Statistical Machine Translation Based on Hierarchical Phrase Alignment

Taro Watanabe †, Kenji Imamura and Eiichiro Sumita

taro.watanabe@atr.co.jp

ATR Spoken Language Translation Research Laboratories
Introduction to SMT  
(refer to TMI Tutorial)

\[ e = \arg \max_e P(e|f) \]

\[ e \leftarrow \text{Translator} \leftarrow f \]
Introduction to SMT
(refer to TMI Tutorial)

Apply the Bayes Rule:

\[ e = \arg \max_e P(e|f) \]

Source Model \[ \rightarrow e \rightarrow \text{Channel Model} \rightarrow f \]

\[ e = \arg \max_e P(e|f) = \arg \max_e P(e) \times P(f|e) \]
Introduction to SMT
(refer to TMI Tutorial)

Apply the Bayes Rule:

\[ e = \arg \max_e P(e|f) \]

\[ e = \arg \max_e P(e|f) \]
\[ = \arg \max_e P(e) \times P(f|e) \]

- \( P(e) \) — Language Model
- \( P(f|e) \) — Translation Model
Translation Model

How to represent $P(f|e)$? (a correspondence between e and f)
Translation Model

- How to represent $P(f|e)$? (a correspondence between $e$ and $f$)
- Introduction of $a$: alignment

$$P(f|e) = \sum_a P(f, a|e)$$
Translation Model

- How to represent $P(f|e)$? (a correspondence between $e$ and $f$)
- Introduction of $a$: alignment

$$P(f|e) = \sum_a P(f, a|e)$$

- An example of alignments

```
NULL  The_1  poor_2  don’t_3  have_4  any_5  money_6

Les_1  pauvres_2  sont_3  demunis_4
```
Translation Model

- How to represent $P(f|e)$? (a correspondence between $e$ and $f$)
- Introduction of $a$: alignment

$$P(f|e) = \sum_a P(f, a|e)$$

- An example of alignments

```
NULL  The_1  poor_2  don’t_3  have_4  any_5  money_6
|      |      |      |      |      |      |
Les_1 pauvres_2 sont_3 demunis_4
```

$a = (1, 2, 4, 5)$
Translation Model

- How to represent \( P(f|e) \)? (a correspondence between \( e \) and \( f \))
- Introduction of \( a \) : alignment

\[
P(f|e) = \sum_a P(f, a|e)
\]

- An example of alignments

\[
\begin{align*}
\text{NULL} & \quad \text{The}_1 \quad \text{poor}_2 \quad \text{don't}_3 \quad \text{have}_4 \quad \text{any}_5 \quad \text{money}_6 \\
\text{Les}_1 & \quad \text{pauvres}_2 \quad \text{sont}_3 \quad \text{dimunis}_4 \\
\end{align*}
\]

\[a = (0, 2, 3, 6)\]
Structure of TM (IBM Model 4)

**Translation Model**

**Lexical Model**
\[ \prod t(f_j|e_i) \]

**Fertility Model**
\[ \prod n(\phi_i|e_i) \]

**Distortion Model**

- **Head**
  \[ \prod d_1(j - k|\mathcal{A}(e)\mathcal{B}(f)) \]

- **Non-Head**
  \[ \prod d_{1>} (j - j'|\mathcal{B}(f)) \]

**NULL Translation Model**
\[ (m-\phi_0) p_0^{m-2\phi_0} p_1^{\phi_0} \]
Structure of TM (IBM Model 4)

<table>
<thead>
<tr>
<th>Lexical Model</th>
<th>Translation Model</th>
<th>Fertility Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\prod t(f_j</td>
<td>e_i)$</td>
<td>$\prod n(\phi_i</td>
</tr>
</tbody>
</table>

**Distortion Model**

- Head: $\prod d_1(j-k|\mathcal{A}(e)\mathcal{B}(f))$
- Non-Head: $\prod d_{1>}(j-j'|\mathcal{B}(f))$

**NULL Translation Model**

$\left(m-\phi_0\right)p_0^{m-2\phi_0}p_1^{\phi_0}$

**Lexical Model**

| e ... t(f|e) e ... |
|-------------------|
| f ... f ...      |

Statistical Machine Translation Based on Hierarchical Phrase Alignment – p. 4/19
Structure of TM (IBM Model 4)

Translation Model

Lexical Model
\[ \prod t(f_j | e_i) \]

Fertility Model
\[ \prod n(\phi_i | e_i) \]

Distortion Model

Head
\[ \prod d_1(j - k | \mathcal{A}(e) \mathcal{B}(f)) \]

Non-Head
\[ \prod d_{1 >}(j - j' | \mathcal{B}(f)) \]

NULL Translation Model
\[ (m - \phi_0) p_0^{m - 2 \phi_0} p_1^{\phi_0} \]

Fertility Model

\[ e \quad \cdots \quad e \]

\[ n(\phi | e) \]

\[ f \quad \cdots \quad f_1 (\text{Head}) \quad \cdots \quad f_2 \quad f_3 \]
Structure of TM (IBM Model 4)

**Lexical Model**
\[ \prod t(f_j|e_i) \]

**Fertility Model**
\[ \prod n(\phi_i|e_i) \]

**Distortion Model**
- **Head**
  \[ \prod d_1(j-k|\mathcal{A}(e)\mathcal{B}(f)) \]
- **Non-Head**
  \[ \prod d_{1>}(j-j'|\mathcal{B}(f)) \]

**NULL Translation Model**
\[ (m-\phi_0)p_0^{m-2\phi_0}p_1^{\phi_0} \]

**Distortion Model (Head)**
- **e** \[ \ldots \]
- **f** \[ \ldots \]
  \[ f_k \]
- **A(e)** \[ \ldots \]
- **d_1(j-k|A(e)B(f_j))**
- **B(f_j)** \[ \ldots \]
Structure of TM (IBM Model 4)

Translation Model

Lexical Model
\[ \prod t(f_j|e_i) \]

Fertility Model
\[ \prod n(\phi_i|e_i) \]

Distortion Model

Head
\[ \prod d_1(j - k|\mathcal{A}(e)\mathcal{B}(f)) \]

Non-Head
\[ \prod d_{1>} (j - j'|\mathcal{B}(f)) \]

NULL Translation Model
\[ (m-\phi_0)p_0^{m-2\phi_0}p_1^{\phi_0} \]

Distortion Model (Non-Head)

\[ e \rightarrow \ldots \rightarrow e \]

\[ d_{1>} (j - j'|\mathcal{B}(f_j)) \]

\[ f \rightarrow \ldots \rightarrow f_{j'} \rightarrow \ldots \rightarrow \mathcal{B}(f_j) \rightarrow \ldots \]
Structure of TM (IBM Model 4)

Translation Model

Lexical Model
\[ \prod t(f_j|e_i) \]

Fertility Model
\[ \prod n(\phi_i|e_i) \]

Distortion Model
Head
\[ \prod d_1(j - k|A(e)B(f)) \]
Non-Head
\[ \prod d_{1>}(j - j'|B(f)) \]

NULL Translation Model
\[ e \quad \text{NULL} \quad \ldots \]
\[ (m-\phi_0)p_0^{m-2\phi_0}p_1^{\phi_0} \]
\[ f \quad \ldots \quad f_j \quad \ldots \quad f_j' \quad \ldots \]
Problems of SMT — Modeling

- Good statistical translation model?
Problems of SMT — Modeling

- Good statistical translation model?
- No syntactical knowledge
- Basically, word-for-word translation considering reordering
  - Phrasal constraints implicit in IBM Model 4 and 5
  - Very good for similar language pairs
  - What about Japanese and English or others?
Problems of SMT — Modeling

- Good statistical translation model?
- No syntactical knowledge
- Basically, word-for-word translation considering reordering
  - Phrasal constraints implicit in IBM Model 4 and 5
  - Very good for similar language pairs
  - What about Japanese and English or others?

An example of viterbi alignment for F-E (from Mathematics of SMT)

the program has been implemented

le programme a été mis en application
Problems of SMT — Modeling

- Good statistical translation model?
- No syntactical knowledge
- Basically, word-for-word translation considering reordering
  - Phrasal constraints implicit in IBM Model 4 and 5
  - Very good for similar language pairs
  - What about Japanese and English or others?

An example of viterbi alignment for J-E

do you have some good medicine for a fever

熱 の 薬 は あり ません か
Problems of SMT — Training

- Possible to estimate good parameters?
Problems of SMT — Training

- Possible to estimate good parameters?
- EM-algorithm with bootstrapping
  - start with simpler models, such as
    - IBM Model 1 or 2 — word-for-word translation model
    - HMM Model — alignment with 1st order dependency
to determine initial parameters
- Impossible to enumerate all the possible alignments
  (inevitable for IBM Model 3 – 5)
  Pegging
  - \( \sum \) over *neighbours* of probable alignments
  - *probable alignments* derived from IBM Models 1 or 2
Problems of SMT — Search

- Given an input, can we translate it?

NP-complete problem — Traveling Salesman Problem

visit all the cities (input words)

visit some of the hotels in a city (output words)

(Almost) linear alignment (with local reordering) for G-E, F-E etc.

What about J-E? — drastical reordering
Problems of SMT — Search

Given an input, can we translate it?

- input length = 10, output length = 11 and 20,000 vocabulary
  - $20,000^{11}$ possible translations
  - $(11 + 1)^{10}$ possible alignments

NP-complete problem — Traveling Salesman Problem

- visit all the cities (input words)
- visit some of the hotels in a city (output words)

(Almost) linear alignment (with local reordering) for G-E, F-E etc.

- What about J-E? — drastic reordering
Introduction to HPA

- Align bilingual text phrase-by-phrase
**Introduction to HPA**

- Align bilingual text phrase-by-phrase
- An example

I have just arrived in Kyoto

京都 に 着い た ばかり です

| in Kyoto | — | 京都 に |
| arrived in Kyoto | — | 京都 に 着い |
| have just arrived in Kyoto | — | 京都 に 着い た ばかり です |
An Example of HPA

- Pairing of nodes by syntactic categories starting from word-linkage
- Phrase alignments which maximize the number of aligned phrases
An Example of HPA

- Pairing of nodes by syntactic categories starting from word-linkage
- Phrase alignments which maximize the number of aligned phrases
An Example of HPA

- Pairing of nodes by syntactic categories starting from word-linkage
- Phrase alignments which maximize the number of aligned phrases
An Example of HPA

- Pairing of nodes by syntactic categories starting from word-linkage
- Phrase alignments which maximize the number of aligned phrases
An Example of HPA

- Pairing of nodes by syntactic categories starting from word-linkage
- Phrase alignments which maximize the number of aligned phrases
An Example of HPA

Pairing of nodes by syntactic categories starting from word-linkage

Phrase alignments which maximize the number of aligned phrases
An Example of HPA

- Pairing of nodes by syntactic categories starting from word-linkage
- Phrase alignments which maximize the number of aligned phrases
Chunking by HPA

- Chunking by extracting low-level phrases
Chunking by HPA

Chunking by extracting low-level phrases
Chunking Model

- Create a model by treating each chunk as a token

  business class $\rightarrow$ business:class
  is fully booked $\rightarrow$ is:fully:booked
  予約 で いっぱい です $\rightarrow$ 予約:で:いっぱい:です

- Bootstrapping from IBM Model 1 and create IBM Model 4
HPA Model

- A set of alignments hypothesized by HPA

\[
\mathbf{A}_{\text{HPA}} = \{(1, 2), \{0\}, \{3, 4, 5\}, \{3, 4, 5\}, \{3, 4, 5\}, \{3, 4, 5\}\}
\]

- Directly compute IBM Model 4 parameters w/o pegging

\[
tc(f|e; \mathbf{f}, \mathbf{e}, \mathbf{A}_{\text{HPA}}) = \sum_{\mathbf{a} \in \mathbf{A}_{\text{HPA}}} P(\mathbf{a}|\mathbf{f}, \mathbf{e}) \sum_{i,j} \delta(f, f_j)\delta(e, e_{a_j})
\]

\[
t(f|e) \leftarrow \sum_{s \in \text{train}} tc(f|e; \mathbf{f}_s, \mathbf{e}_s, \mathbf{A}_s)
\]
HPA+train Model

- Use the HPA model (= IBM Model 4) as initial parameters for further training of IBM Model 4
- Use pegged alignments
Overview of Models

- bilingual text
- phrase alignment
- chunks of words
- HPA Model
- HPA+train Model
- Baseline Model
- Chunking Model

HPA based SMT

training model 1
... model 4

training model 1
... model 4

bootstrapping from HPA model

training model 4
## Experimental Results — Settings

### Corpus

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Japanese</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of sentences</td>
<td>145,432</td>
<td></td>
</tr>
<tr>
<td>number of words</td>
<td>835,048</td>
<td>896,302</td>
</tr>
<tr>
<td>vocabulary size</td>
<td>13,162</td>
<td>20,348</td>
</tr>
<tr>
<td>average sentence length</td>
<td>5.74</td>
<td>6.16</td>
</tr>
<tr>
<td>trigram perplexity</td>
<td>36.03</td>
<td>32.93</td>
</tr>
</tbody>
</table>

### Chunking

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Japanese</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of chunks</td>
<td>7,604</td>
<td>6,750</td>
</tr>
<tr>
<td>vocabulary size (of chunks)</td>
<td>2,166</td>
<td>1,624</td>
</tr>
<tr>
<td>average number of chunks per sentence</td>
<td>0.759</td>
<td>0.673</td>
</tr>
<tr>
<td>average number of words per chunk</td>
<td>2.21</td>
<td>2.52</td>
</tr>
<tr>
<td>trigram perplexity</td>
<td>72.36</td>
<td>72.07</td>
</tr>
</tbody>
</table>
### J-E Translation Results (1)

- Tested on 150 inputs

<table>
<thead>
<tr>
<th>Model</th>
<th>WER</th>
<th>PER</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>A</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>B</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>C</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>D</strong></td>
</tr>
<tr>
<td>baseline</td>
<td>70.2</td>
<td>59.2</td>
<td>12.7</td>
</tr>
<tr>
<td>chunking</td>
<td>64.0</td>
<td>53.1</td>
<td>21.3</td>
</tr>
<tr>
<td>HPA</td>
<td>64.5</td>
<td>58.1</td>
<td>17.3</td>
</tr>
<tr>
<td>HPA+train</td>
<td>71.0</td>
<td>59.3</td>
<td>16.0</td>
</tr>
</tbody>
</table>

WER: word error rate
PER: position independent word error rate
SE: subjective evaluation (A: perfect, B: fair, C: acceptable, D: nonsense)
# J-E Translation Results (2)

<table>
<thead>
<tr>
<th>Model</th>
<th>WER</th>
<th>PER</th>
<th>SE(A+B+C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>length</td>
<td>6 8 10</td>
<td>6 8 10</td>
<td>6 8 10</td>
</tr>
<tr>
<td>baseline</td>
<td>66.6 67.5 76.6</td>
<td>56.8 60.7 60.0</td>
<td>66.0 64.0 52.0</td>
</tr>
<tr>
<td>chunking</td>
<td>54.5 57.0 80.6</td>
<td>48.4 48.9 62.0</td>
<td>78.0 72.0 48.0</td>
</tr>
<tr>
<td>HPA</td>
<td>59.5 65.7 68.4</td>
<td>55.3 60.7 58.4</td>
<td>72.0 66.0 56.0</td>
</tr>
<tr>
<td>HPA+train</td>
<td>64.3 72.6 76.2</td>
<td>55.8 62.5 59.7</td>
<td>78.0 72.0 60.0</td>
</tr>
</tbody>
</table>
Sample Translations

ステーキの焼き具合はどうされますか

baseline: (D) can you steak
chunking: (A) how do you like your:steak
HPA+train: (A) how do you like your steak

ゴルフ場の予約できますか

baseline: (C) can i make-a-reservation
chunking: (A) can:i make-a-reservation:for golf
HPA+train: (A) could you make-a-reservation for the golf course

シカゴからシアトルまでどのくらい時間かかりますか

baseline: (A) how-long does it take to seattle from chicago
chunking: (A) how-long will:it:take to seattle from chicago
HPA+train: (B) do you how-long will it take to seattle from chicago
(please be sure to secure the best available seats for us)

base line:  (B)  i would like a seat in a great place please
chunking:  (D)  what 's the maximum area for sends providing seats
HPA+train: (D)  my best regards to your seat find a place please

(i am a beginner may i join)

base line:  (D)  do you have may but take beginner
chunking:  (D)  can:i join beginners ring
HPA+train: (D)  it is but i am a beginner