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

Hrafn Loftsson[1] Hannes Högni Vilhjálmsson[1]

[1]School of Computer Science, Reykjavik University

September 2010

Outline

1. Language Technology/Natural Language Processing
2. Language Technology Projects
3. The disciplines of linguistics
4. Why is LT difficult?

Outline

1 Language Technology/Natural Language Processing

2 Language Technology Projects

3 The disciplines of linguistics

4 Why is LT difficult?

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"
- 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"
- 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

1 Language Technology/Natural Language Processing

2 Language Technology Projects

3 The disciplines of linguistics

4 Why is LT difficult?

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/dialogue.html`

A speaking robot

HAL

- The movie *2001: Space Odyssey*. Director: Stanley Kubric. Made in 1968.
- A computer which talks and understands English (`http://en.wikipedia.org/wiki/HAL_9000`).
- 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 language generation and language understanding cababilities?

Outline

1 Language Technology/Natural Language Processing

2 Language Technology Projects

3 The disciplines of linguistics

4 Why is LT difficult?

The disciplines of linguistics – 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

1 Language Technology/Natural Language Processing

2 Language Technology Projects

3 The disciplines of linguistics

4 Why is LT difficult?

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. 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.

Ambiguity in anaphora resolution

Definition

In linguistics, *anaphora* is an instance of an expression referring to another.

Example

- Input 1: The City Council refused the women a permit because they feared violence.
- Input 2: The City Council refused the women a permit because they were communists.
- **Question**: To which noun phrase does the pronoun "they" refer to?

Ambiguity in anaphora resolution

Definition

In linguistics, *anaphora* is an instance of an expression referring to another.

Example

- Input 1: The City Council refused the women a permit because they feared violence.
- Input 2: The City Council refused the women a permit because they were communists.
- **Question**: To which noun phrase does the pronoun "they" refer to?

Selection and implementation 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.