Annotation Guidelines for Chinese-Korean Word Alignment

2008. 5. 28

POSTECH, R. Korea

Jin-Ji Li

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

Previous work (2)

- 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 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+ㄴ 적+이 있+다.

eojeol

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.

Content word

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.

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 clarity 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

Special Guidelines based on Korean Verbal System (2)

- 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+ㄴ 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*.

Corpus data

	Chinese	Korean
# of sentences	50	50
# of words	1,323	1,502
# of singletons	741	645
Avg. length	26.5	30.4

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

	Kappa Value
A1 vs. A2	0.892
A1 vs. B	0.799
A2 vs. B	0.805
A1 vs. B_acquainted	0.858
A2 vs. B_acquainted	0.844

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 ←→ a **pay-card**
      Do: une **carte de paiement** ←→ 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** ←→ a **pay card**
      Do: une **carte** de paiement ←→ 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

Backup Slide

Order	Type
1	Verb Stem
2	Causative & Passive
3	Honorific
4	Aspect Tense *Modality*
5	Negation
6	*Modality* - Evidential
7	*Mood* - Illocutionary Force

Relative orderings of verbal affixes in Korean (Lee, 1991)

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.

- Kruijff-Korbayova, Chvatalova, and Postolache (2005). Annotation Guidelines for Czech-English Word Alignment, Proceedings of LREC 2006 (pp. 1256--1261). Genova.

- Lambert, P., Gispert, DE A., Banchs, R., and Marino, B. J. (2005). Guidelines for Word Alignment Evaluation and Manual Alignment. Language Resources and Evaluation. 39(3), 267-285.

- Lee, H.-S. (1991). Tense, aspect, and modality: A discourse-pragmatic analysis of verbal affixes in Korean from a typological perspective, PhD thesis, Univ. of California, Los Angeles.

- Li, Charles N. and Thompson A. S. (1996). Mandarin Chinese: A functional reference grammar, University of California Press, USA.

References(2)

- Li, J.-J., Roh, J-E., Kim, D.-I., and Lee, J.-H. (2005). Contrastive Analysis and Feature Selection for Korean Modal Expression in Chinese-Korean Machine Translation System. International Journal of Computer Processing of Oriental Languages, 18(3), 227--242.
- Melamed, I. D. (1998). Annotation Style Guide for the Blinker Project. IRCS Technical Report #98-06, University of Pennsylvania.
- Merkel, M. (1999). Annotation Style Guide for the PLUG Link Annotator. Version 1.0, PLUG report, Magnus Merkel Linkping university.
- Och, F. and Ney, H. (2003). A systematic comparison of various statistical alignment models. Computational Linguistics, 29(1), 19--51.
- Xiao, R. Z. (2002). A corpus-based study of aspect in Mandarin Chinsese, PhD thesis, University of Lancaster.
- Veronis, J. and Langlais, P. (1999). Evaluation of parallel text alignment systems, In Parallel Text Processing (ed. J. Veronis), Kluwer.

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 있+다*.

  - [cn] 我(I)/*曾(already)/去(go)/过(Prt.)/*北京(Beijing)/。
  - [ko] 나(I)+는 북경(Beijing)+에 *가(go) 보+ㄴ 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 *똑바로(clearly) 적(write)+어 주+세+요*.