JSNACS: Adposition and Case Supersenses for Japanese Joshi

Tatsuya Aoyama, Chihiro Taguchi, Nathan Schneider
1Georgetown University, Department of Linguistics
2University of Notre Dame, Department of Computer Science and Engineering
{ta571, nathan.schneider}@georgetown.edu, ctaguchi@nd.edu

Introduction

- Semantic Network of Adposition and Case Supersenses (SNACS: Schneider et al., 2018) now applied to various typologically different languages.
- Japanese 部属 (joshi), which is often translated as particles, do not map to English prepositions in a straightforward manner.
- This aims at extending SNACS annotation to Japanese.

Construal Analysis (SceneRole=Function) (Hwang et al., 2017) and SNACS:

(1) It's a gift НеBENEFICIARY Tom.
(2) It's sad for EXPERIENCER-BENEFICIARY Tom.

Research Questions

1. How can we characterize the semantics of Japanese particles using the SNACS framework?
2. Can we use supersense distributions to compare the semantics of adpositions/case markers within and across languages?

Data & Annotation

Japanese translation of Le Petit Prince (The Little Prince), freely available online!

- The extracted texts were tokenized and UPOS and XPOS tags were manually corrected where relevant.
- Supersense was annotated manually by the author in consultation with the original SNACS guideline (Schneider et al., 2020). Korean SNACS guideline (Hwang et al., 2020), and the SNACS website (http://www.sposotion.org)

Many-to-many mapping between POS and XPOS: particle (binding) was included; particle (adverbial) was included when it can modify an NP: particle (nominal), particle (case): particle that maps to CCOP, and particle (case-final) were all excluded.

- Similar to Korean (Hwang et al., 2020), topic marker and case markers (ACC, NOM) are among the most frequent.
- Genitive marker が (no) and dative particle に (no) are among the most polysemous examples (below).

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Count</th>
<th>Type-Level Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>- A part of the 15 most common particles.</td>
<td>- 90</td>
<td></td>
</tr>
<tr>
<td>- Korean and Japanese (75) and English (97).</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Function Type</th>
<th>Count</th>
<th>Type-Level Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Its a gift НеBENEFICIARY Tom.</td>
<td>- 90</td>
<td></td>
</tr>
<tr>
<td>- It's sad for EXPERIENCER-BENEFICIARY Tom.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Construal analysis capturing the subtle differences in the usage of the quotation particle わ (wak). Contextual meaning is captured in the construal (SceneRole=Function).

Corpus Statistics

- 10 Chapters
- 619 Sentences
- 9,500 Tokens
- 1,810 Annotations

<table>
<thead>
<tr>
<th>Phrasal (Ch 1-5)</th>
<th>Adposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>143</td>
</tr>
<tr>
<td>2</td>
<td>143</td>
</tr>
<tr>
<td>3</td>
<td>143</td>
</tr>
</tbody>
</table>

Table 1: XPOS to USP mapping of Japanese particles. The independent annotation is a combination of the XPOS and USP whose inclusion is lexicon-dependent.

Table 2: Descriptive statistics of the corpus. Left columns represent count data, and right columns represent type-level frequencies.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Count</th>
<th>Type-Level Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>- The construction of the corpus.</td>
<td>- 90</td>
<td></td>
</tr>
<tr>
<td>- Japanese translation of Le Petit Prince (The Little Prince), freely available online!</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Diagram:

- Polysemy of the particle に (ni). Also many stacking particles is very common in Japanese.

Experiments

CWE-based metric

SS-based metric

Figure 1: Frequency breakdown by word type. Figure 2: Number of distinct construal pairs for the 15 most common particles. 15 most polysemous particles.

Results

- Many-to-many mapping between POS and XPOS: particle (binding) was included; particle (adverbial) was included when it can modify an NP: particle (nominal), particle (case): particle that maps to CCOP, and particle (case-final) were all excluded.
- Similar to Korean (Hwang et al., 2020), topic marker and case markers (ACC, NOM) are among the most frequent.
- Genitive marker が (no) and dative particle に (no) are among the most polysemous examples (below).

Construal analysis capturing the subtle differences in the usage of the quotation particle わ (wak). Contextual meaning is captured in the construal (SceneRole=Function).

Evaluation

- SS-based metric seems to be capturing what is not captured by the CWE-based metric (fewer gray-out cells). This is expected, given that SS is manually annotated and CWE is learned in a self-supervised manner for general NLP purposes.

References