from the peeled surface of the **laminated wafer**.

**Correct!**

(TRANSLATION: Subsequently, after the oxide film was removed by washing with hf, and oxygen ions were injected from the peeled surface of the **laminated wafer**.)

**Incorrect!**

(TRANSLATION: Then, after the oxide film was removed by hf cleaning, oxygen was injected from the peeled surface of the **bonded UNK**.)

Japanese technical term “**貼り合わせウェーハ**” (laminated wafer) is translated as Chinese technical term “**贴合晶片**” by SMT
NMT with a Large Vocabulary of Technical Terms

- Step 1. training NMT model with technical term tokens

- Step 2. applying NMT model with technical term tokens
  - Approach 1. NMT decoding and SMT technical term translation
  - Approach 2. NMT rescoring of 1,000-best SMT translations (not as fluent as Approach 1)
Evaluation Results (2) (automatic evaluation)

- **BLEU score**
  - NMT decoding and SMT technical term translation: 55.3
  - NMT rescoring of SMT 1,000-best translation: 55.6
  - Slightly higher

- **RIBES score**
  - NMT decoding and SMT technical term translation: 90.8
  - NMT rescoring of SMT 1,000-best translation: 89.3
  - much higher
Evaluation Results (2)  
(human evaluation)

- **Pairwise Evaluation**
  (scores range from -100 to 100)

  - NMT decoding and SMT technical term translation: 36.5
  - NMT rescoring of SMT 1,000-best translation: 31.0

- **JPO adequacy evaluation**
  (scores range from 1 to 5)

  - NMT decoding and SMT technical term translation: 4.3
  - NMT rescoring of SMT 1,000-best translation: 4.1
Conclusion and Future Work

- Translating patent sentences with a large vocabulary of technical terms by training an NMT system on a bilingual corpus, wherein technical terms are replaced with tokens.

- Evaluation experiments on Japanese-Chinese patent sentences proved the effectiveness of the proposed method.

Future Work:
- Evaluate the proposed method with a bidirectional NMT system (Bahdanau et al. 2015).
- Rescore 1,000-best NMT translations by using SMT system.
Thank you for your attention!