Crowdsourcing for Large Scale Texts Annotation

Jozef Harinek Supervisor: Marián Šimko

Problem

→ Large amount of information on the web in raw texts
→ Better processing - annotated corpus (syntactic annotations)
→ Processing large amount of data - crowdsourcing

→ Students need to perform syntactic annotations at school anyway
→ Large crowd (209 103 + 76 711 students)

Data – 3 datasets

→ DATA-TEST: school book
→ DATA-SNK: Slovak national corpus (beletry, Wikipedia)
→ DATA-NEWS: publicist style (dennikN)

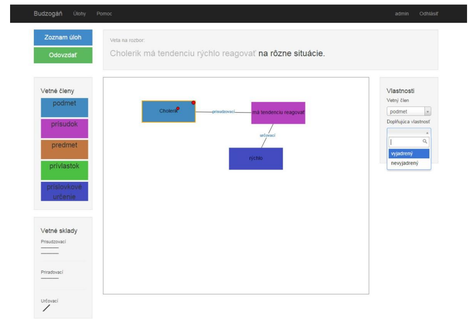
→ Slovak national corpus
  → very detailed
  → for linguists
  → 2 years of annotating (2005 - 2007)
  → many years of verification (up till 2014)
  → still not released

→ Annotations for texts processing
  → smaller granularity is sufficient for some tasks
  → we can obtain sufficient granularity

Method description

→ Huge crowd of potential annotators, that need to perform these annotations anyway

→ Motivation – grades, knowledge, need of practice
→ Quality control – multiple annotations per sentence
→ Data aggregation – gradually building up annotated dataset
→ Human skill – knowledge of syntactic analysis
→ Cardinality – many-to-many



Experiment

→ Qualitative experiment
  → positive feedback from teachers and students

→ Annotations collection
  → 65 participants
  → 488 collected annotations so far
  → experiments in progress (annotations and participants number is growing)

→ Several evaluation criteria:
  → sentence types
  → sentence length
  → students
    → primary/high school
    → grade evaluation

	Correct solution
Tokens	92,68 %
Relations	90 %
Sentences – tokens	85,71 %
Sentences – tokens + relations	71,43 %

Metrics

LAS (Labeled Attachment Score) – percentage of correctly assigned relations with correctly assigned relation types

	LAS
Relations	90 %

$$element = max\{\frac{|opt_1|}{|evaluators|}, \frac{|opt_2|}{|evaluators|}, ... \frac{|opt_n|}{|evaluators|}\}$$

Conclusion

→ Identify the power of crowd
  → elementary school students - basic annotations
  → high school students - little more detailed
→ Annotations quality
  → not as detailed as from experts but sufficient
→ Real-world application
  → web texts processing
  → verification of corpora?
  → building new corpora?

STU FIIT
j.harinek@gmail.com
PeWe@FIIT personalized web group