Ambiguity Resolution for Vt-N Structures in Chinese

Yu-Ming Hsieh1,2  Jason S. Chang2  Keh-Jiann Chen1
1 Institute of Information Science, Academia Sinica, Taipei
2 Department of Computer Science, National Tsing-Hua University, Taiwan
morris@iis.sinica.edu.tw  jason.jschang@gmail.com  kchen@iis.sinica.edu.tw

Problem Statement and Analysis

- In Chinese, the structure of a transitive verb (Vt) immediately followed by a noun (N) may form a verb phrase (VP), a noun phrase (NP), or there may not be a dependent relation, as shown below.

- Parsers generally prefer to the VP reading due to the statistical majority.

Table 1. Statistical data from the Sinica Treebank.

<table>
<thead>
<tr>
<th>VP</th>
<th>NP</th>
<th>Other Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>58%</td>
<td>16%</td>
<td>26%</td>
</tr>
</tbody>
</table>

- Unlike inflectional languages, Chinese verbs modify nouns without morphological inflection, e.g., 結婚/farming 池/pond.
- Linguistically motivated features to build a Vt-N classifier include lexical words, semantic knowledge, the morphological structure of verbs, neighboring parts-of-speech, and the syllabic length of words.

- Analysis of NP(Vt-N) structures in the Sinica Treebank reveals the following four types of semantic structures.

<table>
<thead>
<tr>
<th>Types</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telic(Vt) + Host(N)</td>
<td>研究工具/ research tool</td>
</tr>
<tr>
<td>Host-Event(Vt) + Attribute(N)</td>
<td>冷藏箱/attacking strategy</td>
</tr>
<tr>
<td>Agentive(Vt) + Host(N)</td>
<td>彈射/ shouted sound</td>
</tr>
<tr>
<td>Apposition(Vt) + Affair(N)</td>
<td>動靜/ fried chicken</td>
</tr>
</tbody>
</table>

- Difficulties:
  - Feature selection. Decision on which features to adopt and their combination is a difficult task in classification.
  - Unknown word issue. Unknown word processing has technical problems that affect the prediction of the semantic types and morph-structures of unknown words.
  - Data sparseness problem. Due to the limited size of the current Treebank, we should mine useful information from all available resources.

Proposed Models

- We integrate the Vt-N models into the PCFG parser. The formula of the integrated structural evaluation model is as follows:

\[
\text{Score}(T, S) = \sum \left( \phi_1 \times R_P + \phi_2 \times V_{NP} \right)
\]

Table 2. Semantic types of NP(Vt-N) and examples.

Table 3. The features labels of Vt-N pair in Figure 1.

Table 4. The results of using different feature combinations. P1(%) is the 10-fold cross validation accuracy of the training data; P2(%) is the accuracy of the test data.

Table 5. Experiment results of classifiers with different training data.

Evaluation of the Vt-N Classifier

- The results of Table 4 show that contextual information (M2) and lexical feature (M3) are the most important feature used to disambiguate VP, NP and independent structures. Results also demonstrate the benefits of using the semantic type (M4), verb morph-structure and noun length features (M5).

Knowledge from Large-scale Unlabeled Data

- We applied the data selection method (distance=1, with an intraverbive verb followed by an object noun) and data correction method to learn more useful knowledge.
- The results in Table 5 show that the proposed methods can improve the accuracy.

Table 6. The performance of the PCFG parser with and without model M5. The BF (bracketed f-score) is the parsing performance metric.