Increasing the reliability of a part-of-speech tagging tool for use with learner language

Sylvie Thouësny
Toward more accurate pos-tags: Presentation overview

- Creation and annotation of a learner language corpus
- Tagging improvements: procedure
- Tagging accuracy: results
Toward more accurate pos-tags:

Corpus

- 17 learners of French
- 67 texts (email, reflective writing, wiki)
- 18,039 tokens in total
- 2,102 grammatical and lexical errors in total
Learners’ errors annotated corpus

- learners’ errors annotated manually

- 4 error type classification
  - selection
  - syntactic
  - morphosyntactic
  - spelling

- storage
Part-of-speech annotated corpus

- TreeTagger - TC project - University of Stuttgart
tagging accuracy of 96.34% (Schmidt 1994)
www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

- 10,108 tokens were manually and automatically part-of-speech tagged
8,417 words
1,440 error types
learner’s error rate of 17.11%

- storage
Toward more accurate pos-tags:

Presentation overview

- Creation and annotation of a learner language corpus
- Tagging improvements: procedure
- Tagging accuracy: results
Toward more accurate pos-tags:

Procedure

- Identification of unknown lemmas
- Rule-based part-of-speech tagging
- Cross-reference between pos and error tags
Unknown lemmas

- upper case to lower case conversion

- unknown lemmas identified: N=97

<table>
<thead>
<tr>
<th>input</th>
<th>Cette</th>
<th>semaine</th>
<th>en</th>
<th>Allemagne</th>
</tr>
</thead>
<tbody>
<tr>
<td>original input</td>
<td>*NAM</td>
<td>NOM</td>
<td>PUN</td>
<td>NAM</td>
</tr>
<tr>
<td></td>
<td>&lt;unknown&gt;</td>
<td>&lt;semaine&gt;</td>
<td>&lt;,&gt;</td>
<td>&lt;en&gt;</td>
</tr>
<tr>
<td></td>
<td>pos:</td>
<td></td>
<td></td>
<td>&lt;Allemagne&gt;</td>
</tr>
<tr>
<td></td>
<td>lemma:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>input converted into lowercases</td>
<td>DET:DEM</td>
<td>NOM</td>
<td>PUN</td>
<td>*NOM</td>
</tr>
<tr>
<td></td>
<td>&lt;ce&gt;</td>
<td>&lt;semaine&gt;</td>
<td>&lt;,&gt;</td>
<td>&lt;unknown&gt;</td>
</tr>
<tr>
<td></td>
<td>pos:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>lemma:</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Token: N=10,108
Unknown lemmas in total: N=589 (5.82%)
Treetagger’s correct guesses: N=233 (39.55%)

Before: 9 correct, 88 incorrect
After: 82 correct, 15 incorrect
Unknown lemmas

- word bank

<table>
<thead>
<tr>
<th>Word</th>
<th>Part of Speech</th>
<th>Correct Tag</th>
<th>Incorrect Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>instruction</td>
<td>NOM</td>
<td>unknown instruction</td>
<td>NOM</td>
</tr>
<tr>
<td>vetements</td>
<td>NOM</td>
<td>unknown vétement</td>
<td>NOM</td>
</tr>
<tr>
<td>comfortable</td>
<td>ADJ</td>
<td>unknown comfortable</td>
<td>ADJ</td>
</tr>
<tr>
<td>peut-être</td>
<td>VER:inf</td>
<td>unknown peut-être</td>
<td>ADV</td>
</tr>
<tr>
<td>parceque</td>
<td>VER:subp</td>
<td>unknown parce que</td>
<td>KON</td>
</tr>
<tr>
<td>français</td>
<td>VER:imf</td>
<td>unknown</td>
<td>ADJ</td>
</tr>
</tbody>
</table>

Unknown lemmas identified: N=488

Token: N=10,108
Unknown lemmas in total: N=589 (5.82%)
Treetagger’s correct guesses: N=233 (39.55%)
Rule-based approach

- 44 rules presented using First-order predicate logic (FOPL) framework, “the lingua franca of knowledge representation” (Lager & Nivre, 2001)

- One standard syntax to represent the knowledge is by using rules that are expressed in the form of if...then statements (Coppin, 2004).
Rule-based grammar

\[ \forall p (\text{Position}(p) \land \text{Tag}(p, \ldots) \land \text{Tag}(p-1, \ldots)) \rightarrow \text{Tag}(p, \text{Nom}) \]

For all \( p \), if \( p \) is a valid word position in the input text, and if the tag corresponding to the word at position \( p \) in the text is equal to \( \text{Ver:conj} \), and if the tag corresponding to the word prior to the word at position \( p \), i.e. at position \( p-1 \), matches any determiners \( \text{Det} \) then the tag at position \( p \) corresponding to the word at position \( p \) is replaced by the tag \( \text{Nom} \).
for each token pos-tagged{
  check whether this token has been marked as incorrect
  if marked as incorrect{
    look at the error tag
    adjust the pos tag in accord with the error type
  }
  else if not marked as incorrect{
    leave the pos as such
  }
}
learner input

[...] l'article *a du* être [...]  
...the article must have been...

- l'[DET]
  - word tagged as correct
    - l'[DET]

- article[Noun]
  - word tagged as correct
    - article[Noun]

- a[VER:conj]
  - word tagged as [mo_fo_co]
    - a[VER:conj]

- du[DET]
  - word tagged as [mo_fo_co]
    - du[VER:pper]

error type and pos-tag not related

update
Toward more accurate pos-tags:
Presentation overview

- Creation and annotation of a learner language corpus
- Tagging improvements: procedure
- Tagging accuracy: results
Toward more accurate pos-tags:

Results

- Krippendorff’s (1970) alpha coefficient
- Cohen’s (1960) kappa coefficient
- Recall, precision and F-measure
Krippendorff’s alpha coefficient

- N=10,108 tokens

- machine vs. human before improvements

Run MATRIX procedure:
Krippendorff's Alpha Reliability Estimate

<table>
<thead>
<tr>
<th></th>
<th>Alpha</th>
<th>LL95%CI</th>
<th>UL95%CI</th>
<th>Units</th>
<th>Observrs</th>
<th>Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>0.8962</td>
<td>0.8898</td>
<td>0.9025</td>
<td>10108.0000</td>
<td>2.0000</td>
<td>10108.0000</td>
</tr>
</tbody>
</table>

- machine vs. human after improvements

Run MATRIX procedure:
Krippendorff's Alpha Reliability Estimate

<table>
<thead>
<tr>
<th></th>
<th>Alpha</th>
<th>LL95%CI</th>
<th>UL95%CI</th>
<th>Units</th>
<th>Observrs</th>
<th>Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>0.9760</td>
<td>0.9729</td>
<td>0.9792</td>
<td>10108.0000</td>
<td>2.0000</td>
<td>10108.0000</td>
</tr>
</tbody>
</table>

For an SPSS version of the KALPHA macro, visit Andrew F. Hayes' website http://www.comm.ohio-state.edu/ahayes/
Cohen’s kappa coefficient

### Machine vs. Human before improvements

<table>
<thead>
<tr>
<th>Measure of Agreement</th>
<th>Value</th>
<th>Asymp. Std. Error</th>
<th>Approx. T</th>
<th>Approx. Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kappa</td>
<td>.896</td>
<td>.003</td>
<td>306.858</td>
<td>.000</td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>10108</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

### Machine vs. Human after improvements

<table>
<thead>
<tr>
<th>Measure of Agreement</th>
<th>Value</th>
<th>Asymp. Std. Error</th>
<th>Approx. T</th>
<th>Approx. Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kappa</td>
<td>.976</td>
<td>.002</td>
<td>326.262</td>
<td>.000</td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>10108</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.
Recall, precision and F-measure

\[
\text{recall} = \frac{\text{number of correct part-of-speech tags in Treetagger}}{\text{number of correct part-of-speech tags in gold data}}
\]

\[
\text{precision} = \frac{\text{number of correct part-of-speech tags in Treetagger}}{\text{number of total part-of-speech tags in Treetagger}}
\]

\[
F\text{-Measure} = \frac{(\beta + 1.0)(\text{Precision})(\text{Recall})}{\beta(\text{Precision}) + (\text{Recall})}
\]

Widely used metrics in part-of-speech tagging (Voutilainen, 2003)
Recall, precision and F-measure

- N=10,108 tokens

### Table 1: TreeTagger vs. Human

<table>
<thead>
<tr>
<th></th>
<th>precision</th>
<th>recall</th>
<th>F-Measure $\beta=1$</th>
<th>F-Measure $\beta=2$</th>
<th>F-Measure $\beta=0.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>78.43%</td>
<td>77.67%</td>
<td>78.05%</td>
<td>77.92%</td>
<td>78.18%</td>
</tr>
</tbody>
</table>

### Table 2: TreeTagger++ vs. Human

<table>
<thead>
<tr>
<th></th>
<th>precision</th>
<th>recall</th>
<th>F-Measure $\beta=1$</th>
<th>F-Measure $\beta=2$</th>
<th>F-Measure $\beta=0.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>97.21%</td>
<td>96.01%</td>
<td>96.61%</td>
<td>96.41%</td>
<td>96.81%</td>
</tr>
</tbody>
</table>
Conclusion

- **alpha**
  - before improvements: 97.60
  - after improvements: 89.62
- **kappa**
  - before improvements: 97.60
  - after improvements: 89.60
- **precision**
  - before improvements: 97.21
  - after improvements: 78.43
- **recall**
  - before improvements: 96.01
  - after improvements: 77.67
- **F-measure**
  - before improvements: 96.61
  - after improvements: 78.05