Chunk-based Statistical Translation

Taro Watanabe†, Eiichiro Sumita and Hiroshi G. Okuno
taro.watanabe@atr.co.jp

ATR Spoken Language Translation Research Laboratories
Contents

- Statistical Machine Translation
- Word Alignment Based Statistical Translation
- Chunk-based Statistical Translation
- Experiments
- Summary
Translation from $J$ into $E$

\[ \hat{E} = \arg\max_E P(E|J) \]
\[ = \arg\max_E P(E)P(J|E) \]
Statistical Machine Translation

Translation from $J$ into $E$

$$
\hat{E} = \arg\max_E P(E|J)
= \arg\max_E P(E)P(J|E)
$$

Diagram:
- Translation Model: $P(J|E)$
- Language Model: $P(E)$
- Decoder:
  $$
  \arg\max_E P(J|E)P(E)
  $$

$\hat{E}$
Word Alignment Based Statistical Translation

\[ P(J|E) = \sum_A P(J, A|E) \]

\[ E = \text{NULL}_0 \quad \text{show}_1 \quad \text{me}_2 \quad \text{the}_3 \quad \text{one}_4 \quad \text{in}_5 \quad \text{the}_6 \quad \text{window}_7 \]

\[ J = \quad \text{uindo}_1 \quad \text{no}_2 \quad \text{shinamono}_3 \quad \text{o}_4 \quad \text{mise}_5 \quad \text{tekudasai}_6 \]

\[ A = ( \quad 7 \quad 0 \quad 4 \quad 0 \quad 1 \quad 1 \quad 1 ) \]
\[ P(J|E) = \sum_A P(J, A|E) \]

\[ E = \text{NULL}_0 \text{ show}_1 \text{ me}_2 \text{ the}_3 \text{ one}_4 \text{ in}_5 \text{ the}_6 \text{ window}_7 \]

\[ J = \text{uindo}_1 \text{ no}_2 \text{ shinamono}_3 \text{ o}_4 \text{ mise}_5 \text{ tekudasai}_6 \]

\[ A = (\begin{array}{cccccc}
7 & 0 & 4 & 0 & 1 & 1
\end{array}) \]

- Generative Process of \( P(J, A|E) \)
An Example — IBM Model 4

show₁ → show → show → mise → uindo₁
me₂ → show → NULL → no → no₂
the₃ → one → show → tekudasai
one₄ → window → NULL → o → o₄
in₅ → one → shinamono
the₆ → window → uindo
window₇ → Fertility

Fertility:
\[
n(2|E₁) \quad n(0|E₂) \quad n(0|E₃) \quad \ldots
\]

Lexicon:
\[
t(J₅|E₁) \quad t(J₆|E₁) \quad t(J₃|E₄) \quad \ldots
\]

Distortion:
\[
d₁(1 - \frac{3}{1}|E₄, J₁) \quad d₁(3 - \frac{5+6}{2}|E₁, J₃) \quad d₁(5 - \frac{2+4}{2}|NULL, J₅) \quad d_{>1}(6 - 5|J₆)
\]
An Example — IBM Model 4

- **show**
- **me**
- **the**
- **one**
- **in**
- **window**

**Fertility**

\[ n(2|E_1) \]

\[ n(0|E_2) \]

\[ n(0|E_3) \]

... 

**Lexicon**

\[ t(J_5|E_1) \]

\[ t(J_6|E_1) \]

\[ t(J_3|E_4) \]

... 

**Distortion**

\[ d_1(1 - \left\lfloor \frac{3}{1}\right\rfloor|E_4, J_1) \]

\[ d_1(3 - \left\lfloor \frac{5+6}{2}\right\rfloor|E_1, J_3) \]

\[ d_1(5 - \left\lfloor \frac{2+4}{2}\right\rfloor|NULL, J_5) \]

\[ d_{>1}(6 - 5|J_6) \]
An Example — IBM Model 4

Fertility

| n(2|E₁) | n(0|E₂) | n(0|E₃) |
|--------|--------|--------|
|         |         |        |

Lexicon

| t(J₅|E₁) | t(J₆|E₁) | t(J₃|E₄) |
|--------|--------|--------|
|         |         |        |

Distortion

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
An Example — IBM Model 4

show$_1$ → show → show → mise → uindo$_1$
me$_2$ → show → NULL → no → no$_2$
the$_3$ → one → show → tekudasai → shinamono$_3$
one$_4$ → window → NULL → o → o$_4$
in$_5$ → one → shinamono → mise$_5$
the$_6$ → window → uindo → tekudasai$_6$

Fertility:
\[
n(2|E_1) \\
n(0|E_2) \\
\frac{4}{(2)}P_0^{4-2}P_1^2 \\
(0|E_3) \\
\ldots
\]

Lexicon:
\[
t(J_5|E_1) \\
t(J_6|E_1) \\
t(J_3|E_4) \\
\ldots
\]

Distortion:
\[
d_1(1 - \left\lceil \frac{3}{1} \right\rceil|E_4, J_1) \\
d_1(3 - \left\lceil \frac{5+6}{2} \right\rceil|E_1, J_3) \\
d_1(5 - \left\lceil \frac{2+4}{2} \right\rceil|NULL, J_5) \\
d_{>1}(6 - 5|J_6)
\]
An Example — IBM Model 4

show₁ → show → show → mise → uindo₁
me₂ → show → NULL → no → no₂
the₃ → one → show → tekudasai → shinamono₃
one₄ → window → NULL → o → o₄
in₅ → one → shinamono → mise₅
the₆ → window → uindo → tekudasai₆
window₇

Fertility
n(2|E₁)
n(0|E₂)
(4)₄ P₀⁴−2 P₁²
n(0|E₃)
...

Lexicon
t(J₅|E₁)
t(J₆|E₁)
t(J₃|E₄)
...

Distortion
d₁ (1 − |E₄, J₁|
(d₁ (3 − |E₁, J₃|
d₁ (5 − |NULL, J₅|
(d₁ (6 − 5|J₆)
Problems

- **Strategy:** Generate a set of words from each source word and reorder them.
Problems

- **Strategy**: Generate a set of words from each source word and reorder them.

- **Insertion/Deletion Modeling**
  - ♦ Fertility Model to select deletion
  - ♦ A binomial distribution to determine insertion
Problems

- **Strategy:** Generate a set of words from each source word and reorder them.

- **Insertion/Deletion Modeling**
  - Fertility Model to select deletion
  - A binomial distribution to determine insertion

- **Local Alignment Modeling**
  - Collection of Local Reordering $\rightarrow$ Global Reordering
  - Long distance word alignment
Chunk-based Statistical Translation

\[ P(J|E) = \sum_{\mathcal{J}} \sum_{\mathcal{E}} P(J, \mathcal{J}, \mathcal{E}|E) \]

\( \mathcal{J}, \mathcal{E} \): sequences of chunks (\(|\mathcal{J}| = |\mathcal{E}|\))
Chunk-based Statistical Translation

\[ P(J|E) = \sum \sum P(J, J, E|E) \]

\[ P(J, J, E|E) = \sum \sum \sum P(J, J, A, A, E|E) \]

\( J, E \): sequences of chunks (\(|J| = |E|\))

\( A \): chunk alignment

\( \mathcal{A} \): word alignment
Chunk-based Statistical Translation

\[
P(J|E) = \sum \sum P(J, J', E|E)
\]

\[
P(J, J', E|E) = \sum \sum P(J, J', A, A', E|E)
\]

\[
E = \begin{align*}
&\text{show}_1 \text{ me}_2 \\
&\text{mise}_5 \text{ tekudasai}_6
\end{align*}
\]

\[
J = \begin{align*}
&\text{uindo}_1 \text{ no}_2 \\
&\text{shinamono}_3 \text{ o}_4
\end{align*}
\]

\[
A = (3, 2, 1)
\]

\[
A = (\begin{bmatrix} 7 & 0 \end{bmatrix}, \begin{bmatrix} 4 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 \end{bmatrix})
\]
Model Structure

Chunking
Deletion & Fertility
Insertion
Lexicon
Reorder

ACL 2003 – p. 8/19
Chunking: Choose Chunk Size — \( \prod_i \epsilon(\varphi_i|E_i) \)

\( \varphi_i = \) chunk size and if \( \varphi_i > 0 \) then, \( E_i \) is a head word
Chunking: Associate Non-Head Words — $\prod_{i: \varphi_i=0} \eta(c(E_{h_i})|h_i - i, c(E_i))$
Deletion — $\prod_{i: \varphi_i = 0} \delta(d_i | c(E_i), c(E_{hi}))$

ACL 2003 – p. 8/19
Model Structure

Chunking

Deletion & Fertility

Insertion

Lexicon

Reorder

Chunk Reorder

Fertility — $\prod_{i : \phi_i > 0} \nu(\phi_i | E_i) / \phi_i$

$\phi_i = \# \text{ of words}$
Model Structure

Insertion — $\prod_{i: \varphi_i > 0} \iota(\phi'_i | c(E_i))$

$\phi'_i = \# \text{ of NULL words}$
Model Structure

\[
\text{Chunking} \quad \text{Deletion} \quad \text{Insertion} \quad \text{Lexicon} \quad \text{Reorder}
\]

Reorder — \( \prod_j P(\mathcal{A}_j | \mathcal{E}_{A_j}, \mathcal{J}_j) \)
Model Structure

Chunk Reorder — $P(A|\mathcal{E}, \mathcal{J})$

ACL 2003 – p. 8/19
Characteristics

- String-to-String translation model with a hidden chunk layer
Characteristics

- String-to-String translation model with a hidden chunk layer
- Bag-of-words to bag-of-words translation
  - Chunking – Translate – Reorder
Characteristics

- String-to-String translation model with a hidden chunk layer
- Bag-of-words to bag-of-words translation
  - Chunking – Translate – Reorder
- Chunk-wise word insertion vs. Sentence-wise insertion
Characteristics

- String-to-String translation model with a hidden chunk layer
- Bag-of-words to bag-of-words translation
  - Chunking – Translate – Reorder
- Chunk-wise word insertion vs. Sentence-wise insertion
- Chunking/Translate/Reorder by hypothesized “head” words
Parameter Estimation

- EM-Algorithm
- E-step: for each pair $E$ and $J$

$$P(J, A, \mathcal{A}, \mathcal{E}|J, E) = \frac{P(J, J, A, \mathcal{A}, \mathcal{E}|E)}{\sum_{J, A, \mathcal{A}, \mathcal{E}} P(J, J, A, \mathcal{A}, \mathcal{E}|E)}$$

Then, computes expectation

- M-step: From expectation, induce parameters
Some Tricks

- Computational problem
- Local maximum problem
Some Tricks

- Computational problem
  - Inside-Outside Algorithm
  - Approximation

- Local maximum problem
Some Tricks

- Computational problem
  - Inside-Outside Algorithm

- Approximation

- Local maximum problem
Some Tricks

- Computational problem
  - Inside-Outside Algorithm
  - Approximation
    - All possible word alignment: $O(lmk^4(k + 1)^k)$
    - All possible chunking/alignment: $O(2^l2^m n!)$

- Local maximum problem
Some Tricks

- Computational problem
  - Inside-Outside Algorithm
  - Approximation
    - All possible word alignment: $O(lmk^4(k + 1)^k))$
    - All possible chunking/alignment: $O(2^l2^m n!)$
    - Viterbi Chunking/Alignment + Neighbours

- Local maximum problem
Some Tricks

- Computational problem
  - Inside-Outside Algorithm
  - Approximation
- Local maximum problem
Some Tricks

- Computational problem
  - Inside-Outside Algorithm
  - Approximation

- Local maximum problem
  - Initial parameters from IBM Model 4
  - Smoothing
Decoding

- **Left-to-right generation breadth-first beam search**
  - Generate possible output chunks for all possible input chunks
  - Generate hypothesized output by consuming input chunks in arbitrary order and combining possible output chunks in left-to-right order
Decoding

- Left-to-right generation breadth-first beam search
  - Generate possible output chunks for all possible input chunks
  - Generate hypothesized output by consuming input chunks in arbitrary order and combining possible output chunks in left-to-right order

- Pruning
  - Beam size pruning
  - Example-based scoring

\[
\log P_{tm}(J|E) + \log P_{lm}(E) + \text{weight} \times \sum_j \text{freq}(E_{A_j}, J_j)
\]

- Chunk-based translation model is a deficient model
- Many model components
Japanese-to-English Translation Experiments

<table>
<thead>
<tr>
<th>Basic Travel Expression Corpus</th>
<th>Japanese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td># of sentences</td>
<td>171,894</td>
<td></td>
</tr>
<tr>
<td># of words</td>
<td>1,181,188</td>
<td>1,009,065</td>
</tr>
<tr>
<td>vocabulary size</td>
<td>20472</td>
<td>16232</td>
</tr>
<tr>
<td># of singletons</td>
<td>82,06</td>
<td>5,854</td>
</tr>
<tr>
<td>3-gram perplexity</td>
<td>23.7</td>
<td>35.8</td>
</tr>
</tbody>
</table>

- model4: IBM Model 4
- chunk3: Chunk-based Statistical Translation (chunk size ≤ 3)
- chunk3+: + Example-based scoring
Sample Viterbi Chunking/Alignment

[i * have]  [the * number]  [of my * passport]
[* パスポートの] [* 番号の] [控え] [はあります]

[i * have]  [a * stomach ache]  [please * give me]  [some * medicine]
[お腹が * 痛い] [* ので] [* 薬を] [* 下さい]

[i * have]  [a * reservation]  [* for]  [two * nights]  [my * name is]  [risa kobayashi]
[二 *泊] [* の] [予約を * し] [ている * のです] [が * 名前は] [小林 * リサです]
Evaluation

**WER:**  Word-error-rate, which penalizes the edit distance against reference translations.

**PER:**  Position independent WER, which penalizes without considering positional disfluencies.

**BLEU:**  BLEU score, which computes the ratio of n-gram for the translation results found in reference translations.

**SE:**  Subjective evaluation ranks ranging from A to D (A:Perfect, B:Fair, C:Acceptable and D:Nonsense), judged by native speakers.

- Tested on 510 sentences
- 16 set of references for non-subjective evaluations
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>WER [%]</th>
<th>PER [%]</th>
<th>BLEU [%]</th>
<th>SE [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>model4</td>
<td>43.3</td>
<td>37.2</td>
<td>46.5</td>
<td>59.2</td>
</tr>
<tr>
<td>chunk3</td>
<td>40.9</td>
<td>36.1</td>
<td>48.4</td>
<td>59.8</td>
</tr>
<tr>
<td>chunk3+</td>
<td>38.5</td>
<td>33.7</td>
<td>52.1</td>
<td>65.1</td>
</tr>
<tr>
<td>Model</td>
<td>WER [%]</td>
<td>PER [%]</td>
<td>BLEU [%]</td>
<td>SE [%] A</td>
</tr>
<tr>
<td>--------</td>
<td>---------</td>
<td>---------</td>
<td>----------</td>
<td>---------</td>
</tr>
<tr>
<td>model4</td>
<td>43.3</td>
<td>37.2</td>
<td>46.5</td>
<td>59.2</td>
</tr>
<tr>
<td>chunk3</td>
<td>40.9</td>
<td>36.1</td>
<td>48.4</td>
<td>59.8</td>
</tr>
<tr>
<td>chunk3+</td>
<td>38.5</td>
<td>33.7</td>
<td>52.1</td>
<td>65.1</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>WER [%]</th>
<th>PER [%]</th>
<th>BLEU [%]</th>
<th>SE [%]</th>
<th>A</th>
<th>A+B</th>
<th>A+B+C</th>
</tr>
</thead>
<tbody>
<tr>
<td>model4</td>
<td>43.3</td>
<td>37.2</td>
<td>46.5</td>
<td>59.2</td>
<td>74.1</td>
<td>80.2</td>
<td></td>
</tr>
<tr>
<td>chunk3</td>
<td>40.9</td>
<td>36.1</td>
<td>48.4</td>
<td>59.8</td>
<td>73.5</td>
<td>78.8</td>
<td></td>
</tr>
<tr>
<td>chunk3+</td>
<td>38.5</td>
<td>33.7</td>
<td>52.1</td>
<td>65.1</td>
<td>76.3</td>
<td>80.6</td>
<td></td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>WER [%]</th>
<th>PER [%]</th>
<th>BLEU [%]</th>
<th>SE [%]</th>
<th>A</th>
<th>A+B</th>
<th>A+B+C</th>
</tr>
</thead>
<tbody>
<tr>
<td>model4</td>
<td>43.3</td>
<td>37.2</td>
<td>46.5</td>
<td>59.2</td>
<td>74.1</td>
<td>80.2</td>
<td></td>
</tr>
<tr>
<td>chunk3</td>
<td>40.9</td>
<td>36.1</td>
<td>48.4</td>
<td>59.8</td>
<td>73.5</td>
<td>78.8</td>
<td></td>
</tr>
<tr>
<td>chunk3+</td>
<td>38.5</td>
<td>33.7</td>
<td>52.1</td>
<td>65.1</td>
<td>76.3</td>
<td>80.6</td>
<td></td>
</tr>
<tr>
<td>input:</td>
<td>一五二便の荷物はここで全部ですか</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>
</tr>
<tr>
<td>reference:</td>
<td>is this all the baggage from flight one five two</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>model4:</td>
<td>is this all you baggage for flight one five two</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chunk3+:</td>
<td>is this all the baggage from flight one five two</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>input:</th>
<th>朝食をルームサービスでお願いします</th>
</tr>
</thead>
<tbody>
<tr>
<td>reference:</td>
<td>may i have room service for breakfast please</td>
</tr>
<tr>
<td>model4:</td>
<td>please give me some room service please</td>
</tr>
<tr>
<td>chunk3+:</td>
<td>i’d like room service for breakfast</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>input:</th>
<th>もしもし三月十九日の予約を変更したいのですか</th>
</tr>
</thead>
<tbody>
<tr>
<td>reference:</td>
<td>hello i’d like to change my reservation for march nineteenth</td>
</tr>
<tr>
<td>model4:</td>
<td>i’d like to change my reservation for ninety days be march hello</td>
</tr>
<tr>
<td>chunk3+:</td>
<td>hello i’d like to change my reservation on march nineteenth</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>input:</th>
<th>二三分待って下さい今電話中なんです</th>
</tr>
</thead>
<tbody>
<tr>
<td>reference:</td>
<td>wait a couple of minutes i’m telephoning now</td>
</tr>
<tr>
<td>model4:</td>
<td>is this the line is busy now a few minutes</td>
</tr>
<tr>
<td>chunk3+:</td>
<td>i’m on another phone now please wait a couple of minutes</td>
</tr>
</tbody>
</table>
Summary

- String-to-String translation model with hidden chunks
- More hidden variables
  - More cost for training + decoding
    - Training Cost \(\approx\) IBM Model 5 with pegging
    - Decoding Cost: moderate with Example-based scoring
- Quality Improvement: Slightly, but (probably) significant
Summary

- String-to-String translation model with hidden chunks
- More hidden variables
  → More cost for training + decoding
  ♦ Trainin Cost ≈ IBM Model 5 with pegging
  ♦ Decoding Cost: moderate with Example-based scoring
- Quality Improvement: Slightly, but (probably) significant
- Other approaches?
Typology of Statistical Machine Translation

- Approach 1: Precomputation of Structure
- Approach 2: Structure-to-String
- Approach 3: Collection of Hierarchical FST
Typology of Statistical Machine Translation

- Approach 1: Precomputation of Structure
- Approach 2: Structure-to-String
- Approach 3: Collection of Hierarchical FST
Typology of Statistical Machine Translation

- **Approach 1: Precomputation of Structure**
  - Templates (Och et al. 1999)
  - Chunks from syntax-based phrase alignment (Watanabe et al. 2002)
  - Direct phrase induction (Marcu and Wong 2002)
    - Bias the training corpus by template, chunk or phrase
    - Works significantly better on observed word sequences, but not for unseen sequences

- **Approach 2: Structure-to-String**

- **Approach 3: Collection of Hierarchical FST**
Typology of Statistical Machine Translation

- Approach 1: Precomputation of Structure
- Approach 2: Structure-to-String
- Approach 3: Collection of Hierarchical FST
Typology of Statistical Machine Translation

- Approach 1: Precomputation of Structure
- Approach 2: Structure-to-String
  - Phrase-to-string Modeling (Wang 1998)
  - Syntax-to-string Modeling (Yamada and Knight 2001)
    - Bias the source part of a training corpus by “structure”
    - Computationally cheaper
    - Relies on the monolingual processing (parser or chunker)
- Approach 3: Collection of Hierarchical FST
Typology of Statistical Machine Translation

- Approach 1: Precomputation of Structure
- Approach 2: Structure-to-String
- Approach 3: Collection of Hierarchical FST
Typology of Statistical Machine Translation

- Approach 1: Precomputation of Structure
- Approach 2: Structure-to-String
- Approach 3: Collection of Hierarchical FST
  - (Alshawi et al. 2000)
  - Deterministic vs. Non-Deterministic
  - Faster decoding + less space vs. Slow decoding + pruning
  - Limited domain vs. Larger domain