Main Trends in Semantic-Research of Estonian Language Technology

Haldur ÕIM [1], Heili ORAV, Kadri KERNER, and Neeme KAHUSK
Institute of Computer Science, University of Tartu

Abstract. The paper gives a general overview of the development of computational semantics in Estonia beginning from the second half of the 20th century. Main focus concentrates on the work we have done so far and on the problems we try to solve at present in our research group of computational linguistics at Tartu University.

Keywords. computational semantics

Our works in the field of semantic analysis can be divided—quite conventionally, of course—into three main themes: lexical semantic resources, sense disambiguation and semantics of sentence. But first we give for the background an overview of the development of computational linguistics in Estonia.

1. A short history of computational linguistics in Estonia

History of computational linguistics in the University of Tartu has been quite long. It started before teaching of computational linguistics—already in early sixties. The first electronic computer in Estonia was established at Tartu University in 1959 and one of the first "non-mathematical" tasks the enthusiasts attacked was machine translation. We failed, of course, but what we learned was very important: that the methods and forms of language description for computer should be quite different from those intended for humans.

At our university a special program of mathematical and structural linguistics was started: the students participating in this program received special teaching in new trends of linguistics (including, of course, classical schools of structural linguistics and generative grammar) and several mathematical disciplines, from mathematical logic to statistics. The first real task in the area that at present is called language technology was, surprisingly, in the field of semantic resources: at the very beginning of 70ties we started to build an information retrieval system for legal texts in Estonian, and in the frames of this we compiled a thesaurus of legal terms (concepts) where the classical semantic relations (synonymy, hyponymy, part-whole, several functional relations, e.g. causality) were fixed.

After the information retrieval project we turned to artificial intelligence and, in the frames of this, to language understanding and human-computer interaction. We started to

[1] Institute of Computer Science, University of Tartu, Liivi 2 50409 Tartu, ESTONIA E-mail: haldur.oim@ut.ee

build a language understanding system called TARLUS (=TARtu Language Understanding System), were actively involved in "All-Union" activities including regular meetings/seminars with the common name Dialogue (these meetings, by the way, occur regularly to this day). For TARLUS we had to create (preliminary) programs for morphological and syntactic analysis of Estonian, but, of course, to continue to develop our semantic resources. And this was the actual beginning of our Research Group of Computational Linguistics. Main trends have been in the fields of morphology, syntactical analysis, semantic analysis and pragmatics (dialogue models). And all kinds of language resources. When in the middle of 90ties EU started the COPERNICUS program, we joined it, as did several research groups from other Baltic countries.

In 2006 started the National Program for Estonian Language Technology[2] (see [1] as well) which by the idea should cover all areas of language processing, from speech technology to pragmatics of human interaction, that are considered relevant "to enable Estonian to function seamlessly in the modern information technology infrastructure". In the following we describe three trends and their results in the area that we qualify as semantic.

2. Estonian Wordnet

During the last decades, wordnets have been developed for several languages (over 50 languages) in the world[3]. For Estonian there are two concept-based thesauri available. First thesaurus [2] has more of an historic value (compiled by Andrus Saareste as war refugee in Uppsala in 1979) and second, the modern and most famous one is the wordnet-type thesaurus of Estonian. The creation of Estonian Wordnet[4] was started within the project EuroWordNet (EWN, see also [3])[5]. The Estonian team joined the project supported by European Union in 1998 together with Czech, French and German languages. In the framework of the project the Estonian Wordnet has been created during the years 1997–2000. After some discontinuation this project was awaken again. In 2006 started the project for increasing EstWN and is supported by Estonian National Programme on Human Language Technology. Thanks to governmental program our thesaurus has enlarged a lot—the number of concepts in thesaurus is more than 34 000 (June 2010).

The main idea and basic design of all wordnets in the project came from Princeton WordNet (more in [4]). Each wordnet is structured along the same lines: synonyms (sharing the same meaning) 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')). There are 43 semantic relations used in Estonian Wordnet. Different wordnets of each language are connected with each other via special ILI (Inter-Lingual-Index) relations. ILI concepts themselves do not have intra-language relations, this allows handling lexicalization and knowledge (ontology) separately: see [3] for futher details.

[2] www.keeletehnoloogia.ee/

[3] There are currently around 50 wordnets to different languages in the world (see more http://www.globalwordnet.org/).

[4] EstWN, see http://www.cl.ut.ee/ressursid/teksaurus/

[5] See http://www.illc.uva.nl/EuroWordNet/

The wordnet builders all around have applied different compilation strategies. Our chosen approach so far for enlarging has been manual and domain-specific, i.e we have added 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 [13]. For example from the domain of architecture the concept *antiiktempel* ('antique tempel') has 1 hyperonym, 11 hyponyms, 1 has_holo_part and 8 has_mero_part relations.

We have tried some ways to enlarge Estonian Wordnet automatically also. For instance, during the increasing process of EstWN around 3000 noun synsets were automatically transferred from the Estonian Synonym Dictionary [16]. After this attempt we discovered that manual work gives more high-quality result because the revision is too long-standing. Automatically we plan to include an amount of words which have been derived via suffixes. Most frequent suffix between noun and verb is *-mine* (i.e *kõndima* 'to walk' — *kõndimine* 'walking'). This approach gives us thousands of new entries. This work is described in [5].

Besides including domain-specific vocabulary we have started to think about how to supplement metaphors and multi-word units (idioms etc) into EstWN, because it would increase the size and usability of the thesaurus to a remarkable degree. Metaphors and metaphorical meanings of words are a topical issue in linguistics and lexicology and they surely should be considered in building a thesaurus [15] . But their occurrence in text is really rather unpredictable and chaotic. And if we add the metaphorical uses to the thesaurus, then how should we explain them properly. As is known, the understanding of a metaphor depends on the context.

A multi-word unit is a combination of two or more words that occur together to express a single meaning. In English, compound words are often written separately and therefore seen as a kind of multi-word expression. In Estonian, compounds are almost always written as single words and therefore separated from multi-word expressions. The fact that there is no certain definition for neither of these expressions makes it also difficult to include them in wordnets. There are several problems that occur when adding them, for example formal and semantic problems as well as some more specific problems like handling prepositions in the wordnet structure (Fellbaum 1998). Besides, some idiomatic constructions are just too complex and variable to integrate them. It can be said that although there are many multi-word expressions already included, inaccuracies in semantic relations and missing synonyms are rather frequent.

Including compound words into wordnet-type thesaurus is a problem for Estonian language as well as for example for the German language, because in both of these languages words can be combined quite freely while the meaning still stays understandable. Nevertheless, the number of compounds in wordnets should be somehow restricted. There the usage of Corpus of Estonian Written Language can be helpful. It is important to include at least the frequent ones.

To sum up, it appears that the creation of a concept-based thesaurus is not as easy as it seems at first sight. The main problems we face nowadays in setting up a thesaurus include:

- Possibility of automatic extension
- Multi-word combinations.

3. Sense disambiguation

Secondly named task is word sense disambiguation of Estonian language. Currently we are working on the increasing of the Word Sense Disambiguation Corpus of Estonian and we hope to reach to the total amount of words in the corpus of 500 000 by the end of 2010.

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 (see [14]). During the first stage around 110 000 tokens were annotated. There were 43 morphologically analyzed texts of fiction from the Corpus of the Estonian Literary Language[6] and only nouns and verbs were the subject of annotation.

The second project started in 2009. Since the first project dealt with fictional texts, then now we have included newspaper texts, scientific texts, informational texts and legal texts. These texts come from morphologically disambiguated corpus of Estonian[7]. Compared to the previous project we are now annotating nouns, verbs and also adjectives and adverbs, since these parts of speeches are now present in EstWN. The texts are divided into parts of ca 2000 words of each, and annotated by two people. In the first project the disagreement of two annotators was settled by discussion, now we have decided that it is more effective if the disagreements are resolved by the third annotator.

As a sense division we are using Estonian Wordnet and for disambiguation there has been developed a tool KYKAP [7] which is meant to assist the human annotator and speed up the annotation process.

The annotation-task is divided into three parts. Firstly texts are pre-annotated. For speeding up the annotation process we pre-annotate the words that are monosemous in EstWN. Also many of the highly polysemous word forms indicate to a certain sense and now these word forms are included in the pre-annotation task. From the sense annotated corpus it is possible to extract word pairs which tend to have one sense per one collocation [8] and these collocations are then being used in pre-annotation as well.

After pre-annotation human annotators tag the words which have not been tagged by the pre-annotation system or correct tags added by pre-annotation process. And finally, third person solves the disagreements.

This number of words and different text types makes WSDCEst hopefully a valuable resource for WSD systems as training and testing data, also for some basic statistics about word sense distributions.

4. Semantic analysis of sentences

The third direction in our research we would like to give an overview of is semantic analysis of (simple) sentences of Estonian. One of the distant goals in natural language processing has been the semantic analysis of language, so that in addition to the recognition of structure of words and sentences, the computer could also understand the meaning of sentences (ultimately, of texts). Let us note that the solution of this task is also a precondition of the solution of several pragmatic tasks (human-computer interaction in natural language). We have worked at this problem about 5 years. The input of the corresponding

[6] http://www.cl.ut.ee/korpused/baaskorpus/ (21.03.2010)

[7] http://www.cl.ut.ee/korpused/morfkorpus (10.06.2010)

program is the syntactic tree of a sentence and the output is its representation in the form of a frame where the syntactic roles (Subject, Object etc) are replaced with the semantic ones (Agent, Patient, Recipient etc) using a lexicon of verb frames where each frame is organized according to these semantic roles. The lexicon contains verbs that can function as predicates in sentences and thus determine the possible semantic roles that can /must occur in corresponding sentences, and the task of the program is to match the units from syntactic trees with these semantic roles. The principles and general structure of our approach were described on the third Baltic HLT conference [9]. Thus far, we have restricted our research to the domain of motion, i.e. to sentences which express events where some entity changes its location.

Here we want to give a short overview of the main points of the development in our work (and in our understandings of what is crucial in the task of semantic analysis of sentences at the present stage; there will be a more detailed presentation at the conference: “Semantic analysis of sentences: the Estonian experience” by Õim et al [10]). These points can be summarized as follows: first, the organization of the frame lexicon; second, inferences as part of the meaning of a sentence; third, the role of ontological information (world knowledge) in sentence understanding

1. With respect to the frame lexicon the first thing to point out is that the frames in it are in fact not frames of verbs but frames of EVENTS represented/designated by the corresponding verbs: the central semantic unit in text semantics is not a word nor even a sentence but an event (in our domain of motion). The details of one such event (information about the fillers of the roles) can be picked up from different sentences but they should be collected and integrated into the frame of this individual event. For instance, let’s take a string of sentences (not necessarily in immediate succession in the real text): Yesterday, Mari went to Tallinn. This time she took her own car because she had to be in Tallinn very early. She left Tartu already at six o’clock. These sentences describe (pieces) of a concrete traveling event, but its different role fillers (AGENT—Mari, TIME—yesterday, INSTRUMENT—car, LOCFROM—Tartu, LOCTO—Tallinn, TIMEFROM—six o’clock) are given in different sentences (the role names in capital letters are from the list of our semantic roles).

The second aspect worth mentioning in connection with our frames is the use of so-called hidden arguments (as fillers of certain roles; this term—and the whole idea—we took from conceptual semantics, e.g. Jackendoff 2002 [11]). The idea is that some predicates incorporate in their meaning the information about the fillers of certain roles: e.g. walking and running imply that AGENT’s legs are used as the (immediate, bodily) INSTRUMENT, in the same way as seeing and looking imply the use of eyes. This information has not to be explicitly expressed in a sentence, unless something special is said about these instruments; and this specific information can come in another sentence, cf. He looked at me. His eyes were blue. Because of this the information about such “hidden” roles-fillers has to be included already into the frames of the corresponding verbs; and into the frame representations of concrete sentences, too, even when they are not explicitly given in the syntactic structure of the first sentence.

2. Inferences are a necessary part of the whole event expressed by a sentence. In our domain of motion most important inferences concern information about moving entities, especially, where the entity was located before the event and where it is after the event—to be able to answer such questions as “where was/is X?”. In our frames this problem has been solved by attaching corresponding rules to the roles/entities which move, using

the information from the roles LOCFROM (starting place) and LOCTO (end place). The point is that there are three critical roles whose fillers can move: AGENT, OBJECT and INSTRUMET. But in case of different verbs the entities in these roles move differently. Compare, for instance verbs like walk (AGENT moves), throw (OBJECT moves, but not AGENT), bring (AGENT and OBECT move, and if an INSRUMENT is used, it moves, too).

3.Ontological knowledge and its relationship to "pure" linguistic-semantic knowledge is becoming more and more important today and especially, of course, in modeling understanding of sentences and texts, but the solution of the problems starts in building lexical-semantic resources (see [12]) In case of sentence analysis the ontological information is connected, in particular, with the problem of inferences. To give just one example: when someone throws a stone onto a street we know (infer) that it will be there until we learn that somebody moved it somewhere else; but if somebody throws a stone into the air, we know (infer) that it will not stay there but falls back down. This is not connected with the frame of the verb throw; instead, these inferences are connected with our knowledge of what is a stone, what is a street, what is air. This is ontological knowledge about the corresponding entities and their possible interactions. We are dealing with these problems using the concept of qualia structure [11] but at present there is little to report about practical results.

In sum, what we have learned thus far in the field of sentence/text semantic analysis is:

1. when "already" in semantics (after the analysis of sentences in a text) we in a sense can forget about sentences and have to build semantic structures in terms of semantic units (in our case, events);
2. the events structures (and they are the structures which remain in our memory after reading a text) are not compiled from the information gathered from the analyzed sentences only;
3. in addition, we use inferences to fill in certain gaps in the event structure; and
4. we use ontological knowledge to do this.

Acknowledgements

This work is supported by the European Regional Development Fund through the Estonian Center of Excellence in Computer Science, EXCS and the National Programme of Estonian Language Technology.

References

[1] E.Meister, J.Vilo and N.Kahusk, National Programme for Estonian Language Technology: a pre-final summary. In *This volume*, (2010).

[2] A.Saareste, *Eesti keele mõisteline sõnaraamat I–IV. Dictionnaire analogique de laEstonienne I–IV*. Kirjastus Vaba Eesti, Stockholm, 1958–1968.

[3] P.Vossen, Eurowordnet: a multilingual database of autonomous and language-specific wordnets connected via an inter-lingual-index. *Semi-special issue on multilingual databases. International Journal of Linguistics*, (2004).

[4] G.Miller, R.Beckwith, C.Fellbaum, D.Gross and K.Miller, WordNet: An on-line lexical database. *International journal of lexicography*, 3(4), (1990) 235–244.

[5] N.Kahusk, K.Kerner and K.Vider, Enriching Estonian WordNet with Derivations and Semantic Relations. In *This volume*, (2010).

[6] C.Fellbaum, Towards a representation of idioms in WordNet. In S. Harabagiu (ed), *Proceeding of the Workshop on Usage of WordNet in Natural Language Processing Systems*. Montreal: COLING/ACL, (1998), pp. 52–57.

[7] N.Kahusk, Eurown: An EuroWordNet Module for Python. In *Principles, Construction and Application of Multilingual Wordnets. Proceedings of the 5th Global Wordnet Conference*. P. Bhattacharyya, C. Fellbaum, P. Vossen (Eds) Mumbai: Narosa Publishing House, (2010), pp. 360–364.

[8] W.Gale, K.Church, D.Yarowsky, One Sense Per Discourse. *DARPA Workshop on Speech and Natural Language*, New York, (1992), pp. 233–237.

[9] K.Müürisep, H.Orav, H.Õim, K.Vider, N.Kahusk, P.Taremaa, Fom Syntax Trees in Estonian to Frame Semantics. In *The Third Baltic Conference on Human Language Technologies. October 4–5, 2007*, Kaunas. Proceedings,Vilnius, (2008), pp. 211–218.

[10] H.Õim, H.Orav, N.Kahusk and P.Taremaa, Semantic analysis of sentences: the Estonian experience. In *This volume*, (2010).

[11] R.Jackendoff, *Foundations of Language*. Oxford: Oxford University Press, 2002.

[12] N.Kahusk, K.Kerner, H.Orav, Toward Estonian Ontology. In: *LREC 2008 Proceedings: LREC 2008*, Marrakesh; Maroko; 26. mai–1. juuni 2008. Eds: Oltramari, A. ; Prevot, L.; Huang, C.-R.; Buitelaar, P.; Vossen, P.. Elite Imprimerie, (2008), 20–24.

[13] K.Kerner, H.Orav, S.Parm, Semantic Relations of Adjectives and Adverbs in Estonian WordNet. In: LREC 2010 Proceedings: LREC 2010, Malta, Valetta, ELRA, (2010), 33–37.

[14] K.Kerner, K.Vider, Word Sense Disambiguation Corpus of Estonian. *The Second Baltic Conference on Human Language Technologies, April 4–5, (2005), Proceedings*, pp. 143–148.

[15] K.Vider, H.Orav, Concerning the difference between a conception and its application in the case of the Estonian wordnet. In: *Proceedings of the second international wordnet conference: Second international wordnet conference*; Brno; 2004. Ed. by Sojka, P.; Pala, K.; Smrz, P.; Fellbaum, Ch.; Vossen, P.. Brno:, (2004), 285–290.

[16] A.Õim, Sünonüümisõnastik, Tallinn, 1991.