Semantic Analysis of Sentences: The Estonian Experience

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

Abstract. This paper describes the work done on computational semantics of Estonian simple sentences. Main attention is paid to inferring knowledge from text. Differences from widely used FrameNet is discussed.

Keywords. Syntactic semantics, semantic roles, frame semantics, FrameNet, Estonian, inference

Introduction

For about 5 years we at Tartu University are working on a LT project called “Semantics of simple sentences”. Concerning the title of the project we would like to stress at once that our real goal is to move from sentence semantics to semantics of coherent texts, that is, to modeling of the “real” process of human language understanding, so that what we call now semantic representation of a sentence or text in fact would be a representation of what the reader/hearer *knows* after having read/heard this text. This should explain why we in our project, when speaking of semantic analysis of simple sentences (= translating syntactic trees of isolated sentences into some kind of semantic structure), quite intentionally deal with such theoretically-oriented problems as drawing different kinds of inferences or using ontological knowledge.

In the analysis process itself we really use as input the syntactic dependency trees of sentences from the Estonian Treebank[2] and the output is the representation of the sentence in the form which we call a (sentence) frame where the syntactic dependencies (Subject, Object, Adverbial etc) are replaced by semantic roles (Agent, Instrument, Recipient etc; see e.g. [1]). But this is the first step only. And even this is not that simple as it seems.

The main problems we have found to be of critical importance can be summarized as follows: First, compiling the inventory of semantic roles; because at the present state of semantics there is little hope to create a universal inventory, we have to restrict ourselves to some semantic domain. At present the domain of our semantic analysis program is motion—self-motion as well as caused-motion events.

Second, in the process of transition from a syntactic tree of a sentence to its semantic frame (representing the corresponding event) it often appears necessary to add so-called

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

[2] www.clarin.eu/arborest

hidden arguments/roles (in the sense of Jackendoff's conceptual semantics), i.e. arguments that do not always appear in the surface sentence as syntactic elements but will be needed in its semantic representation, e.g. when some specific information has to be added to their description later (e.g. walking and legs, throwing something somewhere and hands, looking-seeing and eyes—as implicit Instruments).

Third, the problem of inferences: the full meaning of a sentence includes, for the recipient, not only the data explicitly represented in it but also the knowledge s/he can derive from these data by means of inferences; and this is particularly important when we start to model the understanding of coherent texts where the knowledge derived from the previous sentences by inferences cannot be distinguished from the information explicitly conveyed.

And fourth, there is the need to take into account, along with "pure" linguistic meaning of language expressions, also the world knowledge (domain ontology).

Since it is clear that in a short overview it would not be possible to treat all these problems at the reasonably informative level we will concentrate below on the kernel of our system, the frame lexicon; by describing its organization it is possible to show also how it helps to solve other problems, e.g. the problem of inferences.

1. Frame lexicon, semantic roles, inferences

Frames in our system are structures consisting of a head—a motion verb which in a sentence can function as predicate—and its possible arguments as fillers of certain semantic roles. Thus, semantic roles are the main structuring elements of a frame. The original idea behind the concept of frame came, of course, from frame semantics and specifically from FrameNet[3] (see e.g. [2] for overview). But for purposes of our project which deals with the interaction between syntax (sentences, texts) and semantics we had to work out our own inventory of semantic roles. One reason for this was, for instance, the need to draw inferences from frames: FrameNet does not deal with inferences, at least not explicitly. But in case of semantic analysis of sentences it is impossible to ignore this problem; and certain kinds of inferences are directly connected with semantic roles (we will discuss the problem below). This, by the way, does not mean that FrameNet structures cannot be used to draw inferences from sentences with e.g. motion verbs as predicates. We have tried this, in parallel with our frame structures. But the role inventory in FrameNet is too complicated and domain-dependent to be taken as a regular basis of sentence/text semantic analysis program at the very beginning.

The first conceptually important point we want to make clear is that although the heads of frames are verbs, the frames are in fact not frames of verbs but frames of EVENTS represented/designated by the verbs as possible predicates of corresponding sentences. The basic 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 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:

> 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.

[3] http://framenet.icsi.berkeley.edu/

These sentences describe (pieces) of a specific traveling event the frame of which is evoked by the verb 'went' in the first sentence, 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 idea of this differentiation and, concretely, the concept of event in our case we have taken from Conceptual Semantics [3] where the complex problems of word-sentence-text semantics (including the background knowledge) are dealt within a common framework. Such complex treatments are quite rare in today's theoretically oriented linguistics.

Apparently, the most direct way to explain our ideas connected with frames and their structure—semantic roles and inferences—in the analysis and representation of the meaning of sentences would be to use a concrete example. Below, we give (in basic details) the frame structure of agentive self-motion (AGENTIIVNE LIIKUMINE) represented by verbs like *kõndima* 'to walk', *lendama* 'to fly' (like a bird) *ujuma* 'to swim', *sõitma* 'go (using a vehicle), travel, ride. . . ' and then explain the reason of its structural elements. See Fig 1 for an overview of the general role structure.

There are two features in the structure of this frame that are of importance here and need explanation.

First, the ASETSEMA-subframes are attached to the roles whose fillers move in the event described by the frame. In the agentive self-motion event AGENT and INSRUMENT are the entities that move. ASETSEMA_1 and ASETSEMA_2 fix the location of the entity before and after the motion event, accordingly, taking the corresponding information from the LOCFROM and LOCTO roles of the main frame. Thus, this is our present (preliminary) solution to the problem of inferences concerning the location of entities participating in a motion event before and after the event. The reason why we have chosen such a straightforward solution is that in different motion events different participant move. For instance, in case of an agentive caused-motion event where AGENT throws an OBJECT from place L1 to place L2, only OBJECT moves to L2, AGENT stays at L1, and therefore in the frame corresponding to predicate throw ASETSEMA1/2 subframes are attached only to OBJECT role and not to AGENT. But in an event where AGENT brings an OBJECT from L1 to L2, both OBJECT and AGENT move, and if AGENT uses an INSTRUMENT, it moves, too. And therefore ASETSEMA1/2 subframes should be attached all these roles in the bring frame.

The second feature which needs explanation is the use of in- and at-subroles by LocRoles. It may be remarked at the outset that this is also connected with the problem of inferences, but in quite different way; and it brings in the ontological dimension. The critical point here is that in our folk ontology of the world we differentiate, among other aspects, between entities that have inside and those that do not, the difference being, that other object can be moved into the first ones (and kept there), but not into the second ones. For instance, bags, baskets, boxes, boats, cupboards, have inside, but stones, chairs, trees, etc do not in the same sense. Of course, there is an indefinite number of entities in case of which this difference simply does not make sense. In the context of motion domain this difference appears relevant in the following way. Both kinds of entities can function as fillers of the role LOCTO, that is, as reference points of where the motion ended. But there is a principal difference, in case of "entities with inside", whether the moving object moved into them (like in sentence 'I put the shoes in the basket') or somewhere near it ('I put the shoes behind the basket'). The difference becomes important, among other

```
AGENTIVE SELF-MOTION
HYPERONYM: MOTION

ROLE STRUCTURE
  Participant Roles
    AGENT (participant who controls his/her activity,
  the instigator of the event)

    FRAME: ASETSEMA_1 ’be located’
      Object:  = Agent
      Loc = Locfrom
      Time = Timefrom
    FRAME: ASETSEMA2
      Object = Agent
      Loc = Locto
      Time = Timeto

  INSTRUMENT
  [the same ASETSEMA subframes attached as by AGENT,
    only Object = Instrument,
    which means that INSTRUMENT is supposed to move
    the same way as AGENT]

  Loc-Roles
    LOCFROM (starting place, e.g. from the garden,
      from under the table, from the box)
      Locfrom-in
      Locfrom-at
   LOC (where the motion takes place, e.g. on the street,
in the garden, under the table)
      Loc-in
      Loc-at
   LOCTO (the ending place, e.g. onto the street,
    into the garden, into the box)
      Locto-in
      Locto-at

  /---/

  Time-roles
    [The same system: TIMEFROM, TIME, TIMETO, DURATION]
  /---/
    Other roles
      Not important in the given context: DIRECTON, PATH,
      MANNER, about 30 in total.
```

Figure 1. Frame structure of agentive self-motion.

things, when in the later text it is said that the corresponding entity with inside (basket in our examples) moves to another place (e.g. is taken somewhere). Then, by inference, one (e.g. the computer program) should conclude that all things that were in it (e.g. my shoes) are also at this place. But things that were ‘at“ it (behind, before, etc) have not moved. Of course, this concerns a very specific aspect of the motion, but our intention was just to demonstrate that once we start a serious task of semantic analysis of sentences and (coherent) texts we cannot avoid ”landing“, former or later, at such specific problems.

The last aspect we would like to touch 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). 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 walked to the table. He was barefoot.’ 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 the predicate of which triggered the frame (walked in the given case). In the corresponding frame under the role in question such information should explicitly formulated. For instance when we take the Estonian verb *kõndima* ‘to walk’ then under the role AGENT as one of the semantic requirements to its possible fillers should be given ”has legs,“ e.g.:

```
KÕNDIMA
  AGENT
  SEMREQ: Living being
  HAS_BODYPART: legs
```

And there should be in the frame the specific INSTRUMENT role by which is the information that as this instrument function the legs of the AGENT:

```
INSTRUMENT-B[odypart]
  Legs = BODYPART-of-AGENT
```

Summary

Our main aim was here not give technical details of our project but to give an outline of the solutions to problems we consider critical in the semantic analysis of text and, further, of coherent texts. We described our approach to them: in the center of it is the frame lexicon, where the principal elements are semantic roles, including ”hidden“ semantic roles. And second, the treatment of inferences, which constitute an inevitable part of sentence semantics.

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] Müürisep, Kaili; Orav, Heili; Õim, Haldur; Vider, Kadri; Kahusk, Neeme; Taremaa, Piia (2008). 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.

[2] *International Journal of Lexicography*. Special issue. 16 (3) Sept., 2003. Thierry Fontenelle, ed.

[3] Jackendoff, Ray 2002. *Foundations of Language*. Oxford: Oxford University Press.