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

December 2009

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
- Subroutines
- 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)
- Morphemes – 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 and predicate logic

Knowledge units

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

Lexical semantics

Knowledge units

- Basic terms and concepts: synonymy, antonymy, homonymy,
- Ontology
  -
  - hyponymy, hypernymy
- WordNet
- Word Sense Disambiguation

Outline

Discourse Analysis and Discourse Model

- What is **Discourse** and what is the topic of **Discourse Analysis**?
- What is the difference between **Transactional** and **Interactional** language?
- What is the difference between **Discourse Function** and **Discourse Device**?
- What are the **Gricean Maxims** and why are they important for language interpretation?
- What is **Discourse Context**, **Discourse Model** and a **Discourse Entity**?
- How do you find **Coreference** with a simple **Recency List** method?

Information Structure and Information Status

- What is **Information Structure**
- What is a **Theme** and a **Rheme**?
- What discourse devices may express Information Structure?
- How are Information Structure and **Intonation** related?
- What is **Information Status** and what are the different values it can take?

Discourse Structure and Attentional Stack

- What is **Discourse Structure** and what evidence do we have that such a structure exists?
- What properties does a **Discourse Segment** have?
- What is the difference between the **Intentional** and **Information** view of segments?
- What is a **Discourse Purpose** and a **Discourse Segment Purpose**?
- What is the **Attentional Stack** and a **Discourse State**?
- What is a **Discourse Marker** and how do they relate to the attentional stack?
- What are **Rhetoric Relations** (roughly)?

Simple Dialog Systems, Speech Acts

- Be able to draw an **Finite State Automata** for a simple dialog system.
- Know the difference between **Implicit** and **Explicit** feedback on understanding.
- What are **Adjacency Pairs** and their extended **3 Sentence Intervention** version?
- Know the three parts of **Speech Acts**: **Locutionary**, **Illocutionary**, **Perlocutionary**.
- You do not need to memorize all speech act types, but it helps to know examples of a few top level categories.
- Understand the need for talking about **Dialog Acts** instead of speech acts.
- What is an **Agent Model** and how does that relate to **Grounding**?

Nonverbal Behavior

- What does **Multimodal Communication** mean?
- What do we mean by **Comprehensive Communicative Act**?
- What are some of the evidence that nonverbal behavior **Communicates**?
- How does nonverbal behavior relate to **Discourse Functions and Devices**?
- How does one go about studying the relationship between functions and devices?
- What are some examples of function to **nonverbal behavior mappings**?
- Be ready to describe how knowledge of nonverbal behavior can benefit NLP based systems.