The Second Baltic Conference on HUMAN LANGUAGE TECHNOLOGIES, April 4 - 5, 2005

WORD SENSE DISAMBIGUATION CORPUS OF ESTONIAN

Kadri Kerner, Kadri Vider
University of Tartu (Estonia)

Abstract

The research group of computational linguistics of the University of Tartu has developed Word Sense Disambiguation Corpus of Estonian (WSDCEst). During four years 100 000 running words are looked over and all content words in texts are manually annotated according to EstWN word senses. The source texts are mostly fiction. Paper gives quantitative analysis of corpus and focuses on some inspiring and linguistically relevant ideas and hypoteses.

There were significant inconsistencies in opinions of these people, who disambiguated the texts. This shows us the most problematic entries in EstWN, the need to reconsider the borders of meaning of some concepts. Part of our research focuses on exploiting agreement and disagreement of human annotators: are there any remarkable and important sense clusters.

Manual sense tagging refers to problems in EstWN like missing examples, overlapping synsets of explanations and over-grained senses. Sense clusters are made by processing the disagreement files and the most frequent words are analyzed by describing lexical relations like autohyponymy and sisters (co-hyponyms).

Keywords: word sense disambiguation, semantic annotation, corpora, Estonian Wordnet, sense clusters

1. General[1]

For several language technology applications, it is important to make sure in which sense each word is meant. Demand for systems able to resolve this problem originated SensEval – international organization devoted to the evaluation of Word Sense Disambiguation Systems (SensEval). Part of Word Sense Disambiguation Corpus Estonian was distributed as Gold Standard as well as Test Corpus for Estonian task in Senseval-2 competition in 2001 (Kahusk et al. 2002).

The problem of semantic disambiguation (tagging and annotation) is tightly connected to morphological and syntactic disambiguation, but is more complicated. It is even argued, that the concept of a word sense is questionable and depends on goals of

[1] This paper is based on work supported in part by the Estonian Science Foundation under grant No 5534 "Concept based resources and processing tools for the Estonian language" and by Estonian State Target Financing R&D project No 0182541s03 "Computational models and language resources for Estonian: theoretical and applicational aspects" and by Governmental programme "Eesti keel ja rahvuslik mälu" sub-project "Language Technology: Semantic analysis of Estonian simple sentence".

disambiguation (Kilgarriff 1997). However is word sense disambiguation (WSD) task of interest in lexicography and lexical semantics.

2. Sense-tagged corpora

There are two main approaches one can take to the order in which words are tagged in texts (Langone et al. 2004; Kilgarriff 1998). In the sequential approach (also termed 'textual' or 'all-words-task') annotator tries to assign the context-appropriate sense to each open class word as it is encountered. The targeted approach (also termed 'lexical' or 'lexical sample task') involves tagging all corpus instances of a pre-selected word.

Some semantically annotated corpora, e.g. SEMCOR in English, are tagged sequentially and we chose same approach for WSDCEst. Other corpora, for example HECTOR in English, use lexical choice and receive very detailed sense-distinctions of particular word. It is good to know, that SEMCOR as well as HECTOR corpora use English WordNet sense distinctions.

It should be kept in mind, that not all words can be disambiguated, but only content words. Although normally nouns, verbs, adjectives and adverbs are considered as content words (see e.g. Stevenson, Wilks 2001), in WSDCEst only nouns and verbs were subject to disambiguation. We use EstWN as sense distinction resource (see Kahusk et al. 2005 in this volume), but adverbs are not represented in EstWN yet, and there is too little number of adjectives.

3. Multilevel approach to WSD

Word sense disambiguation is closely connected to morphological and syntactic disambiguation. Stevenson and Wilks (2001) propose multilevel approach to WSD. Semantic, sometimes even pragmatic information can be derived from hypernymy hierarchies, and syntactic information can be read from morphological analysis.

3.1. Morphology

Lexical entries (literals) in EstWN are presented nominal singular form for nouns and supine form for verbs. In real texts, the words are mostly in their full richness of forms. Lemmatizing and part-of-speech-tagging are made with Estmorf tagger (Kaalep 1997). In sense annotating we considered only nouns (_S_ com) and non-auxiliary verbs (_V_main or _V_ mod).

The modal verbs are explicitly marked in the output of the morphological disambiguator (_V_ mod). When a verb is marked as such, then the senses that don't correspond to the modal senses could be removed, e.g. the word 'saama' has all together 12 senses in the thesaurus, but only 2 of them ('can' or 'may') correspond to the modal use of the word.

The output of the morphological analyzer often contains valuable information for word sense disambiguation. In some cases the word-form used in the text can uniquely specify the sense of the word, although its lemma is ambiguous, e.g. the word 'palk' can either mean salary or log of a tree, but its genitive form is different in each meaning (either 'palga' or 'palgi'). By using only the lemma we ignore this distinction that can be explicitly present in the text.

3.2. Syntax

At the moment the input text contains no information about its syntactic structure, most importantly the verbal phrases and other multi-word units are not marked as such. Also, the syntactic structure can help to reduce the number of possible senses to choose from. For example the most frequent word 'olema' (be, have) has five more frequent senses. Only one sense is present in complementary clauses; 3 senses appear in existential sentences and one in possessive sentences. Linguistic knowledge about the nature of the sentence can help the disambiguation process of human annotator.

4. Texts for WSD Corpus

We chose 43 texts for word sense disambiguation from Corpus of the Estonian Literary Language (CELL) subcorpus of Estonian fiction from 1980s. Each text file contains about 2500 tokens. Most of the texts that are annotated for word senses, are fiction. Total amount of tokens in texts is around 110,000 (depends on calculating punctuation in or out) at present. About 34,5% of them (see Table 1) are annotated content words, whereas its impossible to disambiguate word senses without context we counted items of other part-of-speeches together.

Table 1. Words and senses in WSD Corpus of Estonian

	Nouns (S com)		Verbs (main and modal)	
	Total	Mean per text	Total	Mean per text
Tokens	21373	497,05	17947	417,37
Lemmas	6536	311,98	1649	177,51
Lexical entries found in EstWN	2585	200,07	1261	160,51
Polysemous words	-	223,65	-	227,07
Senses per annotated lexical entry	-	1,12	-	1,42

5. Manual annotation

Twelve linguists and students of linguistics tagged nouns' and verbs' senses in the texts, each text was disambiguated by two persons. Pre-filtering system added lexeme and number of senses for each annotating word found in EstWN. Annotators marked in brackets the sense number of EstWN which matched best with used sense of a word by their opinion. If the word was missing from the EstWN, "0" was marked as sense number, and if the word was found in EstWN, but missed appropriate sense, "+1" was marked. Example (1) presents a sentence „*Neid kentsakaid mõtteid põimin jälle suvel oma kirjutistesse*" (I'm going to weave these weird thoughts into my writings in summer again.) as it occurs in WSD Corpus of Estonian.

If inconsistencies were met, they were discussed until agreement was achieved. On about 20% of cases the disambiguators had different opinions. This shows us the most problematic entries in EstWN, the need to reconsider the borders of meaning of some concepts. Part of our research (Kerner 2004) focuses on exploiting agreement and

disagreement of human annotators: are there any remarkable and important sense clusters.

(1) <s>
Neid
see+d //_P_ dem pl part //
kentsakaid
kentsakas+id //_A_ pos pl part //
mõtteid
mõte+id //_S_ com pl part // **mõte(1)#@5**
põimin
põimi+n //_V_ main indic pres ps1 sg ps af // **põimima(3)#@3**
jälle
jälle+0 //_D_ //
suvel
suvi+l //_S_ com sg ad // **suvi(1)#@1**
oma
oma+0 //_P_ pos sg gen //
kirjutistesse
kirjutis+tesse //_S_ com pl ill // **kirjutis(1)#@1**
.
. //_Z_ Fst //
</s>

6. Sense clusters

Sense clusters are made up by processing the disagreement files. The most frequent words are analyzed by looking over different sense numbers that annotators proposed as in Table 2.

Table 2. Sense clusters of HAKKAMA

Combination of sense numbers	Frequency
2 -- 3	17
2 -- 5	10
2 -- 6	9
3 -- 5	3

There is little disagreement among word senses that doesn't include autohyponymy and/or sisters (co-hyponyms). Also a very important observation is that human annotators disagree less when all the representation fields of EstWN are properly filled (hyperonym (s), synset, definition, explanation). The research referred to the fact that explanations seem to be very important for human annotators. When adding missing explanations, the difference between senses becomes more definite. The fact that some words are not highly polysemous indicates usually (but not always) to a minor disagreement of human annotators.

Manual sense tagging refers to problems in EstWN like missing examples, overlapping synsets or explanations and over-grained senses. In many cases it is

impossible to determine the one and only sense. Sometimes it is even not necessary (Vider et al 2003: 316-317) and sometimes the nearby context allows different senses.

It is difficult to distinguish word senses that are detectable in EstWN but not visible in the real usage of text (or language). In some cases the disagreement between human annotators arises, because of the lack of lexicographical knowledge (or the human annotator is somewhat superficial).

Exploiting sense clusters can be helpful in referring to insufficiency of EstWN. For example, if all the sense numbers of a word combine with each other (Table 2), it can be assumed that the distribution of senses is incomplete and needs to be improved. If there is no disagreement among human annotators, then there are no remarkable sense clusters and therefore the senses of this particular word are reasonably distributed (or divided). Also our research showed that words with abstract meanings are difficult to annotate and make up essential sense clusters.

Some researchers (Vider et al. 1998; Vossen et al. 1998:6) claim that words representing so-called Base Concepts are difficult to annotate semantically (apparently because of their broad meanings). This research also confirmed this fact. The boundaries and the area of the usage of a hyperonym or hyponym should be very precisely represented in EstWN. The tendency seems to be that hyponyms as narrower meanings are better to distinguish than hyperonyms. That is the reason, why top concepts tend to combine with many of the different word senses (example in Table 3).

Table 3. Sense clusters of SAAMA

Combination of sense numbers	Frequency
10 -- 11	24
10 -- 9	9
10 -- 2	8
10 -- 6	5
10 -- 3	4
10 -- 7	4

Importance for automatic WSD

Important sense clusters can be effective for improving the processing work of word sense disambiguation system. It is easier for this system to choose the appropriate sense if the word senses are not too over-grained; the speed and accuracy of the system will increase. If there are any remarkable sense clusters, it might be useful for the applications of language technology to join these senses. For example, in a machine translation system – if a human annotator can not distinguish all the senses of a particular word, then maybe the machine translation system also should not.

References

CELL = Corpus of the Estonian Literary Language. Retrieved February 15, 2005, from http://test.cl.ut.ee/korpused/morfkorpus/index.html.en

Kaalep, Heiki-Jaan 1997. An Estonian morphological analyser and the impact of a corpus on its development. *Computers and the Humanities*, 31, 115–133

Kahusk, Neeme; Orav, Heili; Õim, Haldur 2002. Sensiting inflectionality: Estonian task for SENSEVAL-2. In: *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, 25-28

Kahusk, Neeme; Vider, Kadri 2005. TEKsaurus – the Estonian WordNet online. *This volume*.

Kilgarriff, Adam 1998. Gold standard datasets for evaluating Word Sense Disambiguation programs. *Computer Speech and Language*, 12(3), 453–472.

Kilgarriff, Adam 1997. "I don't believe in word senses." In: *Computers and the Humanities*, 31(2), 91–113.

Kerner, Kadri 2004. Sõnatähendused tekstides ja tesauruses ühestajate erimeelsuste põhjal. (English title: Word senses in texts and in thesauri based on human annotators disagreement) B.A. thesis. (Manuscript.) University of Tartu, Dept. of General Lingustics.

Langone, Helen; Haskell, Benjamin R.; Miller, George A. 2004. Annotating WordNet. In: Meyers, A. (ed). *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*. Boston: Association for Computational Linguistics. 63–69

SensEval = Senseval web page. Retrieved February 28, 2005 from http://www.senseval.org/

Stevenson, Mark; Wilks, Yorick 2001. The interaction of knowledge sources in word sense disambiguation. In: *Computational Linguistics* 27 (3), 321–349.

Vider, Kadri; Orav, Heili 2003. Idee ja rakenduse vahe tesauruse näitel. In: *Eesti Keele Instituudi toimetised 14. Toimiv keel I. Töid rakenduslingvistika alalt*. Tallinn: Eesti Keele Sihtasutus. 313–322.

Vider, Kadri; Orav, Heili 1998. Sõna tasandilt mõiste ruumi. In: *Keel ja Kirjandus* 1. 57–64.

Vossen, Piek; Kunze, Claudia; Wagner, Andreas; Dutoit, Dominique; Pala, Karel; Sevecek, Pavel; Vider, Kadri; Paldre, Leho; Orav, Heili; Õim, Haldur 1998. *Set of Common Base Concepts in EuroWordnet-2*. Amsterdam: Deliverable 2D001, WP3.1, WP 4.1; EuroWordNet, LE4-8328.

KADRI VIDER is a researcher and a PhD student at Department of General Linguistics, University of Tartu. She received her M.A. in 1999 dealing with senses of Estonian verbs in semantic database such as wordnet. Her research interests concern computational lexicology, lexical semantics and word sense disambiguation. Her doctoral study focuses on senses of Estonian verbs and possibilities to distinguish them in texts. She is member of the board of the Estonian Association of Applied Linguistics. e-mail: kadri.vider@ut.ee

KADRI KERNER is a M.A student at Department of General Linguistics, University of Tartu. She received her B.A in 2004 with the graduation thesis: Word Senses in Texts and in Thesauri based on Human Annotators Disagreement. Her M.A thesis deals with sense clusters of Estonian nouns, grammaticalization, Estonian Wordnet and FrameNet. e-mail: kadri.kerner@ut.ee