Automatic WSD: Does it make sense of Estonian?

Kadri Vider and **Kaarel Kaljurand**
University of Tartu
Department of General Linguistics
Tiigi 78, 50410 Tartu, Estonia
kvider@psych.ut.ee and kaarel@ut.ee

Abstract

This paper describes a fully automatic Estonian word sense disambiguation system called *semyhe* which is based on Estonian WordNet (EstWN) hyponym/hypernym hierarchies and meant to disambiguate both nouns and verbs.

1 Short description of the system

The main inspiration for our system is Agirre and Rigau (1996) similar system that disambiguates the English noun senses based on WordNet hyponym/hypernym hierarchy, taking into consideration the distances between the nodes corresponding to the word senses in the WordNet tree as well as the density of the tree. They have also experimented with using meronyms/holonyms in addition to hyponyms/hypernyms but report that it does not improve the results.

Our main object was not to focus on the homonymous words only (lexical sample), but to try to disambiguate all nouns and verbs in the text. The Estonian WordNet (EstWN) also contains adjectives but they are not linked by hyponym/hypernym relations. The word sense disambiguation could also try to describe a unique sense for adverbs but in our case such words have not yet been included in the thesaurus.

As far as we know this is the first attempt on automatic Estonian word sense disambiguation.

1.1 Input

The input text for our system must be morphologically analyzed, meaning that each word is provided with its lemma and morphological reading. Taking those two into account we can localize the senses that correspond to the word in EstWN hyponym/hypernym tree (Vider et al., 1999). It must be mentioned that although the morphological description in the input can be quite detailed, we only use the information on whether the word is a noun or a verb.

A simple morphological analysis that only looks at the word-form and not its context can result in very ambiguous output. On average 45% of the words are morphologically ambiguous in Estonian texts (Kaalep, 1997). The ambiguity can be greatly reduced by also applying the Estonian morphological disambiguator (Kaalep and Vaino, 1998) to the text before the word sense disambiguation. Since even then the words can in principle stay morphologically ambiguous, our system doesn't require each word to have exactly one morphological reading assigned to it in the input text.

1.2 Output

Similarly to the morphological analysis, we do not try to provide each word with exactly one sense. In case two (or more) senses have equal evaluation results then both of those prevail in the output.

1.3 Sense disambiguation algorithm

We apply the exact same algorithm for both nouns and verbs. Nouns and verbs cannot be compared with each other since in terms of hyponym/hypernym hierarchy they are located in different trees in the thesaurus. So the disambiguation is carried out in 2 runs, first nouns are disambiguated, then verbs, or vice versa.

A window is shifted on the text and as a word moves through the window its senses are compared with the senses of other words in the window. The context is either made out of nouns or verbs depending on which part of speech is being disambiguated.

The basis of the comparison is the similarity between the senses which is defined through the notion of conceptual distance, the distance between the nodes corresponding to the senses in EstWN tree. Winners are the senses that minimize the total distance between the word senses in the window, all the rest are removed from the list of candidates for the correct reading. *semyhe* leaves the word ambiguous when there are more than one senses with equal result. This usually happens when the senses of the context words are located in different hierarchies and hence can not be compared. Currently there are 108 different top nodes in EstWN, 29 corresponding to nouns and 79 to verbs.

In addition, the work of the system can be modified via several options in the configuration file:

- The window-size can be changed, increasing it makes the output less ambiguous since there is a higher possibility that the comparable senses end up in one window. On the other hand a bigger window may span across several sentences making the compared words possibly irrelevant to each other. For the moment we have used window of 5 words.
- Since we use no syntactic analysis before word sense disambiguation, the context of any word under observation is unstructured, the only syntactic information that we can use is therefore only the distance of the words from each other in terms of running text. A set of weights can be defined that is mapped to the distances, so that the similarity of the senses of the words that are far away from each other is less relevant for the total score.
- We can also take into account the average depth of the compared nodes in the tree — the bigger the depth the more reliable the score.

So far we haven't yet experimented with any of those options much.

2 Analysis of the results

For the purposes of analyzing the quality of disambiguation, tests were made with 12 manually sense tagged texts. These text samples were mainly from fiction, in a part also from newspapers and they contained approximately 10,000 tokens that corresponded to either nouns or verbs.

Manual tagging naturally had to remove all the morphological ambiguity, therefore the results obtained on those texts should be better than on the texts that have only been treated automatically before the word sense disambiguation. Words that occurred in the text but were not present in EstWN were marked as having 0 senses, approximately 30% of such words are proper names.

Manual tagging also recognized multi-word units, which in our case are mostly noncontiguous verbal phrases that are hard to detect automatically even if we had a complete list of such units.

Using *semyhe*, we set out to disambiguate all the nouns and verbs contained in the texts. Since *semyhe* can leave a word ambiguous, it makes sense to evaluate its work in terms of recall and precision. Table 1 lists *semyhe* results when the window of context words has size 5. The table also shows the results obtained with a random method which chooses exactly one sense for every word randomly (in this case recall and precision have equal values).

The row groups of the table refer respectively to the results with polysemous words and the overall results. Note that the words which were manually marked as having 0 senses were considered monosemous and so they are always correctly analyzed, with unique sense selected for every word.

	POS	recall	precision	random
polysem	nouns	0.543	0.347	0.423
	verbs	0.495	0.249	0.251
	both	0.514	0.283	0.292
overall	nouns	0.839	0.700	0.773
	verbs	0.601	0.338	0.412
	both	0.745	0.522	0.630

Table 1: *semyhe* results with 10,000 nouns and verbs

Figure 1 shows the distribution of words between different number of senses according to those texts. This shows the ambiguity that any automatic analysis has to cope with.

Note that there is an unusually large number of words with 9 different senses. The main

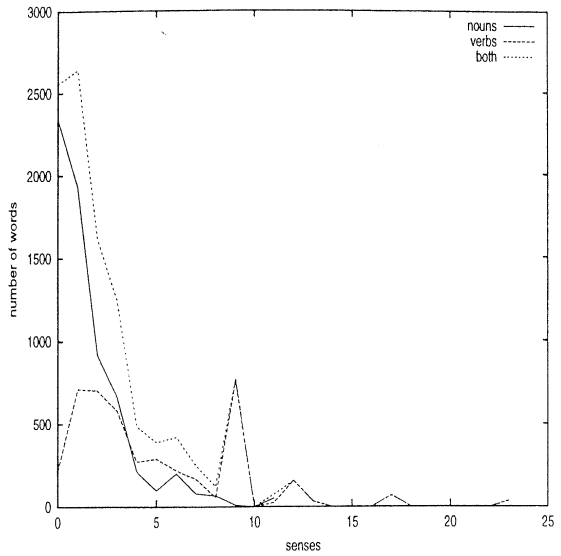

Figure 1: distribution of words in running text

reason for this is the frequent word 'olema' (*to be, to have*). Fortunately the distribution of its senses is highly skewed, meaning that mostly this word is used in one or two senses. Including the sense frequency information in the disambiguation process could considerably improve the results.

3 Problems and solutions

Several problems have been discovered concerning the relatively simple approach described above.

The output of the morphological analyzer often contains valuable information for word sense disambiguation which we have currently ignored.

- 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*, *tree trunk*, 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. The number of words behaving this way, though, is not very large.
- the modal verbs are explicitly marked in the output of the morphological disambiguator, when a verb is marked as such, then the senses that don't correspond to the modal senses could be removed and the winning sense should be chosen from the prevailing ones, e.g. the word 'saama' has all together 12 senses in the thesaurus, but only 2 of them correspond to the modal use of the word (either *can* or *may*).

Right now the frequency information of the senses has not been used. Most probably the results that could be obtained with the "commonest" baseline (Kilgariff and Rosenzweig, 2000) would beat the results of *semyhe*. We think that even the frequency information that could be calculated using the 10,000 manually sense-tagged words can be very useful for disambiguating purposes.

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, therefore the analyzer tries to disambiguate all the components of a multi-word unit separately, this of course results in an incorrect analysis. Also, having the information about the syntactic structure of the sentences could help to reduce the number of possible senses to choose from. For example the word 'olema' that was already mentioned above has five more frequent senses:

1. *be* — copula, used with an adjective or a predicate noun
2. *exist* — have an existence, be extant
3. stay in place, be stationary or spend a certain length of time
4. be somewhere, occupy a certain area, occupy a certain position
5. *have*, *have got*, *hold* — have or possess, either in a concrete or an abstract sense

The first sense is present in complementary clauses; senses 2, 3 and 4 appear in existential sentences and the last one in possessive sentences. If the information about the nature of the sentence was present in the input text it would certainly help the disambiguation process.

The output of *semyhe* stays often very ambiguous. This either happens when the sense-nodes of the context words are located in different trees so that their similarity cannot be calculated; or when different nodes of one word have the same parent node and are equally distant from the rest of the sense-nodes so that the similarity measure for them will be equal. The second reason may not be a big problem considering WordNet's fine-grainedness and the fact that for some applications a detailed sense distinction is not needed. The disambiguation result in this case can be simply seen as the union of the prevailed senses. Often, though, this approach does not hold, e.g. it is crucial for translation that the senses of the word 'naine' which can either stand for *woman*, *wife* or generally *female person*, are fully disambiguated, although the senses stand for more or less the same thing.

References

E. Agirre and G. Rigau. 1996. Word Sense Disambiguation using Conceptual Density. In *COLING-96*.

H.-J. Kaalep and T. Vaino. 1998. Kas vale meetodiga õiged tulemused? Statistikale toginev eesti keele morfoloogiline ühestamine. *Keel ja Kirjandus*, 1:30–36. In Estonian. English title: Getting right result with a wrong method? Statistical morphological disambiguation of Estonian.

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

A. Kilgariff and J. Rosenzweig. 2000. Framework and Results for English SENSEVAL. *Computers and the Humanities*, 34:15–48.

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.