T-(538|725)-MALV, Natural Language Processing Introduction

Hrafn Loftsson[1] Hannes Högni Vilhjálmsson[1]

[1]School of Computer Science, Reykjavik University

August 2008

Outline

Outline

Learning outcome and content

Learning outcome

The course objectives are that students:

- Know the main methods used in the field of natural language processing
- Are familiar with the main research areas in the field
- Are able to implement a system which processes a natural language

Content

The goal of language technology (LT) is to develop systems which allow people to communicate with computers using natural languages. LT is an interdisciplinary field, requiring knowledge from subjects like linguistics, statistics, psychology, engineering and computer science. This course discusses fundamentals of natural language processing (NLP), which is one of the subfields of LT . . . (see the course web page).

Learning outcome and content

Learning outcome

The course objectives are that students:

- Know the main methods used in the field of natural language processing
- Are familiar with the main research areas in the field
- Are able to implement a system which processes a natural language

Content

The goal of language technology (LT) is to develop systems which allow people to communicate with computers using natural languages. LT is an interdisciplinary field, requiring knowledge from subjects like linguistics, statistics, psychology, engineering and computer science. This course discusses fundamentals of natural language processing (NLP), which is one of the subfields of LT . . . (see the course web page).

Text books

Main text

An Introduction to Language Processing with Perl and Prolog

Other books - available in the RU library

- Foundations of Statistical Natural Language Processing
- Speech and Language Processing
- Handbook of Natural Language Processing
- Learning Perl

Text books

Main text

An Introduction to Language Processing with Perl and Prolog

Other books - available in the RU library

- Foundations of Statistical Natural Language Processing
- Speech and Language Processing
- Handbook of Natural Language Processing
- Learning Perl

Course assessment

Individual parts

- Programming projects: **40%**
  - 5 for MSc students, 4 for BSc students. Can be worked on in a group of two students.
- A final written exam: **30%**
  - The grade 5.0 is needed to pass the course.
- Participation in course: **15%**
  - MSc students: Reading and presentation of a paper.
  - Participation in class and in forum discussion.
- Three assignments: **15%**
  - Assignments are worked on individually.

The programming project

Consists of five parts

- Part I: Tokenisation
- Part II: POS tagging
- Part III: Shallow parsing
  - BSc students receive this part for "free"
- Part IV: Discourse model
- Part V: Final project

The project schedule (and days to prepare)

- Week 3: Assignment I: 10.09.2008 (8 days)
- Week 6: Programming Project I: 29.09.2008 (15 days)
- Week 8: Assigment II: 13.10.2008 (10 days)
- Week 9: Programming Project II: 20.10.2008 (15 days)
- Week 10: Programming Project III: 29.10.2008 (10 days)
- Week 11: Assigment III: 05.11.2008 (8 days)
- Week 12: Programming Project IV: 12.11.2008 (10 days)
- Week 14: Final Project: 26.11.2008 (15 days)

Outline

Language Technology

Goal

- The goal of (human) language technology (HLT, LT) is to develop systems which allow people to communicate with computers using natural languages.
- The Icelandic term is "Máltækni (tungutækni)"
- Interdisciplinary field — interplay of fields like linguistics, statistics, psychology, engineering and computer science.

Two main subfields

- Text (Language) Processing (í. Textavinnsla)
- Speech Processing (í. Talvinnsla)

Language Technology

Goal

- The goal of (human) language technology (HLT, LT) is to develop systems which allow people to communicate with computers using natural languages.
- The Icelandic term is "Máltækni (tungutækni)"
- Interdisciplinary field — interplay of fields like linguistics, statistics, psychology, engineering and computer science.

Two main subfields

- Text (Language) Processing (í. Textavinnsla)
- Speech Processing (í. Talvinnsla)

Natural Language Processing (NLP)

LT vs. NLP

- Language Technology (LT) $\approx$ Natural Language Processing (NLP)
- í. Máltækni $\approx$ málvinnsla
- In NLP, the emphasis is on:
  - The analysis (í. greining) of structure (í. formgerð) and semantics (í. merking) of a language
  - The generation (í. myndun) of language from structure/semantics.
- NLP $\approx$ Computational Linguistics (í. tölvufræðileg málvísindi)

Outline

LT Projects

Examples

- **Grammar checking** (í. Málfræðileiðrétting)
  - `http://en.wikipedia.org/wiki/Grammar_checker`
- **Information retrieval** (í. Upplýsingaheimt) and **Information Extraction** (í. Upplýsingaútdráttur)
  - `http://en.wikipedia.org/wiki/Information_extraction`
- **Question-Answering Systems** (í. Fyrirspurnarkerfi)
  - `http://en.wikipedia.org/wiki/Question_answering`
- **Machine Translation** (í. Vélrænar þýðingar)
  - `http://en.wikipedia.org/wiki/Machine_Translation`

LT Projects

More examples

- **Speech recognition** (í. Talkennsl/Talgreining)
  - `http://en.wikipedia.org/wiki/Speech_recognition`
- **Speech synthesis; text-to-speech** (í. Talgerving)
  - `http://en.wikipedia.org/wiki/Speech_synthesis`
- **Dialogue Systems** (í. Samræðukerfi)
  - `http://nlp.shef.ac.uk/research/areas/dialogue.html`

A speaking robot

HAL

- The movie *2001: Space Odyssey*. Director: Stanley Kubric. Made in 1968.
- A computer which talks and understands English.
- The movie made a prediction 33 years into the future.
- How close is this prediction to reality?
- What is needed to construct an agent, like HAL, which possesses langugae generation and language understanding cababilities?

Outline

The disciplines of lingustics – from sounds to meaning

- Phonetics and Phonology (í. Hljóðfræði og hljóðkerfisfræði)
- Morphology (í. Orðhlutafræði)
- Syntax (í. Setningafræði)
- Semantics (í. Merkingarfræði)
- Discourse and Dialogue (í. Orðræða og samræða)

Theses disciplines comprise the different levels of LT.

Outline

Ambiguity (í. Margræðni)

- Ambiguity occurs when more than one linguistic structure is associated with a particular input.
- In other words, when different kinds of meanings can be associated with the input.
- In most cases, humans remove the ambiguity unconsciously.
- On the other hand, ambiguity is a major obstacle in language processing and can occur in all the different levels of LT.
- Ambiguity is removed by applying disambiguation (í. einræðing).

Ambiguity in speech recognition

Example

- Input: The boys eat the sandwiches.
- Possible output:
  - The boy seat the sandwiches.
  - The boy seat this and which is.
  - The boys eat this and which is.
  - The boys eat the sand which is.
  - etc.

Ambiguity in part-of-speech tagging (í. (orðflokks)mörkun)

Example

- Input: Hann á við (he owns wood).
- Tags of individual words:
  - Hann=fpken_fpkeo
  - á=aþ_ao_sfg1en_sfg3en_aa_nven_nveo_nveþ
  - við=ao_fp1fn_aþ_aa_nkeo

Meaning of individual letters in tags:

```
n=nominative,nefnifall, o=accusative,þolfall,
þ=dative,þágufall, e=genitive,eignarfall
n=noun,nafnorð, f=pronoun,fornafn, p=personal pronoun,persónufornafn,
a=adverb,atviskorð, s=verb,sögn
k=masculine,karlkyn, v=feminine,kvenkyn
e=singular,eintala, f=plural,fleirtala
f=indicative mood,framsöguháttur, g=active voice,germynd
```

Ambiguity in syntax/semantic analysis

Example

- Input: I saw the boy with the telescope.
- Meaning:
  - I used a telescope to see the boy.
  - I saw the boy who had a telescope.

Selection and implementatin of a model

When a natural language is analysed:

- A formal model needs to be developed.
  - A good model is difficult to design.
  - A language is closely tied to human thought and understanding.
- The model needs to be implemented in a program.

Outline

Tungutækni - Report by a committee

Ministry of Education, Science and Culture, 1999

- `http://www.tungutaekni.is/news/Skyrsla.pdf`
- Main question: "Why should a population of only 300,000 people strive to make the Icelandic language suitable for use in an information technology society?"
- Proposals:
  - Corpora should be built and made accessible for research and development of LT tools.
  - A special fund should be established to support research in the field of LT.
  - Companies should be sponsored in order to develop LT tools.
  - Educational programs in the field of LT should be established.

Forum discussion of the week

LT in your country

- What is the status of LT in your country?
- Which resources/tools are available?

Forums

- `http://malv2008.proboards57.com/`

Outline

Web sites

- The Icelandic Centre for Language Technology (í. Tungutæknisetur): `http://www.tungutaekni.is`
- Center for Analysis & Design of Intelligent Agents (í. Gervigreindarsetur HR): `http://ailab.ru.is`
- IceNLP: `http://nlp.ru.is`
- Foreign Language and Culture Training: `http://alelo.com/`