Low Inter-Annotator Agreement
=
An Ill-Defined Problem?

Vojtěch Kovář, Pavel Rychlý, Miloš Jakubíček

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno

{xkovar3,pary,jak}@fi.muni.cz

Supported by the Czech-Norwegian Research Programme within the HaBiT Project 7F14047.

Motivation

- Current perception
  - well defined task ⇒ high inter-annotator agreement
- Problematic agreement in NLP tasks
  - English tagging: 97%
  - English parsing: < 95% (Sampson)
  - Czech parsing: < 90% (PDT research)
  - collocation extraction, topic detection, term extraction... ?

Results of the discrepancy

- Try to claim high agreements
    - extremely extensive manuals
    - but we want to find out how people understand language, without any manuals
    - agreement numbers not published
- We need to be able to work with tasks with low IAA
    - testing
    - training

What we want

- Imagine a tool solving a low-IAA binary classification task
- We want it to
    - give positive answer where all annotators agreed on positive
    - same for negative
    - any answer is good if the annotators did not agree
- N-ary classification task
    - by disagreement, we may want to check if the tool agreed at least with one human annotator

Proposal

$$precision = \frac{\#true\_positives}{\#true\_positives + \#false\_positives}$$

$$recall = \frac{\#true\_positives}{\#true\_positives + \#false\_negatives}$$

But only take into account 100% agreements among people, ignore the other cases

Random agreements

- Can be minimized by adding more annotators
  - binary 50:50 task
  - 7 annotators
  - <1% random agreements
- Unevenly distributed tasks
  - large number of annotators needed to minimize random agreements
  - topic for discussion

Conclusions

- We need to work on low-IAA tasks
- We have introduced a straightforward methodology