Contents

- Motivation
  - Why annotation guidelines for word alignment?
- Previous work
- Proposed approach
  - Utilizing contrastive analysis of morpho-syntactic encodings
- Experimental setting & result
- Conclusion
Motivation - why annotation guidelines?

- Chinese and Korean belong to entirely different language families in terms of typology and genealogy
  - Finding correspondence b/w words is quite unclear
  - Differences in verbal systems cause most linking obscurities

- To achieve more objective, correct, and consistent evaluation results of word alignment

- How to systematically describe linguistic phenomena occurring in morpho-syntactically distant language?
  - From the perspective of contrastive analysis of morpho-syntactic encodings
Previous work (1)

- Blinker project (Melamed, 1998)
  - General guidelines
    - Omissions in translation
    - Phrasal correspondence

- ARCADE project (Veronis & Langlai, 1999) & PLUG Link Annotator (Merkel, 1999)
  - General guidelines
    - Mark as many words as necessary on both the target and source side
    - Mark as few words as possible on both the target and source side
Guidelines for Spanish-English word alignment (Patrick and et al., 2005)
- General guidelines
  - Minimum lexical unit size
  - Indivisibility rule
  - Absence of correspondence

Guidelines for Chinese-English word alignment (Upenn, 2006)
- General guidelines
  - Translated vs. Not translated
  - Minimum match vs. maximum match
  - Context-dependent translation
  - Glue approach
Previous work (3)

Detailed guidelines
- Enumerate specific annotation rules classified by lexical categories such as Part of Speech (POS)

Summary of previous work
- General guidelines
  - Also useful for Chinese-Korean word alignment
  - Detailed guidelines
    - Cannot systematically describe linguistic phenomena occurring in morpho-syntactically distant language pairs
Some issues in annotation guidelines

General guidelines summarized by Veronis & Langlais
- Mark as many words as necessary on both the target and source side
- Mark as few words as possible on both the target and source side

S(ure) vs. P(ossible) link
- P link: no need to reach an agreement

‘Not translated’
- Null link
Proposed approach

- Propose guidelines utilizing contrastive analysis of morpho-syntactic encodings

- Most linking obscurities are caused by differences in morphological form of verbs

Proposed approach:
- First, investigating the grammatical categories Korean verbs convey
- Then, finding the corresponding elements in Chinese
General comparison

Chinese is an isolating language, while Korean is an agglutinative one

- Morphological form of Korean is much more complex than that of Chinese

[cn] 我(I) / 曾(already) / 去(go) / 过(Prt.) / 北京(Beijing) / 。
I have been to Beijing.

[ko] 나(I)+는 北京(Beijing)+에 가(go) 보+ㄴ 적+이 있+다.
General comparison

Chinese is an isolating language, while Korean is an agglutinative one

Morphological form of Korean is much more complex than that of Chinese

[cn] 我(I) / 曾(already) / 去(go) / 过(Prt.) / 北京(Beijing) / 。
I have been to Beijing.

[ko] 나는 북경(Beijing)에 가(go) 보+ㄴ 적+이 있다.
eojjeol
General comparison

Chinese is an isolating language, while Korean is an agglutinative one.

Morphological form of Korean is much more complex than that of Chinese.

[cn] 我(I) / 曾(already) / 去(go) / 过(Prt.) / 北京(Beijing) /。I have been to Beijing.
[ko] 나(I)+는 북경(Beijing)+에 가(go) 보+ㄴ 적+이 있+다.
General comparison

🌟 Chinese is an isolating language, while Korean is an agglutinative one

➤ Morphological form of Korean is much more complex than that of Chinese

[cn] 我(I) / 曾(already) / 去(go) / 过(Prt.) / 北京(Beijing) /。
I have been to Beijing.

[ko] 나(I)+는 북경(Beijing)+에 가(go) 보+ㄴ적+이 있다.

Function word
General comparison

- An *eojeol* in Korean
  - One or more stem (content) + function morphemes
  - Function morphemes (inflection): postposition or verbal affixes
  - Function morphemes occupy 41.3% of all Korean morphemes

- Average # of function morphemes inflected by a verb is 1.94, while that of content morphemes is 0.7

→ Korean verbal affixes causes uncertain alignment cases
→ Understanding the organization of Korean verb is crucial
Comparison of verbal systems b/w Chinese and Korean (1)

- A verbal phrase in Korean
  - A verb stem + a series of verbal affixes
  - Verbal affixes are ordered in a relative sequence
  - Express various modality information viz. tense, aspect, mood, negation, and voice

[ko] 먹(stem)고_있(aspect)었(aspect)었(tense)다(mood)
    had been eating

[ko] 잡(stem)히(passive)었(aspect)겠(modality)다(mood)
    may have been captured

→ Correspondences in Chinese are mainly composed of features used to display Chinese modality information
Comparison of verbal systems b/w Chinese and Korean (2)

- Difference of modal expression b/w two languages
  - Korean: intensively by verbal affixes of complex inflectional forms
  - Chinese: discontinuous morphemes around lexical verbs

- Prominence and correlations of modality system increases the annotation ambiguity
  - Chinese is an aspect- and topic- prominent language
  - Tense, aspect, and mood are interconnected within ‘temporal structure’ of an event
  - Some negative particles can imply aspect information in Chinese

→ Need to clarify the method for expressing modality information in Chinese
Special Guidelines based on Korean Verbal System (1)

- General annotation principle
  - First, judge Korean verbal phrases
  - Korean is a verb-final language
  - Then, match the correspondent words in Chinese

- Allow phrasal correspondences and different link types
  - S-link, P-link, and not-translated (Null-link)

- Explicit and unambiguous correspondences are S-linked and implicit correspondences are P-linked
  - Annotators may have disagreements on P-links
Give an explanation based on five grammatical categories such as tense, aspect, mood, negation, and voice

Compose most of the modal expression in Chinese

For example, aspect system in Chinese

An aspect prominent language with a complete set of markers to express distinct aspectual distinctions (Xiao, 2002)

Aspect markers

- Aspectual particles & adverbs
- Verb reduplication
  - Idiosyncratic linguistic form in Chinese
- Resultative Verb Complement (RVC)
  - Ex. Push the door *open*
Aspect system in Chinese

**Aspectual particle & Adverb**
- [cn] 我(I) / 曾(already) / 去(go) / 过(Prt.) / 北京(Beijing) /。
- [ko] 나(I) +는 북경(Beijing) +에 가(go) 보 + 느 적 + 01 있 + 다.

**Verb reduplication**
- [cn] 我(I) / 看(Prt.) / 看(read) / 报纸(newspaper) /。
- [ko] 나(I) +는 신문(newspaper) +을 보(read) + 았 + 다.

**RVC**
- [cn] 大家(everybody) / 把(Prep.) / 作业(homework) / 交(submit) / 上来(RVC) /。
- [ko] 모두(everybody) 숙제(homework) +을 내(submit) 주 + 세 + 요.
Corpus data

<table>
<thead>
<tr>
<th></th>
<th>Chinese</th>
<th>Korean</th>
</tr>
</thead>
<tbody>
<tr>
<td># of sentences</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td># of words</td>
<td>1,323</td>
<td>1,502</td>
</tr>
<tr>
<td># of singletons</td>
<td>741</td>
<td>645</td>
</tr>
<tr>
<td>Avg. length</td>
<td>26.5</td>
<td>30.4</td>
</tr>
</tbody>
</table>

Statistics for test data

- Sentence-aligned parallel corpus from the DongA newspaper
- 101,226 sentence pairs
- Non-literally translated Korean-to-Chinese corpus
Experimental setting

- Validation: Using Kappa statistic

- Scenario:
  1. Kappa value between two skilled annotators (A1 and A2) who are very familiar with the annotation guidelines;
  2. Kappa values between each skilled annotator and a beginner (B) who was never involved in corpus annotation;
  3. Kappa values between each skilled annotator and the beginner acquainted (B_acquainted) with the annotation guidelines;
### Experimental result

<table>
<thead>
<tr>
<th></th>
<th>Kappa Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1 vs. A2</td>
<td>0.892</td>
</tr>
<tr>
<td>A1 vs. B</td>
<td>0.799</td>
</tr>
<tr>
<td>A2 vs. B</td>
<td>0.805</td>
</tr>
<tr>
<td>A1 vs. B_acquainted</td>
<td>0.858</td>
</tr>
<tr>
<td>A2 vs. B_acquainted</td>
<td>0.844</td>
</tr>
</tbody>
</table>

Kappa values b/w annotators

- >0.8: definite conclusion of the assessment scale
- >0.67 & <0.8: tentative conclusion
Conclusion

- Annotation guidelines for Chinese-Korean word alignment
  - Systematic comparison of verbal system by analyzing morpho-syntactic encodings
  - From viewpoint of grammatical category
    - Systematic and consistent annotation instructions

- Adopt Kappa value to validate the reliability of proposed guidelines
  - High reliability: 0.892
  - Produce consistent annotation results

- Applicable to other language pairs from linguistically distant language families
Thank You!
General Guidelines (Backup slide)

- Translated vs. Not translated
  - Correct vs. incorrect
  - Omissions in translation
- Minimum match vs. maximum match (completely-semantically matched link)
  - Phrasal correspondence
- Context-dependent translation
  - Anaphora (pronoun)
  - Demonstrative words
  - Contextual omissions and additions
- Glue approach
  - Glue extra words to its nearest head
General Guidelines (Backup slide)

Mark as many words as necessary on both the target and source side

- Mark as many words as you feel necessary to ensure a two-way equivalence
- Ex)
  - Don't do: une carte de paiement $\leftrightarrow$ a pay-card
  - Do: une carte de paiement $\leftrightarrow$ a pay-card

Mark as few words as possible on both the target and source side

- Mark the smallest number of words possible on each side, while preserving two-way equivalence
- Ex)
  - Don't do: une carte de paiement $\leftrightarrow$ a pay card
  - Do: une carte de paiement $\leftrightarrow$ a pay card
General Guidelines (Backup slide)

- **Minimum lexical unit size**
  - As few words as possible but as many words as necessary
  - The whole group must be considered as an indivisible lexical unit

- **Indivisibility rule**
  - The only valid elements in an alignment are single words and indivisible groups of words
  - A word cannot be aligned to only a part of a group

- **Absence of correspondence**
  - Omissions in translation
  - Not translated
### Relative orderings of verbal affixes in Korean (Lee, 1991)

<table>
<thead>
<tr>
<th>Order</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Verb Stem</td>
</tr>
<tr>
<td>2</td>
<td>Causative &amp; Passive</td>
</tr>
<tr>
<td>3</td>
<td>Honorific</td>
</tr>
<tr>
<td>4</td>
<td>Aspect&lt;br&gt;Tense&lt;br&gt;Modality</td>
</tr>
<tr>
<td>5</td>
<td>Negation</td>
</tr>
<tr>
<td>6</td>
<td>Modality - Evidential</td>
</tr>
<tr>
<td>7</td>
<td>Mood - Illocutionary Force</td>
</tr>
</tbody>
</table>
References(1)

Gispert, Gupta, Popovic, Lambert, Marino, Federico, Ney and Banchs (2006). Improving Statistical Word Alignments with Morpho-syntactic Transformations, FinTAL - 5th International Conference on Natural Language Processing (pp. 368--379), Turku, Finland.


References


Aspect system in Chinese

🌟 Aspectual particle & Adverb
- [cn] 他 (he) / 在 (now) / 写 (do) / 作业 (homework) /。
- [ko] 그 (he) + 는 숙제 (homework) + 를 하 (do) + 고 있으 + 다.
- [cn] 我 (I) / 曾 (already) / 去 (go) / 过 (Prt.) / 北京 (Beijing) /。
- [ko] 나 (I) + 는 북경 (Beijing) + 에 가 (go) 보 + 는 적 + 이 있 + 다.

🌟 Verb reduplication
- [cn] 我 (I) / 看 / 了 (Prt.) / 看 (read) / 报纸 (newspaper) /。
- [ko] 나 (I) + 는 신문 (newspaper) + 을 보 (read) + 았 + 다.

🌟 RVC
- [cn] 大家 (everybody) / 把 (Prep.) / 作业 (homework) / 交 (submit) / 上来 (RVC) /。
- [ko] 모두 (everybody) 숙제 (homework) + 를 내 (submit) 주 + 세 + 요.
- [cn] 写 (write) / 清楚 (clearly) / 你 (your) / 的 (Prt.) / 名字 (name) /。
- [ko] 당신 (your) + 의 이름 (name) + 을 복바로 (clearly) 적 (write) + 아 주 + 세 + 요.