Computerlinguistik II

Vorlesung im SoSe 2019
(M-GSW-10)

Prof. Dr. Udo Hahn

Lehrstuhl für Computerlinguistik
Institut für Germanistische Sprachwissenschaft
Friedrich-Schiller-Universität Jena

http://www.julielab.de

Two Paradigms for NLP

- Symbolic Specification Paradigm
  - Manual acquisition procedures
  - Lab-internal activities
  - Intuition and (few!) subjectively generated examples drive progress based on individual (competence) judgments
    - "I have a system that parses all of my nine-teen sentences!"

Symbolic Specification Paradigm

- Manual rule specification
  - Source: linguist´s intuition
- Manual lexicon specification
  - Source: linguist´s intuition
- Each lab has its own (home-grown) set of NLP software
  - Hampers reusability
  - Limits scientific progress
  - Waste of human and monetary resources (we “burnt” thousands of Ph.D. student all over the world ☹)

Shortcomings of the "Classical" Linguistic Approach

- Huge amounts of background knowledge req.
  - Lexicons (approx. 100,000 – 150,000 entries)
  - Grammars (>> 15,000 – 20,000 rules)
  - Semantics (>> 15,000 – 20,000 rules)
- As the linguistic and conceptual coverage of classical linguistic systems increases (slowly), it still remains insufficient; systems also reveal 'spurious' ambiguity, and, hence, tend to become overly "brittle" and unmaintainable
- More fail-soft behavior is required at the expense of … ? (e.g., full-depth understanding)

Two Paradigms for NLP

- Symbolic Specification Paradigm
  - Manual acquisition procedures
  - Lab-internal activities
  - Intuition and (few!) subjectively generated examples drive progress based on individual (competence) judgments
    - "I have a system that parses all of my nine-teen sentences!"
- **Empirical (Learning) Paradigm**
  - **Automatic acquisition procedures**
  - **Community-wide sharing of common knowledge and resources**
  - **Large and ‚representative' data sets drive progress according to experimental standards**
    - **"The system was tested on 1,7 million words taken from the WSJ segment of the MUC-7 data set and produced 4.9% parsing errors, thus yielding a statistically significant 1.6% improvement over the best result by parser X on the same data set & a 40.3% improvement over the baseline system!"**

Empirical Paradigm

- Large repositories of language data
  - Corpora (plain or annotated, i.e., enriched by meta-data)
- Large, community-wide shared repositories of language processing modules
  - Tokenizers, POS taggers, chunkers, NE recognizers, ...
- Shared repositories of machine learning algos
- Automatic acquisition of linguistic knowledge
  - Applying ML algos to train linguistic processors by using large corpora with valid linguistic metadata (linguist as educated data supplier, „language expert“) rather than manual intuition (linguist as creative rule inventor)
- Shallow analysis rather than deep understanding
- Large, community-wide self-managed, task-oriented competitions, comparative evaluation rounds
- Change of mathematics:
  - Statistics rather than algebra and logics

Paradigm Shift – We Exchanged our Textbooks…