Human Language Technologies – The Baltic Perspective
A. Tavast et al. (Eds.)

doi:10.3233/978-1-61499-133-5-82

A trivial method for choosing the right lemma

Heiki-Jaan KAALEP, Riin KIRT and Kadri Muischnek
University of Tartu

Abstract. This article presents a simple yet efficient method for solving lemma ambiguity as a part of morphological tagging of Estonian. By lemma ambiguity authors mean the situation when a word-form has several (mostly two) possible morphological readings and the only difference between these readings lies in the correct form of lemma, i.e. the POS and grammatical categories are the same, but possible lemmas are different. This type of ambiguity is characteristic of 1.5% of tokens in an otherwise morphologically disambiguated text. A text- and corpus-based method is used to disambiguate this kind of ambiguity. The precision of the method is 0.94 and recall 0.67.

Keywords: morphological disambiguation, lemma disambiguation, Estonian

Introduction

A typical chain of natural language processing consists of the following modules [1]: (1) tokenizer (splits text flow into tokens); (2) POS tagger (marks up each token with its particular part of speech tag) and lemmatizer (determines the basic form of each token), possibly followed by a syntactic analyzer; (3) word sense disambiguator (disambiguates the meaning of each token and assigns a sense to it).

Step (2) in turn consists of finding all the possible analyses of the words and subsequently choosing the most likely ones from these, based on the sentential context. Information used is essentially limited to grammatical tags (e.g. part of speech, plurality, case, mood etc.); lexical information comes into play solely for a few high-frequency words (e.g. copula and personal pronouns) that act as function words. The idea is that from a grammatical tagging point of view, it is the clausal position and inflectional form of a token that are important, not the exact lemma. Only by abstracting away from concrete lemmas, an algorithm will obtain a desired level of generality.

As a result of this general and grammar-oriented approach, however, some instances of lexical ambiguity are bound to remain. This is the issue the present article will concentrate on. The motivation behind our work is a desire to find a principled way for choosing the right lemma, when grammatical information on the sentence level is not sufficient to make a choice.

The article will concentrate on resolving lexical ambiguity in Estonian, i.e. instances when it is hard to tell the correct lemma form, in spite of having disambiguated the POS and grammatical categories (number, case, mood etc.). Knowing that *teod* is a plural nominative case form still leaves us with the possible lemmas *tigu* (snail) and *tegu* (deed).

The same phenomenon occurs frequently with proper names, e.g. a singular comitative case *Liisiga* (either with *Liis* or *Liisi*). Similarly, in Finnish, *säkeistä* is the plural elative case of either *säe* 'line, phrase, verse' or *säkki* 'bag, sack'; *patoissa* is the plural illative case of either *pato* 'dam' or *pata* 'spade; pot'. (Notice that solving this lexical ambiguity does not necessarily free us from the semantic ambiguity of the unique lemma.)

In contrast, in English, because of its impoverished morphology, such a situation would not arise: once we know that the word-form *banks* is the plural of *bank*, there is no form ambiguity left for the lemma, only semantic ambiguity (river side or financial institution).

The rest of the paper is structured as follows. In Section 1 we give a short overview of the lemma ambiguity types in Estonian. Section 2 presents the algorithm we use for lemma disambiguation. Section 3 gives overview of the results and Section 4 concludes.

1. Lemma ambiguity types in Estonian

For the background information, a few words about the inflectional morphology of Estonian are in place. Verbs can be inflected for mood, time and person; nominals can be inflected for case and number. There are 14 nominal cases in Estonian and for several inflectional classes two or three case forms can be homonymous. For example, inflectional forms of the word meaning 'mother' are identical for all three grammatical cases (nominative: *ema*, genitive: *ema* and partitive: *ema*). For a detailed description of Estonian the reader is referred to [3].

It is important to bear in mind that in Estonian, inflecting words belong to different inflectional classes, i.e. there are several different ways to create the same case form, the choice being dependent on the phonological shape of the lemma. On the other side, it is also possible that different lemmas give rise to homonymous wordforms. Basically, there are four types of lemma ambiguity in Estonian.

1) The lemma ambiguity, caused by homonymous case forms of different lemmas in the lexicon. In addition to the ambiguous nominal case forms exemplified by the ambiguous plural nominative form *teod* mentioned in the Introduction, the same kind of "real" lemma ambiguity can also be present in the verb paradigms. For example, the verbs *looma* 'create' and *lööma* 'hit' share all the past personal forms; e.g. the wordform *lõi* can be either 'hit' or 'create' in 3rd person past singular.

This group also contains compounds having a word-form with an ambiguous lemma as their final component. The compound formation in Estonian is free and productive, but, of course limited by semantic constraints. So, the word-form *roo* standing alone can be the singular genitive form of either the lemma *rood* 'nervure' or the lemma *roog* 'reed'. But while being a part of a singular genitive form of a compound noun *suhkruroo* 'sugarcane' only the second option is semantically plausible. But since the morphological analyzer and disambiguator don't take semantics into account, they treat the compound noun form *ruhkruroo* as ambiguous between two lemma readings.

2) It is obvious that a lexicon of the morphological analyzer cannot include all the words of a language. While analyzing these words, the analyzer must try to guess the lemma. A frequent group of such words are nouns used as terms in a specific professional language usage, e.g. the word-form *isolaadid* 'isolates' could theoretically

be the plural nominative form of either a lemma *isolaat* or *isolaad*, both absent from the lexicon.

3) Proper nouns form a numerous group of out-of-vocabulary words. In Estonian, proper names belong to fewer inflectional classes than common nouns, and this makes their treatment easier. However, the variability of names is greater than that of common nouns. A common problem is that a nominative case form of a noun can end with a consonant or with a vowel. If a name has a vowel ending, its nominative and genitive case forms are homonymous. So, if a proper noun occurs in a text in some other case form than consonant-ending singular nominative, one can't immediately tell whether its lemma should have a vowel ending or not.

E.g. the word-form *Endenil* can be analyzed as a proper noun in the singular adessive case, but its lemma can be either *Endeni* or *Enden*.

These three aforementioned types of lemma ambiguity could be called the "real" ambiguities that need to be solved.

4) There are several groups of nominals in Estonian that could be described as having parallel forms in singular nominative case, i.e. they have two different singular nominative forms, the rest of the paradigm being the same.

For example there is a group of words ending either with the suffix *-ke* or with the suffix *-kene* in singular nominative, e.g. *päike* or *päikene* 'sun'; *väike* or *väikene* 'small'; the singular genitive forms being *päikese* and *väikese* and the singular partitive forms *päikest* and *väikest* respectively. Both singular nominative forms are also used in present-day Estonian, the shorter form is somewhat more frequent, but the longer form is also present in texts.

The other group of words having this kind of lemma ambiguity contains nominals that again have two variants of the singular nominative case, one variant of the lemma being slightly archaic; e.g. the word meaning 'winter' has two possible lemmas *talv* and *tali*, the first of the two being stylistically neutral and the latter sounding a bit archaic.

This last type of ambiguity could be called pseudo-ambiguity as these word-forms are not genuinely ambiguous and there are no "right" and "wrong" lemmas depending on the context; even if the context and the meaning of the sentence are taken into account, both lemma readings still remain possible.

2. Algorithm

We can think about a text in the following way. A text contains words in their inflectional forms, and one can collect all these wordforms into (partial) paradigms. Imagine that we have an oracle which looks at every wordform and decides which paradigm it should belong to: what is its lemma and grammatical categories. The ultimate goal is to group all the wordforms into correct paradigmatic constellations. The oracle may use all kinds of different types of information to make its decisions: it may look at a lexicon and at the inflectional endings to make a decision about the possible paradigms, and it may look at the nearby tokens to narrow its choice. These are the traditional morphological analysis and disambiguation steps. In addition, the oracle may look at the potential partial paradigms it has at the moment. When the oracle sees a wordform which could belong to several constellations, it could use this existing evidence to make the choice.

The algorithm used in the described work to choose the correct lemma of a wordform is almost trivial (steps 1 and 2 are added for clarity, they are not part of the lemma disambiguating algorithm):

1. Find all the possible morphological analyses and lemmas of all the words, resorting to guessing in case the analysis cannot be found with the help of a dictionary. As a result, 40% of all the analyzed tokens have more than one reading.

2. Choose the most likely analyses, based on the sentential context. As a result, some tokens will have a unique grammatical reading, but more than one possible lemma.

3. Make a frequency list of all the lemmas in the disambiguated text (LL). Notice that if a token has multiple possible lemmas, they are all counted as separate ones.

4. For every token with multiple lemmas, keep only the most frequent lemma from LL, as frequency represents aggregate evidence from all the tokens in this text.

So if the text contains an ambiguous singular komitative form *Liisiga* (lemma is either *Liisi* or *Liis*) and an unambiguous singular partitive form *Liisit* (lemma is *Liisi*; the singular partitive form of *Liis* would be *Liisi,* because *Liis* belongs to a different inflectional type), we decide that *Liisi* is the correct lemma of *Liisiga* in this text.

The algorithm works because of the well-known data sparseness of natural language texts: in theory, both *Liisi* and *Liis* might be active in the same text, but in reality in overwhelming instances only one of them is. If there is not enough evidence for disambiguating in the same text, one may simply make a larger list of lemmas (LL) from a larger text collection. The task of disambiguation thus appears to be a task of choosing the suitable corpus: for example, choose texts from the same time period, or from the same subject field.

The algorithm assumes that all the possible lemmas had been found in a uniform and predictable way; otherwise, the simple frequency counts would be misleading. In order to guarantee this uniformity and predictability, we rely on a rule-based morphological analyzer and guesser [4].

This algorithm has been used as one of the steps in disambiguating the morphologically tagged National Corpus of Estonian, which can be queried at www.keeleveeb.ee.

3. Evaluation

For testing purposes we used a test corpus of 110000 tokens that contained newspaper, fiction and scientific texts. There were no considerable differences in terms of precision or recall between the different text classes, so we will present the results as an aggregate (see Table 1).

After morphological analysis and disambiguation, the text contained 1670 tokens (1.5%) with unique grammatical analyses, but ambiguous lemmas.

Our system disambiguated 1190 tokens of the 1670 ambiguous ones; 1120 were correct and 80 were incorrect. So the overall precision of the system was 0.94 and recall 0.67.

For a more detailed analysis of the performance of our system we looked at the following groups in more detail:

Table 1. Lemma disambiguation results

	Total	Ambiguous forms from the dictionary	Ambiguous forms from the guesser	Proper names from the guesser	Parallel forms in the nominative
Initially with lemma ambiguity	1670	620	280	640	130
Disambiguated by the system	1190	530	220	340	100
Correct	1120	500	190	330	100
Erroneous	80	30	40	10	0
Unchanged	480	90	60	300	30
Precision	0.94	0.94	0.86	0.97	1.0
Recall	0.67	0.81	0.68	0.52	0.77

1) The "real" lemma ambiguity, as described in Section 1 (group 1), is caused by the homonymous inflectional forms of nouns or verbs; e.g. the word-form *teod* can be the singular nominative form of either lemma *tigu* 'snail' or *tegu* 'deed'; the word-form *lõi* can be the 3rd person past form of either *looma* 'create' or *lööma* 'hit'.

620 instances of this group were present in the test corpus, of them 500 tokens received a correct and 30 an erroneous analysis. These numbers include also compound nouns with an ambiguous final component..

2) Lemmas not present in the lexicon but designated to a wordform by the guesser. As described in Section 1 (group 2), this group consists mostly of nouns used as terms in some specific professional language usage.

280 instances of this group were present in the test corpus, of them 190 tokens received a correct analysis and 40 tokens an erroneous analysis.

3) For proper nouns; as described in Section 1 (group 3), if not present in the lexicon of the morphological analyzer (i.e. the majority of the proper nouns) and if occurring in a text in other case forms than a consonant-ending singular nominative, one can't tell whether the lemma has a vowel ending or not, so the guesser outputs two lemmas.

640 instances of this group were present in the test corpus, of them 330 tokens received a correct analysis and 10 tokens erroneous analysis.

4) The simplex "pseudo-ambiguous" words that can be described as having parallel forms in the singular nominative case, e.g. the noun 'sun' has two variants of the singular nominative case - *päike* and *päikene*, both of them being in active use.

130 such tokens were present in the test corpus, all 100 of them that we disambiguated were correct. Actually, the disambiguator can't fail here as both lemmas are actually the correct ones.

As one can see from the Table 1, the overall precision of our system is quite high, given the extreme simplicity of the approach. The failures are mainly due to the following two factors:

1) The disambiguator that was used previous to the lemma disambiguation chose an incorrect reading with an incorrect lemma, and thus when later the frequency list of the lemmas was created, it tilted the statistics towards an incorrect alternative. E.g. *Endeni* may formally be a singular genitive case of either *Enden* or *Endeni*, but it might also be analyzed as a singular nominative case of *Endeni,* although local sentential context should cancel this last version out. If the morphological disambiguator has made an error and incorrectly tagged *Endeni* as a singular nominative, then the

frequency counts of lemmas *Enden* and *Endeni* will be tilted towards *Endeni* and the wrong lemma will be chosen for all the ambiguous instances.

2) If the supposedly uniform text that has been chosen as the unit for lemma disambiguation in reality is a mixture of texts, then the principle "one sense per discourse" may be violated: the algorithm will just choose the more frequent lemma, and this will be incorrect in the minority number of cases. (If the text were ideally chosen, there would be no minority cases at all.)

The recall of the algorithm depends on the availability of disambiguating information. If the textual unit that is used for creating the frequency list of the lemmas is too small, the algorithm has not enough information for choosing the more frequent lemma. In our experiment with the test corpus, we did not try to look beyond the borders of the text at hand. This may explain the somewhat low recall.

There were no significant differences between the text classes (fiction, newspapers, science) in the disambiguation quality. Nevertheless, one could point out that disambiguating the proper nouns in fiction texts is a little easier task than disambiguating them in newspaper texts, which is probably due to the fact that the names in fiction texts are used more repetitively. The science texts had the greatest amount of guesser-generated ambiguous analyses for nouns used as terms.

4. Related work

We are not aware of other similar lexical disambiguation endeavors. Krister Lindén and Jussi Tuovila [5] describe a related problem, namely guessing both the lemma and the inflectional paradigm of unknown Finnish words, so that they could be added to a morphological lexicon. For guessing them, they generate a number of possible lemmas for these unknown forms, and, based on these, a number of key word forms, which they later compare with actual words from a corpus. Our algorithm differs from theirs in that we do not go for creating an explicit lexicon. Instead, we pretend that the guessed lemmas, coupled with the actual word forms, form the potential (partial) paradigms of some hypothetical words. A wordform with multiple possible lemmas can belong to different constellations. We just choose the more likely (partial) paradigm for the wordform, but we do not try to add these words to a lexicon. (We could add them to a lexicon afterwards, from our corpus, once the lemmas have been disambiguated.) It seems that the algorithm by Linden and Tuovila could be simplified, if the guessing of lemmas would be done as a separate stage, and the resulting set of lemmas pruned similarly to our corpus-based disambiguation.

Hrafn Lofson et. al. [6] use frequency information for analyzing unknown words in the framework of rule-based morphological disambiguation of Icelandic. Namely, if the wordform is still morphologically ambiguous after local rules for initial sentence-level disambiguation and a set of heuristics (global disambiguation) have been applied, then simply the most frequent tag of the word-form is chosen. Also, if an unknown word can't be fully disambiguated, the most frequent tag is chosen, if the frequency information is available. So Lofson et. al. use the frequencies derived from training data, while we are relying on the frequency information gathered from the test text during the test run, or, if the text does not contain enough evidence, the frequency information gathered from a possibly previously unseen subcorpus.

5. Conclusion

Ambiguity, an inherent and pervasive trait of natural language, has multiple manifestations. It can be found in morphology, syntax and semantics; it may be a real linguistic feature in some cases, while in other cases it is just the result of our over-specification of a linguistic phenomenon, or our imperfect way of processing the language. To make things yet more complicated, the exact points in the language processing chain where ambiguity may rise are highly language-specific.

A special case of ambiguity is lexical ambiguity, i.e. instances when it is hard to tell the correct lemma form, in spite of having disambiguated the POS and grammatical categories.

In terms of the language processing chain, this lexical ambiguity becomes visible already in the step of morphological analysis and disambiguation. This is different from the typical situation with word sense disambiguation (WSD), when the ambiguity becomes visible after POS disambiguation, after having looked up the lemmas in a dictionary (or a parallel, translated text), and having seen that they have multiple senses there (or translations).

The other differences from WSD are the following. First, in some seemingly ambiguous cases the lemmas might actually be different in a trivial way (similar to *centre* and *center*). Second, in case of out-of-dictionary words, notably proper names, the differences in lemmas might be a result of the guesser's uncertainty; so they form an open list.

However, whatever the source of the lexical ambiguity, traditional POS-disambiguation methods are of no help here – they look at grammatical readings only (which are the same in this case), and they do not look at a wider context than a sentence.

In the present article, we propose to follow the principle "one sense per discourse" [2]: look at all the instances of the ambiguous lemmas in the text and choose the most frequent reading.

This simple method achieves for lemma disambiguation an overall precision of 0.94 and recall of 0.67.

Acknowledgements

This work is supported by the European Regional Development Fund through the Estonian Centre of Excellence in Computer Science (EXCS), the Estonian Ministry of Education and Research (grants SF0180078s08 and EKT11005) and project META-NORD (ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, Grant agreement no 270899).

References

[1] D. Cristea, I. Pistol. Managing Language Resources and Tools using a Hierarchy of Annotation Schemas. *Proceedings of Workshop 'Sustainability of Language Resources and Tools for Natural Language Processing'*, organized in conjunction with LREC 2008 (2008).

[2] W. Gale, K. Church and D. Yarowsky. One Sense Per Discours., *Proceedings of the 4th DARPA Speech and Natural Language Workshop* (1992).

[3] M. Erelt (editor), *Estonian Language*. Lingusitica Uralica Supplemenatry Series vol 1.Estonian Academy Publishers. (2003).
[4] H-J. Kaalep, T. Vaino. Complete Morphological Analysis in the Linguist's Toolbox. *Congressus Nonus Internationalis Fenno-Ugristarum Pars V*, (2001), 9-16.
[5] K. Linden, J. Tuovila. Corpus-based lexeme ranking for morphological guessers. *State of the Art in Computational Morphology*, edited by C. Mahlow, M. Piotrowski. Springer, Berlin, Heidelberg (2009), 118-135.
[6] H. Loftsson, S. Helgadóttir, E. Rögnvaldsson. Using morphological database to increase the accuracy in PoS tagging. *Proceedings of Recent Advances in Natural Language Processing (RANLP 2011)*. Hissar, Bulgaria, 49-55.