A Graphical Interface for Automatic Error Mining in Corpora

Gregor Thiele, Wolfgang Seeker, Markus Gärtner, Anders Björkelund and Jonas Kuhn
Institut für Maschinelle Sprachverarbeitung, University of Stuttgart
firstname.lastname@ims.uni-stuttgart.de

Variation

Variation can be caused by:

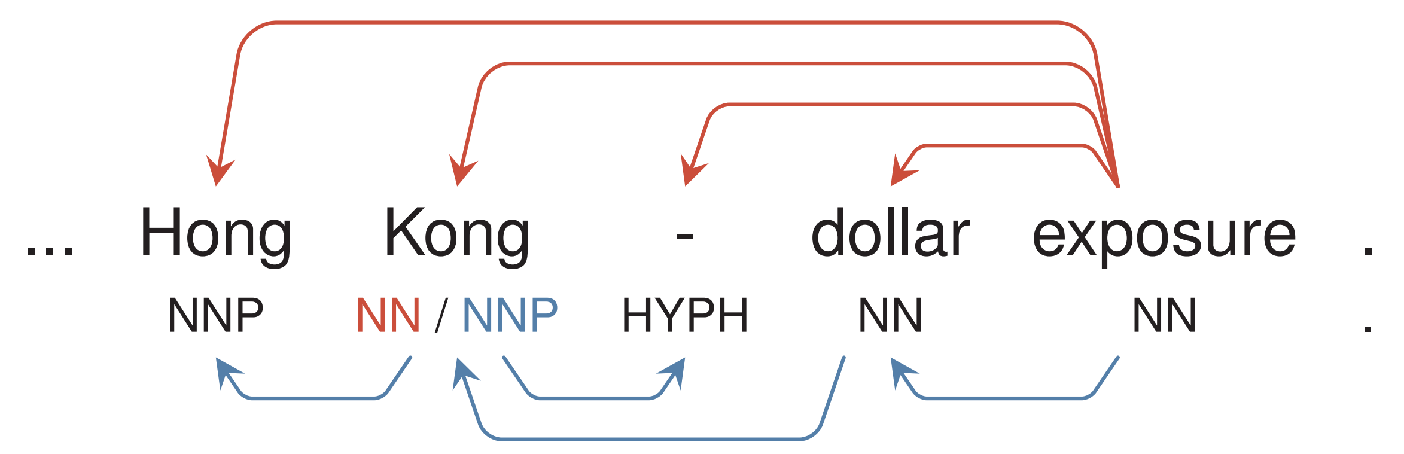
(1) **ambiguity** (different contexts require various tags):
over {IN, JJ, NN, RB or RP}

(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:



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

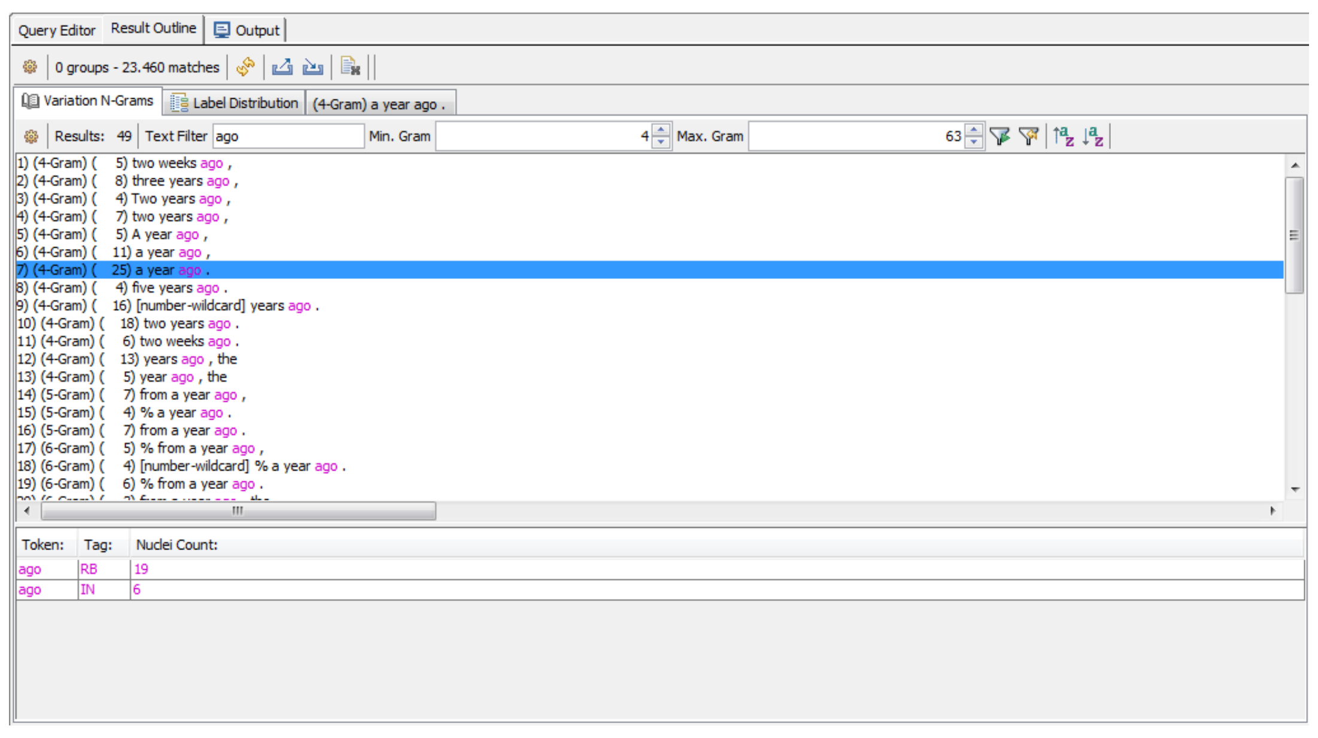
The error mining algorithm by Dickinson and Meurers [1]:

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)

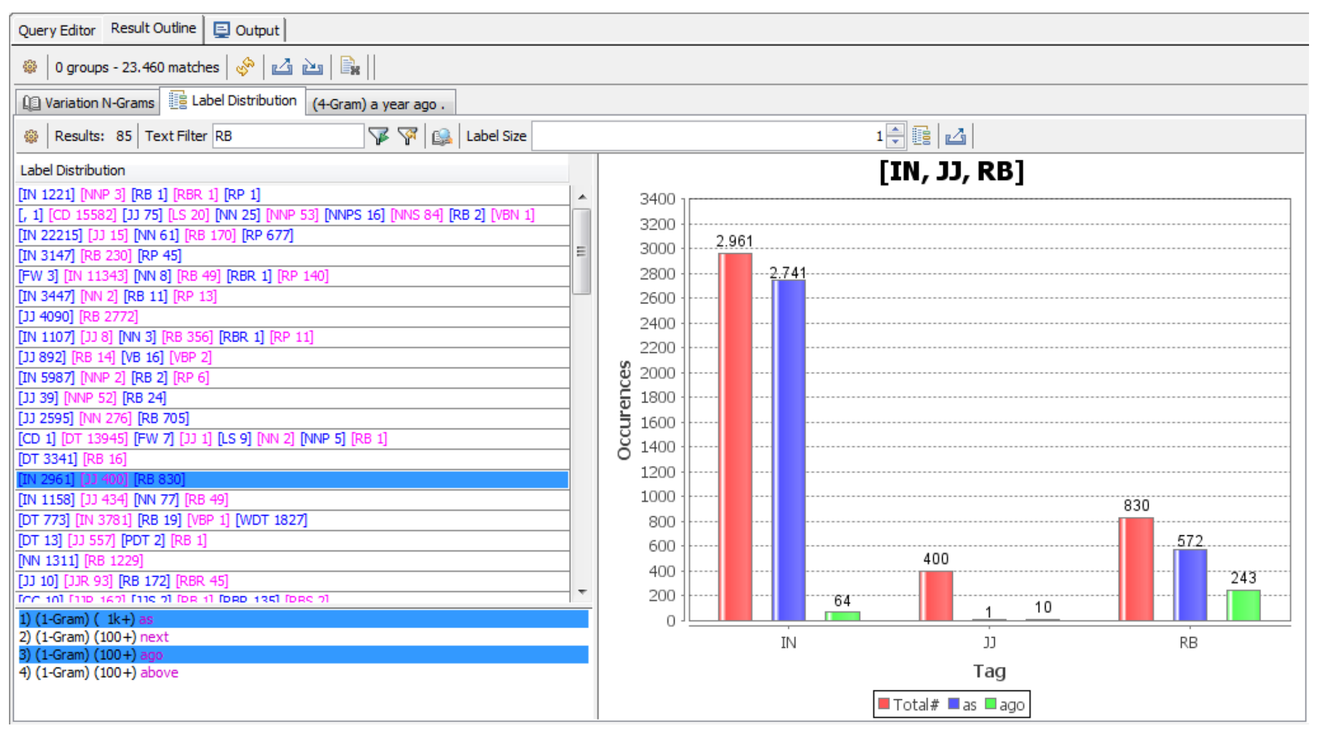
Exploration Views

Figure 1: Variation N-Gram View



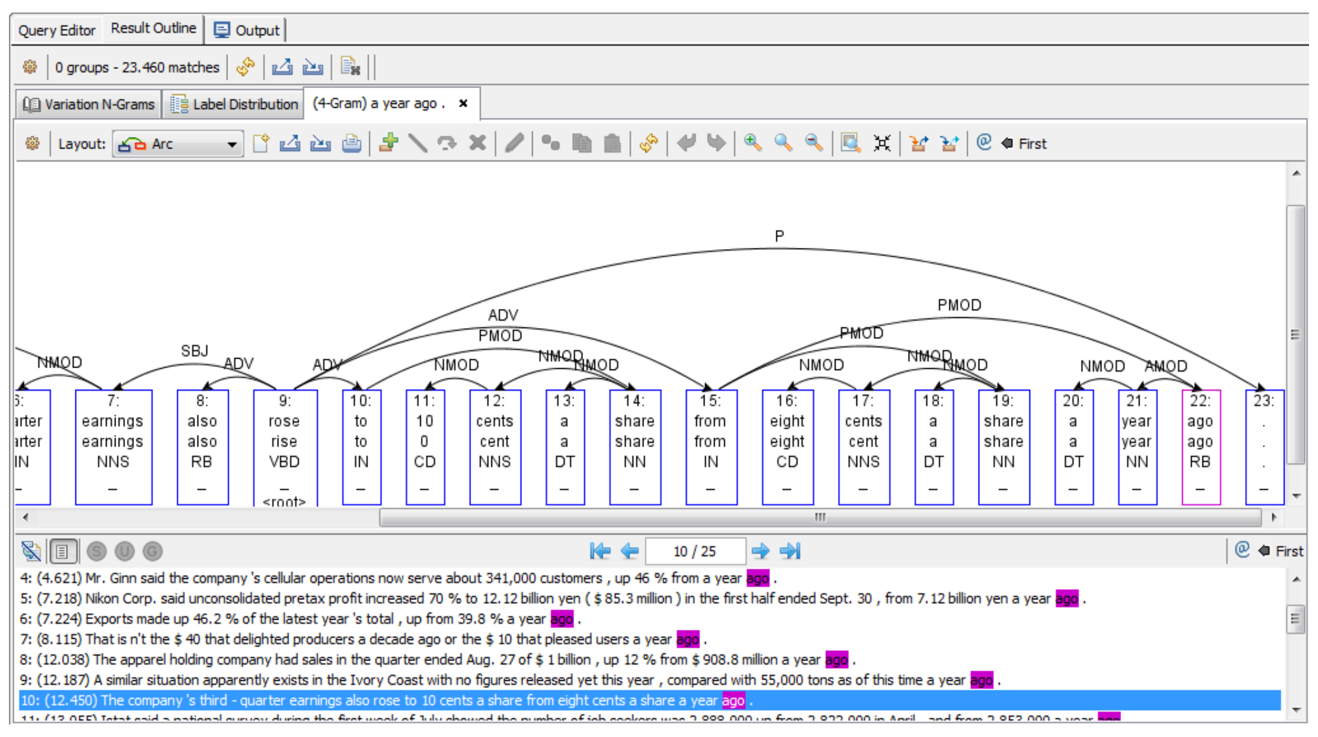
- List of all n-grams
- Nucleus information displayed below
- Various n-gram filter (min/max, sort by length, search specific string)

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.

Click an n-gram, word form or bar chart item in one of the two result views (see Figure 1 and 2) to show all sentences that contain the particular combination of word form/tag in the corpus (see Figure 3).

References

[1] Markus Dickinson and W. Detmar Meurers. Detecting Errors in Part-of-Speech Annotation. In *Proc. of EACL 2003*, pages 107–114, Budapest, Hungary, 2003.

[2] Adriane Boyd, Markus Dickinson, and W. Detmar Meurers. On Detecting Errors in Dependency Treebanks. *Research on Language and Computation*, 6(2):113–137, 2008.

[3] Markus Gärtner, Gregor Thiele, Wolfgang Seeker, Anders Björkelund, and Jonas Kuhn. ICARUS – An Extensible Graphical Search Tool for Dependency Treebanks. In *Proc. of ACL: System Demonstrations*, pages 55–60, Sofia, Bulgaria, August 2013. ACL.

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.

SPONSORED BY THE Federal Ministry of Education and Research
DFG
CLARIN-D