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

Introduction

Hrafn Loftsson¹  Hannes Högni Vilhjálmsson¹

¹School of Computer Science, Reykjavik University

September 2009
Outline

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

1. Language Technology/Natural Language Processing
2. Language Technology Projects
3. The disciplines of linguistics
4. Why is LT difficult?
5. LT in your country
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. Textavinsla)
- Speech Processing (í. Talvinnsla)
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) ≈ Natural Language Processing (NLP)
- í. Máltaekni ≈ 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 ≈ Computational Linguistics (í. tölvufraðileg málvísindi)
Outline

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

- **Grammar checking** (í. Málfræðileiðrétting)

- **Information retrieval** (í. Upplýsingaheimt) and **Information Extraction** (í. Upplýsingaútdráttur)
  - [http://en.wikipedia.org/wiki/Information_extraction](http://en.wikipedia.org/wiki/Information_extraction)

- **Question-Answering Systems** (í. Fyrirspurnarkerfi)

- **Machine Translation** (í. Vélrænar þýðingar)
LT Projects

More examples

- **Speech recognition** (í. Talkennsl/Talgreining)
- **Speech synthesis; text-to-speech** (í. Talgerving)
- **Dialogue Systems** (í. Samræðukerfi)
  - [http://nlp.shef.ac.uk/research/areas/dialogue.html](http://nlp.shef.ac.uk/research/areas/dialogue.html)
A speaking robot

HAL

- 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 capabilities?
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.
1. Language Technology/Natural Language Processing
2. Language Technology Projects
3. The disciplines of linguistics
4. Why is LT difficult?
5. LT in your country
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.
Example

- **Input:** Hann á við (he owns wood).
- **Tags of individual words:**
  - Hann=fpken_fpkeo
  - á=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?
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.
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.
LT in your country

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