Variation

Variation can be caused by:

1. **ambiguity** (different contexts require various tags):
   - over (IN, JJ, NN, RB or RP)

2. **erroneous tagging** (within the same context):
   - a. Erroneous part-of-speech tagging:
     - ... a year ago.
     - ... DT NN RB.
     - ... DT NN IN.
   - b. Erroneous dependency structure:
     - ... Hong Kong dollar exposure.

It is not sufficient to look at a single token to determine if an annotation is wrong (Example 1). Nevertheless variation found within the same context is more likely to be erroneous (Example 2a and 2b). Our interface implements the error mining algorithms introduced by Dickinson and Meurers [1] for pos-tags and Boyd et al. [2] for dependency structures.

Summary

- Interactive graphical interface integrated in ICARUS [3]
- Error mining for part-of-speech and dependency structures
- Supports various levels of user expertise
- Java-based, platform independent, requires no installation

The latest version can be found here:
http://www.ims.uni-stuttgart.de/data/icarus.html

Future Plans: Providing the capability to manual annotate and correct erroneous tags of a given corpus.

Algorithm

The error mining algorithm by Dickinson and Meurers [1]:

**Step 1**: Store tokens with their occurring tag(s), only compute n-grams for tokens with at least two different tags (nucleus)

**Step 2**: Increase the context for all nuclei (include adjacent tokens). Stop when either the context can’t be extended any further or no variation nucleus is left (all instances have the same label)

Exploration Views

**Figure 1**: Variation N-Gram View
- List of all n-grams
- Nucleus information displayed below
- Various n-gram filter (min/max, sort by length, search specific string)

**Figure 2**: Label Distribution View
- Tag distribution over word forms
- Filter tags by string
- Bar chart with frequencies (total / selected token)
- Export bar chart (*.png, *.jpg)

**Figure 3**: Corpus Instances
Shows the corresponding sentences of a selected n-gram or token-tag combination, that have been clicked before in the Variation N-Gram View (Figure 1) or Label Distribution View (Figure 2), including the proper highlighting.

Exploration Views

**Figure 1**: Variation N-Gram View
- List of all n-grams
- Nucleus information displayed below
- Various n-gram filter (min/max, sort by length, search specific string)

**Figure 2**: Label Distribution View
- Tag distribution over word forms
- Filter tags by string
- Bar chart with frequencies (total / selected token)
- Export bar chart (*.png, *.jpg)

**Figure 3**: Corpus Instances
Shows the corresponding sentences of a selected n-gram or token-tag combination, that have been clicked before in the Variation N-Gram View (Figure 1) or Label Distribution View (Figure 2), including the proper highlighting.

References


This work has been funded by the Deutsche Forschungsgemeinschaft (DFG) via Project D8 of the SFB 732 and by the Bundesministerium für Bildung und Forschung (BMBF) via project No. 01UG1120F, CLARIN-D center Stuttgart.