Statistical Morphological Tagging and Parsing of Korean with an LTAG Grammar

Anoop Sarkar
University of Pennsylvania
anoop@cis.upenn.edu

and

Chung-hye Han
Simon Fraser University
chunghye@sfu.ca

TAG+ 6, May 2002 — Venice, Italy
Overview

- Introduction to Supervised Statistical Parsing with LTAG
- LTAG grammar extracted from the Penn Korean Treebank
- Morphological Tagging: Motivation and Experiments
- Statistical parsing of Korean using a Morphological Tagger
Parsings as a machine learning problem

- \( S = \) a sentence
  \( T = \) a parse tree
  A statistical parsing model defines \( P(T \mid S) \)

- Find best parse: \( \arg\max_T P(T \mid S) \)

- \( P(T \mid S) = \frac{P(T, S)}{P(S)} = P(T, S) \)

- Best parse: \( \arg\max_T P(T, S) \)

- e.g. for PCFGs: \( P(T, S) = \prod_{i=1}^{n} P(\text{RHS}_i \mid \text{LHS}_i) \)
Parsing as a machine learning problem

- Training data for English: the Penn WSJ Treebank (Marcus et al. 1993)

- Convert Treebank into LTAG derivations using LexTract (Xia 2001)

- Train statistical LTAG parser from these events

- Evaluate accuracy on test data

- A standard evaluation:
  Train on 40,000 sentences
  Test on 2,300 sentences
Parsing as a machine learning problem

- Training data for Korean: the Penn Korean Treebank (Han et al. 2002)

- Train statistical morphological tagger and statistical LTAG parser

- Evaluate accuracy on test data

- Our evaluation:
  Train on 4,653 sentences (49,473 words)
  Test on 425 sentences (3,717 words)
Statistical Parsing with Tree Adjoining Grammars

- Substitution: \( \sum_\alpha P_s(t, \eta \to \alpha) = 1 \)

- Adjunction: \( P_a(t, \eta \to NA) + \sum_\beta P_a(t, \eta \to \beta) = 1 \)

- Multiple adjunctions at a node (Schabes and Shieber 1994):
  \[
  P_{la}(\tau, \eta \to NA_l) + \sum_{\tau'} P_{la}(\tau, \eta \to \tau') = 1
  \]
  \[
  P_{ra}(\tau, \eta \to NA_r) + \sum_{\tau'} P_{ra}(\tau, \eta \to \tau') = 1
  \]
Statistical Parsing with Tree Adjoining Grammars

- Start of a derivation: $\sum_{\alpha} P_i(\alpha) = 1$

- Probability of a derivation:

  $$Pr(\mathcal{D}, w_0 \ldots w_n) =$$

  $$P_i(\alpha, w_i) \times \prod_{p} P_s(\tau, \eta, w \rightarrow \alpha, w') \times$$

  $$\prod_{q} P_a(\tau, \eta, w \rightarrow \beta, w') \times \prod_{r} P_a(\tau, \eta, w \rightarrow_{\text{NA}})$$
Overview

- Introduction to Supervised Statistical Parsing with LTAG
- LTAG grammar extracted from the Penn Korean Treebank
- Morphological Tagging: Motivation and Experiments
- Statistical parsing of Korean using a Morphological Tagger
I reported the overvation items.'
LTAG Grammar and Derivation Tree using LexTract (Xia 2001)
Korean Treebank

(S (NP-OBJ-1 권한authority/NNC+을acc/PCA)
 (S (NP-SBJ 누구who/NPN+가nom/PCA)
  (VP (VP (NP-OBJ *T*-1)
       가지have/VV+고aux/EAU)
       있be/VX+지int/EFN)))
/?SFN)

→ authority-Acc who-Nom have-AuxConnective be-Int

→ ‘Who has the authority?’
LTAG Grammar for Korean using LexTract

[Diagram of a syntactic tree structure for a sentence in Korean, with nodes labeled for parts of speech (NP, NPP, NNC) and words (누구, who, 권한, authority, VP, VX) along with English translations where applicable.]
LTAG Derivation Tree

α가지 have

α 누구 who{NP}  α 권한 authority{NP}  β 있 be{VP}
Overview

- Introduction to Supervised Statistical Parsing with LTAG
- LTAG grammar extracted from the Penn Korean Treebank
- Morphological Tagging: Motivation and Experiments
- Statistical parsing of Korean using a Morphological Tagger
Motivation for Morphological Tagging

- Each substitution, adjunction is a relation between a pair of words

- Korean is an agglutinative language with a very productive inflectional system

- A fully inflected word seen in the training data will rarely occur in the unseen (test) data

- Sparse data problem is much worse than in English: the part-of-speech tags for inflected word forms are complex and can be novel in unseen data
Motivation for Morphological Tagging

- The morphological tagger provides lemma splitting plus part-of-speech tagging

- Instead of multiplying ambiguity in the parser, we choose to implement a statistical morphological tagger (provides a single-best analysis of the input sentence)

- Both lemma splitting and tagging are trained using the Penn Korean Treebank (same training/test split as in the parser)

- Lexical stem and suffix information as well as part-of-speech information from the morphological tagger is used in the statistical parser
Example input and output from the morphological tagging phase

Input: 제가 관측 사항을 보고했습니다.

Output: 제/NPN+가/PCA관측/NNC사항/NNC+을/PCA보고/VV+했/EPF+했습니다/EFN/EFN/EFN

The part-of-speech tags for inflected word forms are complex and can be novel in unseen data.
Evaluation of the Morphological Analyzer/Tagger

<table>
<thead>
<tr>
<th></th>
<th>unseen test data (3,717 words) precision/recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Treebank trained</strong></td>
<td>95.78/95.39</td>
</tr>
<tr>
<td><strong>Off-the-Shelf</strong></td>
<td>29.42/31.25</td>
</tr>
</tbody>
</table>
Overview

- Introduction to Supervised Statistical Parsing with LTAG

- LTAG grammar extracted from the Penn Korean Treebank

- Morphological Tagging: Motivation and Experiments

- Statistical parsing of Korean using a Morphological Tagger
Morphological Analysis Incorporated into the Statistical Model

In each probability model used in the parser where inflected word forms are used we incorporate the output of the morph tagger as a backoff level.

For example, take the probability model for adjunction:

\[
P_a(t, \eta \rightarrow t') = \Pr(t', p', w' | \eta, t, w, p) \tag{1}
\]

\[
= \Pr(t' | \eta, t, w, p) \times \Pr(p' | t', \eta, t, w, p) \times \Pr(w' | p', t', \eta, t, w, p) \tag{2}
\]
Morphological Analysis Incorporated into the Statistical Model

- $e_1$ = lexicalized model using stems;
  $e_2$ = part-of-speech tags from the morphological tagger:

  $P_{re_1} = Pr(t' \mid \eta, t, w, p)$
  $P_{re_2} = Pr(t' \mid \eta, t, p)$

- The backoff model is computed as follows:

  $\lambda(c) \times P_{re_1} + (1 - \lambda(c)) \times P_{re_2}$
Parsing Experiment: Training and Test Data

- Training data for Korean: the Penn Korean Treebank (Han et al. 2002)

- Train statistical morphological tagger and statistical LTAG parser

- Evaluate accuracy on test data

- Our evaluation:
  Train on 4,653 sentences (49,473 words)
  Test on 425 sentences (3,717 words)
Example derivation reported by the statistical parser

<table>
<thead>
<tr>
<th>Index</th>
<th>Word</th>
<th>Gloss</th>
<th>POS tag (morph)</th>
<th>Elem Tree</th>
<th>Node Address</th>
<th>Subst/Adjoin</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>모든</td>
<td>every</td>
<td>DAN</td>
<td>$\beta NP^*=1$</td>
<td>root</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>호출</td>
<td>call</td>
<td>NNC</td>
<td>$\beta NP^*=1$</td>
<td>root</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>대호+는</td>
<td>sign-topic</td>
<td>NNC+PAU</td>
<td>$\alpha NP=0$</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>매일</td>
<td>everyday</td>
<td>ADV</td>
<td>$\beta VP^*=25$</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>24</td>
<td></td>
<td>NNU</td>
<td>$\beta NP^*=1$</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>시+에</td>
<td>hour-at</td>
<td>NNX+PAD</td>
<td>$\beta VP^*=17$</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>바뀌+게</td>
<td>switch-aux</td>
<td>VV+ECS</td>
<td>$\alpha S-NPs=23$</td>
<td>-</td>
<td>TOP</td>
</tr>
<tr>
<td>7</td>
<td>되+지요</td>
<td>be-decl</td>
<td>VX+EFN</td>
<td>$\beta VP^*=13$</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>.</td>
<td></td>
<td>SFN</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Parser evaluation results

<table>
<thead>
<tr>
<th></th>
<th>On training data</th>
<th>On unseen test data (425 sents)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Current Work</strong></td>
<td><strong>97.58</strong></td>
<td><strong>75.7</strong></td>
</tr>
<tr>
<td>(Yoon et al. 1997)</td>
<td>–</td>
<td>52.29/51.95 P/R</td>
</tr>
</tbody>
</table>
Summary

- First LTAG-based parsing system for Korean.

- LTAG-based statistical parsing is feasible for a language with free word order and complex morphology.

- Our system has been successfully incorporated into a Korean/English machine translation system as source language analysis component.
Summary

- The tagger/analyzer obtained the correctly disambiguated morphological analysis with 95.78/95.39%.

- The statistical parser obtained a dependency accuracy of 75.7%.

- These performance results are better than an existing off-the-shelf Korean morphological analyzer and parser run on the same data.
Grazie ...
Experiments with and without the Morphological Tagger

- Even the part-of-speech tags are often unseen in the test data.

- When we lexicalize trees we use words from the training data and for unknown words the output of a part-of-speech tagger.

- Without a morphological tagger the lexicalization step becomes infeasible. (We can annotate the Treebank with a new smaller tagset, but the number of trees for unknown words explodes.)

- Thus, we could not easily compare parsing with and without a morphological tagger.