Main Trends in Semantic-Research of Estonian Language Technology

Haldur ÕIM, Heili ORAV, Kadri KERNER and Neeme KAHUSK

October 8, 2010
Introduction
A short history of computational linguistics in Estonia

Estonian Wordnet
Structure of WordNet
About EstWN
Enlarging EstWN
Metaphors, Multi-Word Units and Compounds in EstWN

Sense disambiguation
WSD Corpus of Estonian

Semantic analysis of sentences
A short history of computational linguistics in Estonia

- History started before teaching of computational linguistics already in early sixties.
  - The first electronic computer in Estonia was established at Tartu University in 1959
- One of the first non-mathematical tasks the enthusiasts attacked was machine translation.
- A special program of mathematical and structural linguistics was started.
- Beginning of 70ties started to build an information retrieval system for legal texts in Estonian, and in the frames of this a thesaurus of legal terms was compiled.
Turn to artificial intelligence and, in the frames of this, to language understanding and human-computer interaction.

- Language understanding system TARLUS

And this was the actual beginning of the Research Group of Computational Linguistics.

In 2006 started the National Program for Estonian Language Technology.
Structure of WordNet

- Wordnets have been developed for several languages (over 50 languages) in the world.
- The main idea and basic design of all wordnets in the project came from Princeton WordNet.
- Each wordnet is structured along the same lines: synonyms are grouped into synonym sets (synsets).
- Synsets are connected to each other by semantic relations:
  - like hyperonymy (is-a) and meronymy (is-part-of)
  - most of them are reciprocated (e.g. if koer (dog) has hyperonym loom (animal) then loom (animal) has hyponym koer (dog))
Estonian Wordnet

- For Estonian there are two concept-based thesauri available:
  - thesaurus compiled by Saareste has more of an historic value
  - the modern and most famous one is the Estonian Wordnet
- The creation of Estonian Wordnet was started within the project EuroWordNet (1997-2000).
- In 2006 started the project for increasing EstWN and is supported by Estonian National Programme on Human Language Technology.
- The number of concepts in thesaurus is more than 40,000 (nouns, verbs, adjectives, adverbs).
- There are 43 semantic relations used in Estonian WordNet.
Enlarging EstWN

- Enlarging manually and domain-specificly.
- Concepts from semantic fields like architecture, transportation, personality traits and so on.
- Since one person is dealing with one domain at the time, then it makes the relations between different concepts (in one domain) easier to determine:
  - The concept antique tempel has 1 hyperonym, 11 hyponyms, 1 holo-part and 8 mero-part relations.
- Around 3000 noun synsets were automatically transferred from the Estonian Synonym Dictionary.
- Automatically we have included an amount of words which have been derived via suffixes (”Enriching Estonian WordNet with derivations and semantic relations” in Proceedings).
Metaphors, Multi-Word Units and Compounds in EstWN

- How to supplement metaphors and multi-word units (idioms etc) into EstWN.
- Metaphors and metaphorical meanings of words are a topical issue in linguistics and lexicology and they surely should be considered in building a thesaurus.
- In Estonian, compounds are almost always written as single words and therefore separated from multi-word expressions.
- Including compound words into wordnet-type thesaurus is a problem for Estonian language.
- There the usage of Corpus of Estonian Written Language can be helpful, it is important to include at least the frequent ones.
Revision of Estonian Wordnet

- Revising of adverbs using a questionnaire type of mini-test
  - Sense granularity.
  - Clarity of definitions and examples.
- Revising the Taxonomies in EstWN
  - In EstWN word 'inimene' (person) is the word with most hyponyms, more than 800 all together.
  - Study presented solutions of how to decrease the amount of the persons hyponyms.
The first project of creating Word Sense Disambiguation Corpus of Estonian started in 2001 within the Senseval-2 competition and this project lasted for a year.

During the first stage around 110,000 tokens were manually annotated according to senses in Estonian WordNet.

There were 43 morphologically analyzed texts of fiction from the Corpus of the Estonian Literary Language.

Only nouns and verbs were the subject of annotation.
The second project started in 2009.

There are included newspaper texts, scientific texts, informational texts and legal texts.

Texts come from morphologically disambiguated corpus of Estonian.

We are now annotating nouns, verbs and also adjectives and adverbs, since these parts of speeches are now present in EstWN.

We hope to reach to the total amount of words in the corpus of 500 000 by the end of 2010.
Annotation of Senses

- Firstly texts (of 2000 words) are pre-annotated.
  - Monosemous words.
  - Certain word forms.
  - "One sense per one collocation" word pairs.
- Secondly two human annotators tag the words which have not been tagged by the pre-annotation system or correct tags added by pre-annotation process.
  - Annotation tool KYKAP.
- And finally, third person solves the disagreements.
Outline
Introduction
Estonian Wordnet
Sense disambiguation
Semantic analysis of sentences

WSD Corpus of Estonian

KYKAP

Main Trends in Semantic-Research of Est LT
For about 5 years working on a LT project called Semantics of simple sentences.

One of the distant goals in natural language processing has been the semantic analysis of language.

In addition to the recognition of structure of words and sentences, the computer could also understand the meaning of sentences (ultimately, of texts).

More at 11.30 presentation: Semantic Analysis of Sentences: The Estonian Experience