Some thoughts on CLEANEVAL 2 (and 3)

Stefan Evert
Institute of Cognitive Science
University of Osnabrück

UNIVERSITÄT OSNABRÜCK

CLEANEVAL 1.5

- Collect & merge gold standard data
  - FIASCO gold standard (158 pages)
  - CLEANEVAL devset (58 pages)
  - Charles University extension (46 pages?)
  - CLEANEVAL evaluation data (?? pages)
- Efficient standard evaluation with cleaneval.py
- Encourage further research on supervised learning approaches (as well as heuristics)
  - published evaluation results will be comparable

CLEANEVAL 1.5

- Directions for further research & development
  - supervised learning can achieve high accuracy, even on training sets of moderate size
  - sequence tagging is a good idea (→ won the contest)
  - some things are best done with (clever) heuristics
  - find more informative features for ML algorithms
- Collaboration?
  - many complementary approaches & features
- Can we produce a practical open-source tool?

CLEANEVAL 2

- Who?
  - I'm about 60% inclined to organise the next contest …
    … if someone is willing to do it together with me
- When? – 2009
  - in the meantime, research on current gold standard
- What?
  - languages: English & {German, French, Russian, …}
  - get gold standard right (sampling, HTML alignment)
  - once we've learned to clean up English Web pages, CLEANEVAL 3 will focus on multilingual processing

CLEANEVAL 2++

- What I am really interested in is the next step!
  - tokenisation
  - part-of-speech tagging
  - lemmatisation (e.g. for German)
- These aren't solved problems for Web data
  - poor quality (e.g. POS tagging ca. 90% accuracy, compared to 97.5% in published evaluations)

CLEANEVAL 2++

- The formula for Really Useful Web Corpora
  - good WaC spam detection (perhaps the easiest step)
  - high-precision boilerplate removal (> 90% precision)
  - reliable (and consistent) tokenisation
  - accurate part-of-speech tagging (≈ 95% accuracy)
  - acceptable lemmatisation quality