T-(538|725)-MALV, Natural Language Processing
Review and exam preparation

Hrafn Loftsson[1] Hannes Högni Vilhjálmsson[1]

[1]School of Computer Science, Reykjavik University

November 2008

Outline

Outline

Corpora

Knowledge units

- What is it?
- Types of corpora
- Annotated corpora
- Utility of corpora

Finite-state automata (FSA)

Knowledge units

- What is it?
- Types of FSA
- Efficiency
- Operations on FSA

Regular expressions (Regexs)

Knowledge units

- Strings and languages
- Operations on languages
- What is a regex?
- Regex operators
- Connection between regexs and FSA.

Perl

Basics

- Format of a program
- Data types: scalars, arrays, hashes
- Control structures: if, while, for
- File handling
- Regular expressions

Tokenisation

Knowledge units

- What is it?
- Word segmentation/Sentence segmentation
- What are the problems?
- Lexical analyser – JFlex

Word counting and n-grams

Knowledge units

- Language model
- Word types vs. tokens
- n-grams
- Construction of n-gram models
- Maximum likelihood estimation
- Probability of a sentence using bigrams/trigrams
- n-fold cross-validation
- Smoothing

Text processing tools

We have discussed:

- grep
- sed
- tr
- sort
- uniq
- paste, head, tail
- awk (briefly mentioned)

Morphology

Knowledge units

- Part-of-speech (POS)
- Morpheme – stems and affixes
- Morphological analysis – lemmatisation, stemming
- Morphological generation
- Two-level morphology
- Finite-state transducer

POS tagging

Knowledge units

- What is it?
- Tagsets
- Full disambiguation vs. not full
- Baseline tagging
- Accuracy, precision, recall, ambiguity rate of taggers
- Problems with unknown words

POS taggers

Knowledge units

- Rules vs. statistics
- Data-driven vs. linguistic rule-based
- Brill's tagger
- IceTagger
- Statistical taggers like TnT

$$P(t_1)P(t_2|t_1)\prod_{i=3}^{n} P(t_i|t_{i-2},t_{i-1})\prod_{i=1}^{n} P(w_i|t_i)$$

Syntax

Knowledge units

- Constituents
- Syntactic (grammatical) functions
- Verb subcategorisation
- Context-free grammar
- Parsing, parse tree, derivation
- Full parsing vs. partial (shallow) parsing
- IceParser

Parsing techniques

Knowledge units

- Top-down vs. bottom-up parsing
- Chart parsing (but not in detail)
- Probabilistic parsing (but not in detail)

Semantics

Knowledge units

- Principle of compositionality
- λ-calculus
- First-order predicate calculus
- Quantifiers

Discourse and Dialog

Knowledge units

- See the slides that Hannes presented in last class