Sensiting inflectionality: Estonian task for SENSEVAL-2

Neeme Kahusk and **Heili Orav** and **Haldur Õim**
University of Tartu
Research Group of Computational Linguistics
Tiigi 78, 50410 Tartu, Estonia
{nkahusk,horav,hoim}@psych.ut.ee

Abstract

This paper describes the all-word sense disambiguation task provided by Estonian team at SENSEVAL-2. About 10,000 words are manually disambiguated according to Estonian WordNet word senses. Language-specific problems and lexicon features are discussed.

1 Introduction

We got interested in word sense disambiguation (WSD) for two reasons. First, already a couple of years ago it was evident that WSD is becoming one of the new "hot" topics in computational linguistics and language engineering as our knowledge of how to handle semantic parameters of texts and semantic features of words in texts increased. The second reason was purely practical. Since 1996 we have been involved in a large project of building a semantic database of Estonian; participating in the EuroWordNet project has been a part of it (but a very important part, of course). The main source of building this database have been different corpora of Estonian, and in working with corpora the question of whether we are dealing with different meanings of a word in case of its concrete occurrences or not arises constantly. So we got interested in the possibility to use some objective methods here.

Our task was all-words task. This choice is explained with our "practical" interests explained above.

A large amount of work was done to provide training data where disambiguation was done manually. The same kind of work had to be done with test data, of course. The description of this work is given below. Let us note already here that this work appeared to be very useful and informative for us as builders of Estonian WordNet (EstWN).

And let us stress that this was our first attempt of WSD at all.

2 Corpora and lexicon

The test and training texts come from Corpus of the Estonian Literary Language (CELL), the 1980-s. We used this part of the corpus, that was morphologically disambiguated, initially for the syntactic analysis.

The morphological analysis was made with ESTMORF (Kaalep, 1997). Lemma and word class in the output of the program are relevant to our task, but it is impossible to get them without morphological disambiguation, because of frequent homonymy among word forms.

All training texts and most of test texts (5 of 6 total) are fiction. One of the test texts is from newspaper. Six training and six test files provided for the task contain about 2000 tokens each. More information about the texts used in the task is in Table 1.

Table 1: Statistics on training and test corpora

Corpus	Training	Test
Total words	12162	11440
Words to disambiguate	5854	5650
of them being		
verbs	2431	2191
nouns	3423	3459

2.1 Lexicon

The Estonian part of EuroWordNet[1] served as the lexicon. Like other wordnets, EstWN is a lexical-semantic database, the basic unit of which is concept. Concepts are represented as synonym sets (synsets) that are linked to each other by semantic relations. The description of

[1] http://www.hum.uva.nl/~ewn/

EstWN is given in the final document of EuroWordNet (Vider et al., 1999).

EstWN is supposed to cover the Estonian base vocabulary in its initial version. The base vocabulary will be determined by statistical analysis of the reference corpus. Even so it is not always easy (nor appropriate) to stop encoding words with frequencies below a certain threshold. For this reason we expect EstWN to cover more than just the base vocabulary.

Still the EstWN is rather small, there were 9436 synsets, 13277 words and 16961 senses (literals) in it when the disambiguation was done. That makes about 1.28 senses per word as average.

Most of synsets are connected with hyperonym-hyponym relations building corresponding hierarchies.

2.2 Procedure

Four linguists disambiguated the texts, each text was disambiguated by two persons. Only nouns and verbs were disambiguated, as entering adjectives into EstWN is in the very beginning. The sense number was marked according to sense number in EstWN. If the word was missing from the EstWN, "0" was marked as sense number, and if the word was in EstWN, but missed the appropriate sense, "+1" was marked.

If inconsistencies were met, they were discussed until agreement was achieved. On about 28% of the cases the disambiguators had different opinions.

One of the problems that the disambiguators ran into concerned dividing words into different senses in EstWN. It turned out as over-differentiation—word meaning marked as too specific, or over-generalisation—word meaning marked as too general.

2.3 How much the lexicon covers

Not all senses found in EstWN are represented in texts. Maximum number of senses per word found in texts is 13. This is more than appropriate senses in lexicon (see Table 3), but we must remeber about the "+1" that disambiguators had, if they found that there are not enough meanings in EstWN. Table 2 describes distribution of senses in usage and Table 3 shows the top of lemmas according to number of senses.

Table 2: Distribution of lemmas according t number of senses in texts

Corpus	Training	Test
Total number of lemmas	2340	2268
Number of lemmas not in lexicon	819	948
Number of lemmas with 1 sense in texts	2040	2003
Lemmas with 2 senses in texts	215	183
Lemmas with 3 senses in texts	51	50
Lemmas with 4 senses in texts	17	17
Lemmas with more than 4 senses in texts	17	15

Table 3: Comparison of richest words in sense

POS	No of senses in text	Lemma	No of senses in lexicon
verb	13	saama	12
verb	10	pidama	12
noun	10	asi	11
verb	9	olema	9
verb	9	käima	23
verb	7	võtma	7
verb	7	panema	11
verb	7	nägema	7
verb	7	minema	17
verb	7	leidma	8
noun	7	elu	7

It would be the best, if all words to disambiguate were in the lexicon with all their possible meanings. Apparently this presumption is not met.

The number of compounds in Estonian is indefinite. It is quite easy for a writer to invent new compounds that are not in any dictionary, but nevertheless are easily understood by readers. That is one reason, why there are so many sense numbers "0" in the texts. About 46 % of words that are not in EstWN, are compounds.

Another remarkable class of words not in lexicon are proper names, as there are no proper names in EstWN. There are 17.5 % of words proper names.

If we will postpone phrasal verbs and some strange words that contain hyphens (about

7 %), it leaves us with about half thousand words to check why they are not in EstWN.

But why are there missing senses (tagged with "+1")? The reason is simply historical: such words were included into EstWN as synonyms of some base vocabulary word and the other senses of them are not considered yet.

2.4 Phrases and multi-word units

The initial format of text was as it came from ESTMORF and semantic disambiguation: every word on separate line, followed by an aditional line of morphological analysis and sense number, with multi-word phrase marked if word was part of it. The task to convert into Senseval XML format seemed trivial at first, but phrases turned out to be problematic. Unfortunately enough, all the story about phrases is concerning the training corpus only, because in test corpus the multi-word phrases were unmarked.

Estonian is a flective language with a free word order and that makes it complicated to figure out all phrases. The elements of a phrase can be scattered around the sentence in an unpredictable order.

In the initial texts, the disambiguators marked down the whole phrase on the line where the phrase occured. They were not told to mark it on each line, where the non-disambiguatable parts of the phrase were, and it happened that the phrase was not marked on the line, where the head of the phrase was. The algorithm of calculating head or satellite took into account the part of speech and the form. For verb phrases, if both components were verbs, declinable form of verb infinitive was marked as satellite. For noun phrases, substantive makes head and adjective satellite. If both words are substantives, head is the second one... well, mostly.

However, it is known that expressions tend to contain frozen forms, including inflectional endings. For example, one may not say "*Human Right" or "*Humans Right". "Human Rights" is the only correct expression and should be added into thesauri in such form. Phrasal verbs like "ära maksma" (to pay off) and idiomatic verbal expressions like "end tükkideks naerma" (to laugh oneself into pieces) represent a situation that is different from the occasion described above: the verb part may inflect freely, but the other word(s) are frozen forms. Hereby, even if we have determined what is phrase or collocational multi-word unit, we still have a question— are they commonly used and should we add them into the lexicon.

Multiword expressions are included into EstWN if they build up a conceptual unit and are commonly used as lexical units.

3 Results

There were two systems to solve the task on Estonian. The results are in Table 4. Table 5 shows the recall and precision of the COMMONEST baseline

Table 4: Estonian all-words fine-grained scoring results

System	Precision	Recall	Attempted
JHU	0.67	0.67	100
est-semyh	0.66	0.66	100

Table 5: COMMONEST baseline for Estonian all-words task

Data	Recall	Precision
Overall	0.85	0.73
Polysemous	0.69	0.51

As this is the first attempt to disambiguate Estonian nouns and verbs in text, there is no comparison data. These results will set the level that future systems will try to outgo.

4 Conclusions

Results of WSD of corpus texts turned to be a good way to add missing synsets and senses into our wordnet. 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. By now, the last version of EstWN contains 9524 synsets, 13344 words and 17076 senses.

For an inflectional language like Estonian, morphological analysis is extremely important and morphological and semantic disabiguation can help each other.

References

H.-J. Kaalep. 1997. An estonian morphological analyser and the impact of a corpus on its development. *Computers and the Humanities*, 31:115–133.

K. Vider, L. Paldre, H. Orav, and H. Õim. 1999. The Estonian Wordnet. In C. Kunze, editor, *Final Wordnets for German, French, Estonian and Czech.* EuroWordNet (LE-8328), Deliverable 2D014.