Extracting Equivalent Chunks from Chinese-English Bilingual Corpus

Yanjun Ma

National Centre for Language Technology, Dublin City University

Note: This work is part of my master thesis, Tsinghua University, Beijing, China
Outline

- Related research on chunk alignment
- Framework for chunk alignment
- Word alignment using Knowledge base
- Bilingual chunking based on Marker Hypothesis
- Log-linear model for chunk alignment
- Experiment and analysis
- Future work
Related research

- Different levels of alignment
  - Passage, paragraph alignment
  - Sentence alignment
  - Word alignment
  - Chunk (phrase, sub-sentential) alignment
Related Research (cont.)

- Generalized chunk styles
  - Consecutive word sequence (SMT)
  - CoNLL-2000 style chunks
  - Marker chunks (EBMT style chunks)
  - Sub trees in a grammar tree
Outline

- Related research on chunk alignment
- Framework for chunk alignment
- Word alignment using Knowledge base
- Bilingual chunking based on Marker Hypothesis
- Log-linear model for chunk alignment
- Experiment and analysis
- Future work
Framework for chunk alignment

Bilingual chunking

Word alignment

Chunk alignment

Chinese-English sentences

Chinese word segmentation and tagging

English tokenization and tagging

Marker/POS mapping

Marker chunk identification

Chinese BaseNP identification

English BaseNP identification

Hybrid chunk identification

'anchor chunk' alignment

Log-linear model ranking

Ambiguous chunk alignment

In chunk alignment

non-chunk alignment

Word alignment based on dictionary

Word alignment based on EM algorithm

Bilingual dictionary

Chinese lexicon

Bilingual corpus
Outline

- Related research on chunk alignment
- Framework for chunk alignment
- Word alignment using Knowledge base
- Bilingual chunking based on Marker Hypothesis
- Log-linear model for chunk alignment
- Experiment and analysis
- Future work
Word alignment---overview

- Statistical V.S Knowledge base
- Statistical approach
  - Heuristic word alignment
  - IBM model 1~5
- Word alignment using knowledge base
  - Bilingual dictionary
  - thesaurus
Word alignment---1:1 alignment algorithm

\[ C = \{<c_1, 1>, <c_2, 2>, \ldots <c_J', J>\}; \quad \text{// set of Chinese words} \]
\[ E = \{<e_1, 1>, <e_2, 2>, \ldots <e_I', I>\}; \quad \text{// set of English words} \]
\[ A = \emptyset; \quad \text{// set of word alignment} \]

\[
\text{foreach} \quad <c_j', j> \in C, 1 \leq j' \leq j \leq J, \quad <e_i', i> \in E, 1 \leq i' \leq i \leq I \{ \\
\quad \text{if} \quad \text{CEAlignScore} \left( c_j', e_i' \right) > h \\
\quad \quad \text{add} \quad (A, <c_j', j, e_i', i>); \quad \text{// add an alignment} \\
\}
\]

\[
C = C - \{<c_j, j>| \text{Exist} \quad <e_i', i>, <c_j', j, e_i', i> \in A\}; \\
E = E - \{<e_i', i>| \text{Exist} \quad <c_j', j>, <c_j', j, e_i', i> \in A\}; \\
\text{foreach} \quad <c_j', j> \in C \{ \\
\quad S = \text{GetSynonym} \left( c_j' \right); \quad \text{// Get synonym given a word} \\
\quad \text{foreach} \quad c_j \in S, \quad <e_i', i> \in E \\
\quad \quad \text{if} \quad \text{CEAlignScore} \left( c_j, e_i' \right) > h \\
\quad \quad \quad \text{add} \quad (A, <c_j, j, e_i', i>); \\
\}
\]

Output \( A; \)
Word alignment---1:1 alignment algorithm (cont.)

\[
CEAlignScore(c, e) = \begin{cases} 
1 & \text{if } c \neq e \\
\max_{x \in \text{Dict}(c)} \ EESim(x, e) & \text{if } c = e 
\end{cases}
\]

\[
EESim(x, y) = \begin{cases} 
1 & \text{if } x = y \\
0.9 & \text{if } x = \text{WordStem}(y) \\
0.8 & \text{if } x = \text{SubString}(y) \text{ or } y = \text{SubString}(x) \\
0 & \text{else} 
\end{cases}
\]

- **Function description:**
  - **Dict**: Get a set of English translation given a Chinese word
  - **EESim**: Get the similarity between two English words
Word alignment---word anchors

- Define **word anchors** (WA) as follows:
  \[ WA = \{ <c_j', j, e_i', i> | \text{Count}(c_j', j, A) = 1 \]
  \[ \text{and Count}(e_i', i, A) = 1 \} \]

- Ambiguous word alignment (AA)
  \[ AA = A - WA \]
Word alignment---an example

- 而有几个人懂汉语。
  \[ A_{\text{few}} \text{ of } them \text{ know Chinese}. \]

Result of word alignment

- \(<\text{他们}, 1, \text{them}, 4>: 0.9\>
- \(<\text{汉语}, 7, \text{Chinese}, 6>: 1.0\>

Entries in Chinese-English dictionary

- 懂: understand
- 知道: know
Word alignment---an example

- Word expansion based on Chinese Thesaurus
  - $S = \text{GetSynonym}(懂) = \{\text{知道, 了解} \ldots\}$
  - 知道: know
- Get word alignment: <懂, 6, know, 5>
Outline

- Related research on chunk alignment
- Framework for chunk alignment
- Word alignment using Knowledge base
- Bilingual chunking based on Marker Hypothesis
- Log-linear model for chunk alignment
- Experiment and analysis
- Future work
Bilingual Chunking---overview

- Marker hypothesis
- Marker chunk identification
- Difficulties in Chinese-English marker chunk alignment
- Hybrid chunks: a combination of marker chunks and BaseNP
- BaseNP identification using bilingual corpus
Bilingual Chunking---Marker Hypothesis

- Green (1979)
- Natural languages are ‘marked’ for complex syntactic structure at surface form by a closed set of specific lexemes and morphemes.
# Bilingual Chunking---Marker Hypothesis

<table>
<thead>
<tr>
<th>Marker sets</th>
<th>examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;DET&gt;</td>
<td>{the, a, an, those, these, …} determiners</td>
</tr>
<tr>
<td>&lt;PREP&gt;</td>
<td>{in, on, out, with, from, to, under, …} prepositions</td>
</tr>
<tr>
<td>&lt;QUANT&gt;</td>
<td>{all, some, few, many, …} quantifiers</td>
</tr>
<tr>
<td>&lt;CONJ&gt;</td>
<td>{and, or, …} conjunctions</td>
</tr>
<tr>
<td>&lt;POSS&gt;</td>
<td>{my, your, our, …} possessive pronouns</td>
</tr>
<tr>
<td>&lt;PRON&gt;</td>
<td>{I, you, he, she, it, …} pronouns</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Marker sets</th>
<th>examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;DET&gt;</td>
<td>{这, 那, 该, 各, 这些, 任何……}</td>
</tr>
<tr>
<td>&lt;PREP&gt;</td>
<td>{在, 由, 向, 跟, 按照, 关于……}</td>
</tr>
<tr>
<td>&lt;QUANT&gt;</td>
<td>{很多, 大量, 一些……}</td>
</tr>
<tr>
<td>&lt;CONJ&gt;</td>
<td>{不止, 除非, 要是, ……}</td>
</tr>
<tr>
<td>&lt;PRON&gt;</td>
<td>{我, 你, 他, 她们……}</td>
</tr>
</tbody>
</table>
Bilingual chunking ---marker chunk identification

- English marker words can be identified without parsing
- Unfortunately, Chinese marker words can’t!
- Chinese marker words are extremely ambiguous
- One-one mapping between marker sets and POS tags
  marker word identification $\leftrightarrow$ POS tagging
Bilingual chunking --- marker chunk identification

Mapping between POS tags and marker sets

<table>
<thead>
<tr>
<th>English POS</th>
<th>Marker set</th>
<th>Chinese POS</th>
<th>Marker set</th>
</tr>
</thead>
<tbody>
<tr>
<td>IN</td>
<td>PREP</td>
<td>p</td>
<td>PREP</td>
</tr>
<tr>
<td>CD, OD</td>
<td>QUANT</td>
<td>m</td>
<td>QUANT</td>
</tr>
<tr>
<td>CC</td>
<td>CONJ</td>
<td>c</td>
<td>CONJ</td>
</tr>
<tr>
<td>DT</td>
<td>DET</td>
<td>r (determinative pronouns)</td>
<td>DET</td>
</tr>
<tr>
<td>PRP</td>
<td>PRON</td>
<td>r (person)</td>
<td>PRON</td>
</tr>
<tr>
<td>PRP$</td>
<td>POSS</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Bilingual chunking --- marker chunk identification

An example for marker chunk

抱/v 着/u 两/m 个/q 婴儿/n 的/u 妇女/n 正/d 向/p 托儿所/n 走/v 来/f 。/w

With 	two 		baby 		woman 	
to 	nursery 	walk

A/DT woman/NN with/IN two/CD babies/NNS is/VBZ coming/VBG to/TO
the/DT nursery/NN ./.

---

<BEGIN>抱/v 着/u <QUANT>两/m 个/q 婴儿/n 的/u 妇女/n 正/d <PREP>向
/p 托儿所/n 走/v 来/f <PUNC>。/w

<DET>A/DT woman/NN <PREP>with/IN two/CD babies/NNS is/VBZ
coming/VBG to/TO <DET>the/DT nursery/NN <PUNC>./.
Bilingual chunking---difficulties in Chinese-English chunk alignment

- Troubles in Chinese-English marker chunk alignment

<table>
<thead>
<tr>
<th>Marker chunks</th>
<th>Hybrid chunks</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;BEGIN&gt;抱/v 着/u &lt;QUANT&gt;两/m 个/q 婴儿/n 的/u 妇女/n 正/d &lt;PREP&gt;向/p 托儿所/n 走/v 来/f &lt;PUNC&gt;。/w</td>
<td></td>
</tr>
<tr>
<td>&lt;DET&gt;A/DT woman/NN &lt;PREP&gt;with/IN two/CD babies/NNS is/VBZ coming/VBG to/TO &lt;DET&gt;the/DT nursery/NN &lt;PUNC&gt;./.</td>
<td></td>
</tr>
<tr>
<td>&lt;BEGIN&gt;抱/v 着/u &lt;BASENP&gt;两/m 个/q 婴儿/n &lt;BACKNP&gt;的/u &lt;BASENP&gt;妇女/n 正/d &lt;PREP&gt;向/p &lt;BASENP&gt;托儿所/n &lt;BACKNP&gt;走/v 来/f &lt;PUNC&gt;。/w</td>
<td></td>
</tr>
<tr>
<td>&lt;BASENP&gt;A/DT woman/NN &lt;PREP&gt;with/IN &lt;BASENP&gt;two/CD babies/NNS is/VBZ coming/VBG to/TO &lt;BASENP&gt;the/DT nursery/NN &lt;PUNC&gt;./.</td>
<td></td>
</tr>
<tr>
<td>&lt;BEGIN&gt;抱/v 着/u &lt;PREP&gt;with/IN &lt;BASENP&gt;两/m 个/q 婴儿/n: &lt;BASENP&gt;two/CD babies/NNS &lt;BASENP&gt;妇女/n: &lt;BASENP&gt;A/DT woman/NN &lt;BASENP&gt;托儿所/n: &lt;BASENP&gt;the/DT nursery/NN</td>
<td></td>
</tr>
</tbody>
</table>
Bilingual chunking --- Hybrid chunks

- Identification of hybrid chunks
  - Marker chunks identification
    \[
    \text{<DET>A/DT woman/NN <PREP>with/IN two/CD babies/NNS is/VBZ coming/VBG to/TO <DET>the/DT nursery/NN <PUNC>./}.
    \]
  - BaseNP identification
    \[
    \text{<DET> [ A/DT woman/NN ] <PREP>with/IN [ two/CD babies/NNS ] is/VBZ coming/VBG to/TO <DET> [ the/DT nursery/NN ] <PUNC>./}.
    \]
  - Combination with some priority order
    \[
    \text{<BASENP>A/DT woman/NN <PREP>with/IN <BASENP>two/CD babies/NNS <BACKNP>is/VBZ coming/VBG to/TO <BASENP>the/DT nursery/NN <PUNC>./}.
    \]
Bilingual chunking---BaseNP identification using bilingual corpus

- BaseNP identification using bilingual corpus
  - English BaseNP identification
  - Generation of Chinese BaseNP candidates
  - Chinese BaseNP selection
Outline

- Related research on chunk alignment
- Framework for chunk alignment
- Word alignment using Knowledge base
- Bilingual chunking based on Marker Hypothesis
- Log-linear model for chunk alignment
- Experiment and analysis
- Future work
Log-linear model for chunk alignment

- Two-step hybrid chunk alignment
  - Unambiguous chunks (Anchor chunks) --- heuristics
  - Ambiguous chunks --- ranking with a log-linear model
Log-linear model for chunk alignment (cont.)

- Unambiguous chunks (anchor chunks)
  - Premise: anchor words between Chinese chunk and English chunk
  - Constraint: translation of each Chinese words in Chinese chunk is in corresponding English chunk, vice versa

- Ambiguous chunks
Log-linear model for chunk alignment (cont.)

- Ranking with a log-linear model
  - Without anchor words

- Constraints detection
Log-linear model for chunk alignment (cont.)

- Ranking based on log-linear model
  - Candidate generation
  - Feature selection
    - Scores of word alignment
    - Distance distortion
    - Marker transition probability
  - Feature weighting
Log-linear model for chunk alignment (cont.)

- Feature selection and parameter estimation

- Score of word alignment (F-WALI)

\[
Set_w = \{ <c_{j'}, j', 1, e_i, i', 1> | lwc \leq j' \leq rwc, \\
lwe \leq i' \leq rwe, \\
< c_{j'}, j', 1, e_i, i', 1 > \in Set_A \}
\]

\[
Set_w' = \{ <c_{j'}, j', 1, e_i, i', 1 > | lwc \leq j' \leq rwc, \\
lwe \leq i' \leq rwe, \\
< c_{j'}, j', 1, e_i, i', 1 > \in Set_{EM} \}
\]

\[
F - WALI = \sum_{a \in Set_w} \text{GetProb}(a) + \sum_{a \in Set_{w'}} \text{GetProb}(a)
\]
Log-linear model for chunk alignment (cont.)

- Distance distortion (F-DIST)
  - Predict the unknown from the known
  - Guess the probability of ambiguous chunk alignment using its distortion from the chunk alignment without ambiguity (anchor chunk)
Log-linear model for chunk alignment (cont.)

\[
\text{LocDist}(i,j) = \min(|\text{Slope}_L - 1|, |\text{Slope}_R - 1|)
\]

\[
\text{Slope}_L = \frac{(j-j_L)}{(i-i_L)}, \quad \text{Slope}_R = \frac{(j_R-j)}{(i_R-i)}
\]

\[
(i_L, j_L) = \arg \max_{(c_p, i', j', c_p', j')} \quad (i_R, j_R) = \arg \max_{(c_p, i', j', c_p', j')} \quad \text{for } (c_p, i', j', c_p', j') \in \text{Set}_{PA}
\]

\[
\text{LocDist}(2, 2) = 0 \quad \text{Slope}_L = \frac{(2-1)}{(2-1)} = 1; \quad \text{Slope}_R = \frac{(5-2)}{(3-2)} = 3
\]

\[
\text{LocDist}(2, 6) = 2 \quad \text{Slope}_L = \frac{(6-1)}{(2-1)} = 5; \quad \text{Slope}_R = \frac{(5-6/3-2)}{-1} = 1
\]
Log-linear model for chunk alignment (cont.)

- the less LocDist(i,j) is, the more probable (i,j) is

\[
F - DIST = \frac{1}{LocDist(cp_j, ep_{i+n})}
\]

- Marker translation table (F-MARK)
  - Matrix of translation probability from one marker to another
Log-linear model for chunk alignment (cont.)

- Feature weighting
  - YASMET toolkit

- Formula for candidate scoring

\[
\text{Score}(<cp_j, ep_i^{i+n}>) = \lambda_1 * F - \text{WALI}(<cp_j, ep_i^{i+n}>) + \\
\lambda_2 * F - \text{DIST}(<cp_j, ep_i^{i+n}>) + \\
\lambda_3 * F - \text{MARK}(<cp_j, ep_i^{i+n}>)
\]
Extracting n:n chunk alignment from 1:n chunk alignment

- **n:n chunks**

1979年是：1979 was
我儿子出生：my son was born

......
Outline

- Related research on chunk alignment
- Framework for chunk alignment
- Word alignment using Knowledge base
- Bilingual chunking based on Marker Hypothesis
- Log-linear model for chunk alignment
- Experiment and analysis
- Future work
Experiments & analysis

- Word alignment
- BaseNP identification and alignment
- Chunk alignment
Experiments & analysis---

Anchor word alignment

- Evaluation metric: in answer set, sure word alignment is SW, possible word alignment is PW. Word alignment of our system is AW.

\[
\text{precision} = \frac{|AW \cap PW|}{|AW|} \quad \text{recall} = \frac{|AW \cap SW|}{|SW|}
\]

\[
AER(SW, PW; AW) = 1 - \frac{|AW \cap SW| + |AW \cap PW|}{|AW| + |SW|}
\]
### Experiments & analysis---Anchor word alignment

<table>
<thead>
<tr>
<th>experiments</th>
<th>precision</th>
<th>recall</th>
<th>ASR</th>
</tr>
</thead>
<tbody>
<tr>
<td>DictAlign</td>
<td>94.97%</td>
<td>65.27%</td>
<td>0.3799</td>
</tr>
<tr>
<td>SynAlign</td>
<td>89.69%</td>
<td>69.01%</td>
<td>0.3416</td>
</tr>
</tbody>
</table>

#### Graph showing precision, recall, and AER for DictAlign and SynAlign experiments:

- **DictAlign** experiments:
  - Precision: 94.97%
  - Recall: 65.27%
  - AER: 0.3799

- **SynAlign** experiments:
  - Precision: 89.69%
  - Recall: 69.01%
  - AER: 0.3416
Experiments & analysis---BaseNP identification and alignment

Evaluation

In answer set the number of Chinese BaseNP is SNP, the number of Chinese BaseNP which can be aligned to English BaseNP is NNP. The number of Chinese BaseNP identified by our system is MNP, the number of correct BaseNP is CNP, that can be correctly aligned is ANP.

BaseNP identification

\[
\text{precision} = \frac{\text{CNP}}{\text{MNP}} \quad \text{recall} = \frac{\text{CNP}}{\text{SNP}}
\]

BaseNP alignment

\[
\text{precision} = \frac{\text{ANP}}{\text{MNP}} \quad \text{recall} = \frac{\text{ANP}}{\text{NNP}}
\]
Experiments & analysis---BaseNP identification and alignment

- results

<table>
<thead>
<tr>
<th>Experiments</th>
<th>precision</th>
<th>recall</th>
<th>FB1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BaseNP identification</td>
<td>92.77%</td>
<td>93.15%</td>
<td>92.96</td>
</tr>
<tr>
<td>BaseNP alignment</td>
<td>89.26%</td>
<td>73.60%</td>
<td>80.68</td>
</tr>
</tbody>
</table>
Experiments & analysis---chunk alignment

- Evaluation: extract all the aligned fragments in 200 sentence pair corpus as the answer set. The set of sure alignment is SP, possible alignment is PP.

- The set of aligned chunks in our system is AP

\[
\text{precision} = \frac{|AP \cap PP|}{|AP|} \quad \text{recall} = \frac{|AP \cap SP|}{|SP|}
\]
Experiments & analysis---chunk alignment

- Chunk identification methods
  - Marker chunks (MARKER_CHUNK)
  - Hybrid chunks (HYBRID_CHUNK)

- Experiment scheme
  - 1:n chunk alignment with heuristics (SM_AC)
  - 1:n chunk alignment plus log-linear model ranking (SM_AC+ME)
  - n:n chunk alignment: extension of SM_AC (MM_AC)
  - n:n chunk alignment: extention of SM_AC+ME (MM_AC+ME)
Experiments & analysis---chunk alignment

Marker chunk alignment

<table>
<thead>
<tr>
<th>Experiments</th>
<th>precision</th>
<th>recall</th>
<th>FB1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SM_AC</td>
<td>86.84%</td>
<td>25.40%</td>
<td>39.30</td>
</tr>
<tr>
<td>SM_AC+ME</td>
<td>78.57%</td>
<td>26.12%</td>
<td>39.21</td>
</tr>
<tr>
<td>MM_AC</td>
<td>85%</td>
<td>26.19%</td>
<td>40.04</td>
</tr>
<tr>
<td>MM_AC+ME</td>
<td>76.60%</td>
<td>27.78%</td>
<td>40.77</td>
</tr>
</tbody>
</table>

Hybrid chunk alignment

<table>
<thead>
<tr>
<th>Experiments</th>
<th>precision</th>
<th>recall</th>
<th>FB1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SM_AC</td>
<td>88.64%</td>
<td>30.16%</td>
<td>45</td>
</tr>
<tr>
<td>SM_AC+ME</td>
<td>83.67%</td>
<td>31.75%</td>
<td>46.05</td>
</tr>
<tr>
<td>MM_AC</td>
<td>90.59%</td>
<td>50.79%</td>
<td>65.09</td>
</tr>
<tr>
<td>MM_AC+ME</td>
<td>89.69%</td>
<td>56.35%</td>
<td>69.21</td>
</tr>
</tbody>
</table>
Experiments & analysis---chunk alignment

![Graph showing comparison of different chunk alignment methods.](image)

**Experiment Design**

- **SM_AC**
- **SM_AC+ME**
- **MM_AC**
- **MM_AC+ME**

**Markers**

- **Marker chunk alignment**
- **Hybrid chunk alignment**
Experiments & analysis---
chunk alignment

- Chunk alignment based on Hybrid chunk identification outperforms that based on Marker chunks.

- Extraction of n:n chunk alignments can improve recall significantly while the precision is slightly lower.

- In 1:n chunk alignment, ranking module makes slight contribution. While in n:n alignment, ranking module contributes greatly to recall.
Experiments & analysis---chunk alignment

Comparison of chunk alignment with different chunk identification methods

<table>
<thead>
<tr>
<th>experiments</th>
<th>precision</th>
<th>recall</th>
<th>FB1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMT chunks</td>
<td>94.03%</td>
<td>39.68%</td>
<td>55.81</td>
</tr>
<tr>
<td>MARK.MM.AC+ME</td>
<td>76.60%</td>
<td>27.78%</td>
<td>40.77</td>
</tr>
<tr>
<td>HYBR MM.AC+ME</td>
<td>89.69%</td>
<td>56.35%</td>
<td>69.21</td>
</tr>
</tbody>
</table>
Experiments & analysis---chunk alignment

![Graph showing chunk alignment results]

- Baseline
- MARK_MM_AC+ME
- HYBR_MM_AC+ME

Legend:
- Precision
- Recall
- F1

Outline

- Related research on chunk alignment
- Framework for chunk alignment
- Word alignment using Knowledge base
- Bilingual chunking based on Marker Hypothesis
- Log-linear model for chunk alignment
- Experiment and analysis
- Future work
Future work

- Word alignment
  - Improve recall
  - Statistical methods
- Evaluation metric for chunk alignment
- Relationship between different chunk alignment
Thanks!