Part-of-Speech Tagging

Slides: Brigitte Krenn

Sommersemester 2004

ÖFAI, Wien

brigitte@ai.univie.ac.at

Part-of-Speech (PoS) Tagging

- Task: assign a word string with a PoS-string
- Why?
  - Knowledge of (disambiguated) syntactic category is useful in a variety of applications, such as
    - Parsing
    - Collocation Identification
    - Text Categorization
    - Text Summarization
    - Text-to-Speech Synthesis
    - etc.

Tagsets

- Construction Criteria
  - moderate in size: 50 - 100 tags is a good size in statistical tagging
  - discriminatory: information in tagset shall be appropriate for local disambiguation

Tagsets

- Typical information represented by the tags
  - syntactic category: N, V, Adj, Adv, Art, …
  - syntactic-semantic: modal verb, …
  - inflectional: preposition with incorporated determiner, finite verb, nonfinite verb, ...
  - pragmatic: particle
  - positional: attributive adjective, adverbial adjective
  - punctuation: . , ; ? ! ( ) “ …
  - specific categories: foreign material, names, dates, unknown, ...

Accuracy of Taggers

- State of the art tagger: 97 -- 99% accuracy
- 2 measures are of interest
  - recall

$$\frac{\#\text{correctly assigned tags}}{\#\text{ correct tags total}}$$

  - precision

$$\frac{\#\text{correctly assigned tags}}{\#\text{ assigned tags}}$$

Problems of Accuracy Comparison

- Method to count:

  all words vs. only ambiguous words

- tagset compatibility

Preprocessing/Word Splitting

- Typical input format for taggers:

  word per line

- How to identify a word?
  - Simplest: <blank>character sequence<blank>

  – Problems:
    - En: compound nouns
    - How many words? isn’t, Peter’s
    - Hyphenation: Peters-\n

      bergkonferenz
    - Abbreviation vs. full stop
    - complex units: times, dates, names
    - identification of foreign material

Minimalist Approach to Tagging

- Observation: the distributions of tags per word are typically uneven, thus

  - extract a frequency lexicon from a tagged corpus
  - look up the words of the new text in the frequency lexicon, and
  - assign the most frequent PoS

Minimalist Approach to Tagging

Problem: there are unknown words

tag all unknown words as common or proper noun

The larger the lexicon and the closer the new text is to the text used for the lexicon the better the tagging results, e.g. Charniak et al. 93 report 90% tagging accuracy for the Brown Corpus

Statistical Approaches

- Decision trees, e.g. Schmid 94
- Memory-based learning, e.g. Daelemans et al. 96
- Maximum-entropy, e.g. Ratnaparkhi 96
- Hidden Markov Models, e.g. Brants 99

For an overview of statistical taggers see

Manning, Schütze: *Foundations of Statistical Natural Language Processing*. 1999.

See also

- Symbolic approaches, e.g.
  - Constraint grammar: Karlsson et al. 95,
    Voutilainen 97, e.g. ENGCG

    hand-written rules compiled into finite-state automata
  - Transformation-based learning: Brill 95

    possible transformations are manually specified,

    learning = greedy search for optimal sequence

    greedy algorithm: always take best immediate or local solution