Inducing Baseform Models from a Swedish Vocabulary Pool

Eva Forsbom
Department of Linguistics and Philology
Uppsala University/Graduate School of Language Technology
Box 635, SE-751 26 UPPSALA
evafo@stp.lingfil.uu.se

Abstract

In many language technology applications, we need to map wordforms to a citation form or baseform, or the other way around, e.g. for lexicon lookup or for representational purposes.

In this paper, we used a suffix trie mapper with suffix-change probabilities, and computed wordform-baseform and baseform-wordform models from eight subsets of a ranked Swedish vocabulary. All models were evaluated for both directions on a testset, and four of the models were also evaluated for wordform-baseform mapping on five unseen texts.

For wordform-baseform mapping, the best models performed on par with state-of-the-art systems. Most models were useful for some situation—given mapping direction, and time and space restrictions—but no model was best for all situations.

1 Introduction

In many language technology applications, such as machine translation or cross-language information retrieval, words are looked up in a lexicon where only one form of the word—the citation form (usually the baseform)—is present. Even for other applications, such as monolingual information retrieval or lexical cohesion analysis, it can be useful to conflate all forms of a word into one "concept" form. Thus, there is a need for a wordform-baseform mapper.

Going in the other direction—from baseform to wordform—could also be useful for applications such as natural language generation, or query expansion in information retrieval. A mapper that can handle both directions reasonably well, perhaps with different underlying language models, would be an extra treat: two applications for the price of one.

But what kind of information, and how much, should such models contain? In this paper, we test two assumptions: 1) that irregular wordforms either are among the most top-frequent words in a vocabulary, or so rarely used that they are insignificant for a robust application, and 2) that rules for regular forms can be induced from a limited number of examples.

We describe the induction of various wordform-baseform mapping models (Section 3.3) from subsets of a Swedish vocabulary pool (Section 3.1), in the framework of Wicentowski's Base Model (Section 3.2). The models are evaluated both on a testset (Section 4.1) and on unseen texts (Section 4.2).

2 Background

For languages with little inflectional morphology, such as English, stemming can be adequate for finding most baseforms, but for morphologically richer languages, such as Swedish, more morphologic analysis is usually needed.

Several academic systems for morphologic analysis of Swedish words exist (for an overview, see e.g. Dura (1998)), but they are not always suited for simple wordform-baseform mapping and generally not publicly available. At least two commer-

Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (Eds.)
NODALIDA 2007 Conference Proceedings, pp. 51–58

cial systems with publicly available demos exist, SWETWOL (Karlsson, 1992) and Lexware (Dura, 1998), although the demos come with limited access. Both systems are rule-based. SWETWOL is based on two-level morphology, and outputs all possible analyses of a word. It is possible to retrieve a baseform from the analysis, but a disambiguator is needed to choose among the alternative analyses. Lexware is based on inflectional paradigm rules and word-formation rules, and outputs a single analysis (in the demo version). It is generally possible to retrieve the baseform from the analysis.

In Wicentowski's statistically based approach (Wicentowski, 2002), four mapping model types are used, where wordforms are stored in a suffix trie, with varying amount of morphological information in the nodes. In the simplest type, Base Model, the node annotations contain probability estimates for suffix transformations from wordform to baseform, optionally conditioned on a part-of-speech (PoS) tag. Suffix transformations are learnt from a list of ⟨wordform, baseform⟩ tuples (or ⟨wordform, PoS tag, baseform⟩ triples), which could be taken from a dictionary, collected from a corpus, or compiled manually. The probability estimates are also computed from that list.

As the Base Model only considers suffix changes, it is not appropriate for all languages. For suffigating languages like Swedish, however, it has proved to work well: 94.97% type accuracy for a model trained on Stockholm-Umeå Corpus (SUC 1.0) (Ejerhed et al., 1997), evaluated on a testset without part-of-speech tags (13,871 verb forms, 53,115 noun forms, and 53,115 adjective forms) (Wicentowski, 2002).

As a side effect, the mapper can also be used for wordform generation for a baseform given a part-of-speech tag, if the model is reversed and the tag contains enough information. Wicentowski considers this an easier task than wordform-baseform mapping, but did only a minor evaluation on verb forms for English, French and German, and using separate models for each part-of-speech tag. Accuracy ranges from 88.70 to 99.78%.

3 Experimental setup

As we cannot fit all the words in a language into a model, we have to choose the ones that are most useful. Our assumption is that we do not need all regular wordforms in the model, as the ones missing can be handled by analogy. On the other hand, we have to include all irregular wordforms, at least the most frequently used wordforms, or else the application would not be robust enough.

To test our assumptions of the kind and amount of data needed for a good mapper, we used various subsets from a frequency-ranked vocabulary, and our Perl implementation of Wicentowski's Base Model.

3.1 Vocabulary pool

The data used for induction in the experiments come from a Swedish lemma vocabulary pool (Forsbom, 2006) derived from version 2.0 of the 1-million word balanced corpora Stockholm-Umeå Corpus (SUC) (Ejerhed et al., 2006). SUC is compiled in a manner similar in spirit to that of the Brown (Francis and Kučera, 1979) corpus, and is meant to be representative of what a person might have read in a year in the early nineties. Each word is annotated with its baseform and its part-of-speech (mapped to the PAROLE tagset). The texts are also categorised in 9 major categories (genres) and 48 subcategories (domains).

The units of the vocabulary pool are "lemmas", or rather the baseforms from the SUC annotation disambiguated for part-of-speech, so that the preposition *om* 'about' becomes `om.S` and the subjunction *om* 'if' becomes `om.CS`. The lemmas are ranked according to relative frequency weighted with dispersion, i.e. how evenly spread-out they are across the subdivisions of the corpus, so that more evenly-spread words with the same frequency are ranked higher.

The total lemma vocabulary has 69,560 entries, but there is also a genre and domain independent base vocabulary, restricted to entries which occur in more than 3 genres, which has 8,554 entries.

For our experiments, we used only the original SUC baseform in ⟨wordform, PAROLE tag, baseform⟩ (or the reverse) triples as input for inducing the mapping models from various subsets of the vocabulary.

3.2 Base Model mapper

In our implementation of Wicentowski's Base Model, the suffix transformations are conditioned on a part-of-speech tag, to limit the number of possible transformations.

When using the mapper on seen words in our version, the mapper returns only the transformation(s) applicable to those words, i.e. a single transformation for non-ambiguous words, and a set of transformations ranked by their estimated probability for ambiguous words. For unseen words, the mapper follows the design of the original Base Model, and the mapper returns applicable transformations ranked by a weighted back-off probability based on the longest common suffix. The weight is static, and set to 0.1. Our mapper outputs the top-ranked baseforms, optionally with a confidence score.

3.3 Models

In the experiments, two sets of models were trained (see Table 1): one with all entries from the full vocabulary (105,815 entries), and one with only entries from the base vocabulary (28,050 entries). The set based on the full vocabulary is the same as in Wicentowski's experiments, apart from corpus version and the inclusion of PoS tags. Our hypothesis is that it contains all the frequent irregular forms, more than enough samples of regular forms, and some infrequent irregular forms that are not so useful. The set based on the base vocabulary, on the other hand, should contain all the frequent, but no infrequent, irregular forms, and enough samples of regular forms.

The full set obviously takes up more space, takes longer to load and takes longer to search in.

The PAROLE part-of-speech set is rather detailed (153 tags), and an automatic part-of-speech tagger is likely to make a few errors on the more detailed morphological information, while the actual part-of-speech most often is correct (cf. Megyesi (2002)). To see how the Base Model performed in circumstances with a less detailed tagset, we also conflated the PAROLE set to a smaller set (29 tags), i.e. the same set used for disambiguating lemmas in the vocabulary pool. For most words it is simply the part-of-speech, but for common nouns, for example, there is a distinction between neuter and non-neuter gender, as the same baseform could have two differing paradigms depending on gender.

Loading the models with shorter tags used roughly the same amount of time and space, but the lookup time for the evaluation testset (163,999 entries) was about three times longer than for the models with more detailed tags, as there were more alternatives for each node in the trie. For wordform lookup, the shorter tags were not expected to be very useful, as they give no clue about what wordform should be generated.

In addition, we used wordform filtering for four models, since many baseforms had several alternative wordforms connected to a part-of-speech. Wordform filtering was mainly intended for baseform-wordform mapping. Most of the alternatives were antiquated forms (e.g. *hwarandra* for *varandra* 'each other' and *hafva* for *ha(va)* 'have') or forms from reported speech (e.g. *e', e, ä', ä* for *är* 'is'), and some baseforms were not real lemmas, i.e. having the same inflectional paradigm (e.g. *vara* as auxiliary, 'be', or as main verb, 'be' or 'last').

In two of the wordform-filtered models, only wordforms occurring more than once were included, to get rid of wordforms for which statistic information was unreliable. This frequency-based filtering was very aggressive, in particular for the full vocabulary model; reducing its size by more than 50%.

Another filter was used in the two other models, to filter out alternative wordforms for a part-of-speech and baseform, i.e. if their frequency ratio (among all alternatives for that case) were less than or equal to 0.1, to remove the least plausible wordforms. This ratio-based filtering was rather modest for both vocabulary models.

4 Evaluation

We wanted to evaluate the theoretical bounds of the models on a testset which include many words not present in the models, and with many semi-regular and irregular forms, to see their limitations (see Section 4.1). But, as the models are to be used in real applications, mainly for baseform lookup, we also wanted to evaluate them on real, unseen, texts (see Section 4.2).

Model	Set	Filter	Tagset	Size	Loading		Lookup
					Memory	Time	Time
FullFull (bf)	Full	v=full,f=0,r=0	Full	105,815	156MB	24.79s	5m16s
FullFull (wf)	Full	v=full,f=0,r=0	Full	105,815	119MB	22.25s	5m18s
FullShort	Full	v=full,f=0,r=0	Short	105,815	155MB	25.10s	19m57s
FullFiltered1	Full	v=full,f=1,r=0	Full	44,918	66MB	9.77s	3m18s
FullFiltered01	Full	v=full,f=0,r=0.1	Full	105,289	155MB	24.60s	4m47s
BaseFull (bf)	Base	v=base,f=0,r=0	Full	28,050	40MB	5.64s	4m00s
BaseFull (wf)	Base	v=base,f=0,r=0	Full	28,050	23MB	4.34s	3m33s
BaseShort	Base	v=base,f=0,r=0	Short	28,050	39MB	5.30s	13m37s
BaseFiltered1	Base	v=base,f=1,r=0	Full	21,645	32MB	4.21s	2m59s
BaseFiltered01	Base	v=base,f=0,r=0.1	Full	27,540	39MB	5.26s	3m16s

Table 1: SUC baseform/wordform models. Filters: v=vocabulary, f=frequency, r=ratio. Time and memory usage was measured with `top` and `time` on a computer with 4 processors (Intel(R) Xeon(TM) CPU 2.80GHz), i686 Linux kernel 2.6.16-1.2115_FC4smp, 2070kB RAM (k=2^{10}, M=2^{20}).

4.1 In theory: Testset

In the absence of a standardised testset for Swedish morphology, we used the freely available DSSO (Westerberg, 2003)[1], which the Swedish spelling dictionary for `OpenOffice` is based upon.

DSSO contains some morphosyntactic information, such as part-of-speech, case, number, and tense, but misses information on, for example, gender for nouns. In some cases, it has a different view of what part-of-speech a word belongs to (e.g. no determiners, just pronouns, or no subjunctions, just conjunctions), or what the baseform of a word is (e.g. participles have the infinitive verb form as baseform, while in SUC they are mapped to the non-neuter, indefinite, participle form—an adjective form).

In order to make DSSO useful for evaluation of our models, we automatically transformed the DSSO morphosyntactic information into PAROLE tags. In the case of systematic differences, we used a set of rules to do the mapping, and in case of missing information, we used the statistical part-of-speech tagger TnT (Brants, 2000) with a model trained on SUC (Megyesi, 2002) to output all possible tags for each word and then heuristics to choose the right information (e.g. for noun gender, the gender of the most probable noun tag, and non-neuter as default). Obvious errors were corrected, and some erroneous entries in the original DSSO were filtered out, but a few errors may remain.

The transformed testset contains 163,999 entries of ⟨wordform, PAROLE tag, baseform⟩ triples (19.80% in common with the FullFull model. More than half of the entries are common nouns (91,436).

In Table 2, the error rates for the various models on DSSO with only the top 1 alternative are given. The results cover both baseform lookup and, the reverse, wordform lookup. The lower bounds[2] for baseform lookups are given by two baselines: no change of form, and stemming by the freely available Snowball stemmer for Swedish (Porter, 2001). Upper bounds (or state-of-the-art performance) could not be computed for this testset, as we did not have access to state-of-the-art systems other than as demos with limited access. Performance has been reported for, for example, SWETWOL as 0.7 and 0.4% error rate, respectively, for baseform lookup on two texts (Karlsson, 1992): 1) 47,422 tokens (8,432 types) and 2) 54,542 (5,857).[3] The error rates were based on tokens rather than types, which makes comparison hard. Our models are way better than the lower bounds, but also a bit away from the upper bound, although the comparison is skewed, since our error rates are based on types and on the top 1 ranked alternative only.

Among our models, the FullFull model was the best, both for baseform and wordform lookup. And the frequency-filtered models did worse than the unfiltered models, even on wordform lookup, which

[1] `http://dsso.se`

[2] Here, lower is used in the sense worst performance, although the numbers for the error rates are higher.

[3] In a list message summary on PoS-taggers 1993, Lingsoft reports on the performance ("recognised 99.3%", or an error rate of 0.7%) for a list of 300,000 wordforms (`http://www.sfs.uni-tuebingen.de/~abney/taggers.html`) The performance probably refers to recall rather than precision, and tokens rather than types.

Model	Baseform	Wordform
FullFull	4.35	5.52
FullShort	9.66	73.69
FullFiltered1	5.24	7.54
FullFiltered01	4.36	5.51
BaseFull	5.14	6.61
BaseShort	8.49	73.03
BaseFiltered1	5.64	8.42
BaseFiltered01	5.14	6.62
Snowball	57.38	-
NoChange	76.56	-

Table 2: Overall error rates for SUC models and baselines on DSSO (top 1 ranked).

they were supposed to boost performance for, while the ratio-filtered models did about the same as the unfiltered models. For baseform lookup, the models with conflated part-of-speech tags did worse than the models with the full tagset, but much better than the lower bounds. For wordform lookup, on the other hand, they were as lousy as expected.

In Figure 1, we show the error contribution by part-of-speech tag for baseform lookup with the FullFull model. The majority of errors come from genitive forms of common nouns (NC**G@**), where no applicable mapping was found at all and a non-changed form was used (e.g. *bärsens* 'the beers" for *bärs*), or the wordform was missing from the model and the majority regular mapping was used (e.g. *bevi* '(a) proof's" for *bevis*). As it turns out, there are very few genitive forms at all present in the model (1,052 plural forms and 2,619 singular forms), compared to the number of common noun baseforms (21,067). As baseforms ending in *s* or *x, z* are ambiguous for case, writers also usually avoid the synthetic form and use reformulation strategies instead, to be clearer. So, in real-life situations, the genitive errors might not be so important. They might also be possible to remedy in some respect by editing the model to include genitive forms for all baseforms in the model.

Other common errors originate from plural forms of common nouns (NC*P*@**) not found in the model, where the baseform should end with a vowel or have a vowel inserted before *l, n, r*, but the vowel is clipped (e.g. *backarna* 'the hills' to *back* instead of *backe*, or *fablerna* 'the fables' to *fabl* instead of *fabel*). Neuter adjectives (AQPNSNIS) where the baseform should end with *t*, but the *t* is clipped (e.g. *transparen* 'transparent' instead of *transparent*) is another common error source. Deponential verbs (V@**SS) not in the model also have their *s*-suffix chopped off (e.g. *ända* 'to end' instead of *ändas*), an error which can be attributed to the tagset rather than the application, since the tagset does not distinguish between deponential verbs and passive verb forms.

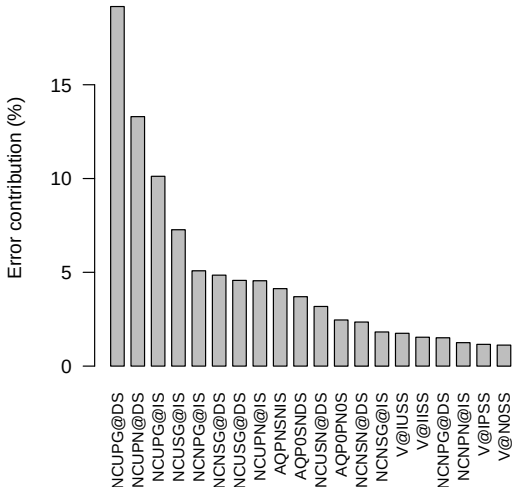

Figure 1: Error contribution by PAROLE tag for FullFull on DSSO baseform lookup (top 1 ranked, error contribution$\geq$1).

As our models are based on statistics, wordforms (or baseforms) that are not in the models are usually assigned the analysis of the most regular mapping, although the correct mapping is also given, but as a lower ranked alternative. When looking at the top 2-10 ranked alternatives, the error rate for all models goes down drastically for the first 2-4 alternatives, and then levels out (see Figures 2 and 3). This is also a more fair comparison with SWETWOL (although the token-type discrepancy is still present).

For baseform lookup, the FullFull, FullFullFiltered01, FullShort and BaseShort models are at the same error rate level as SWETWOL from top 4 onwards, and BaseFull and BaseFullFiltered01 is close (0.8% error rate). The two models with the conflated tagset actually outperform the models with the full tagset from top 4 or 5. This indicates that the full tagset is better for disambiguating regular alternatives with the same suffix within the same part-of-

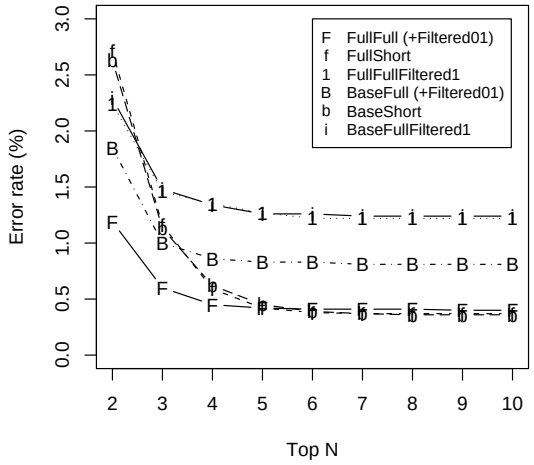

Figure 2: Error rates for SUC models on DSSO baseform lookup (top 2-10 ranked).

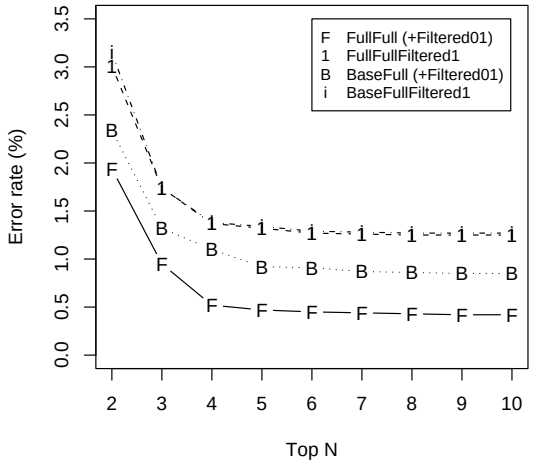

Figure 3: Error rates for SUC models on DSSO wordform lookup (top 2-10 ranked).

speech, while the conflated tagset is better at disambiguating alternatives where the more detailed morphosyntactic information in the full tagset gives too few data points. In those cases, the FullShort and BaseShort models rely more on the suffixes than the tags.

For wordform lookup, all models with the full tagset have error rates between 0.5 and 1.3% from top 5 onward. As mentioned before, the wordform filterings did not help much. On the other hand, the two models with the conflated tagset (not intended for wordform lookup and not shown in the figure) reach error rates of 13.6 (BaseFull) and 33.1% (FullFull) for the top 10 ranked alternatives.

4.2 In practice: Real texts

In the more real-life evaluation, we only used the unfiltered models: FullFull, FullShort, BaseFull, and BaseShort, and only used them for baseform lookup. The selected models were applied to five news texts, randomly sampled from the Scarrie corpus (Dahlqvist, 1999).

As input, we used corrected output from the tagger used for the testset. We corrected the tags as the evaluation should evaluate the models, not the tagger. We also used a single analysis as output, as in the following example for the wordforms *spårlöst* *'without a trace'* and *industrikoncernen 'the industrial concern'*:

```
spårlöst          RG0S     spårlöst
industrikoncernen NCUSN@DS industrikoncern
```

For the comparison, the texts were also analysed with 2 other morphological analysers: SWETWOL and Lexware. The two analysers are not designed for baseform lookup as such, but can be used for the task if the output is post-processed.

SWETWOL is a commercial morphological transducer lexicon description of Swedish[4]. It is based on classical Swedish grammar and can analyse words by inflection, derivation and compounding. The SWETWOL analyser consists of a lexicon with more than 45,000 entries, mostly derived from SAOL[5] and a set of two-level rules compiled into run-time finite-state automata. The output is all possible analyses of the wordform, as in the following examples:

```
"<spårlöst>"
  "spår|lös"  A NEU INDEF SG NOM
  "spår#lös"  A NEU INDEF SG NOM
  "spår#lösa" V ACT SUPINE
  "spår#lösa" V PCP2 UTR/NEU INDEF SG NOM
  "spårlöst"  ADV
"<industrikoncernen>"
```

[4] `http://www.lingsoft.fi/cgi-pub/swetwol`

[5] Svenska Akademiens Ordlista 'Swedish Academy Wordlist'.

```
"industri#koncern"  N UTR DEF SG NOM
```

In the comparison, we used the analysis (or analyses) corresponding to the correct part-of-speech tag given the context. If there were more than one analysis with correct part-of-speech, the correct baseform for the word in context was used.

Lexware is a commercial "language engine" (Dura, 1998) with, *inter alia*, morphological analysis of Swedish wordforms. Its knowledge base is derived from the information in NEO[6], and it has 80,000 entries, inflectional patterns, and 400 word formation rules. For the comparison, we used the Nyckelord tagging demo[7], where the output is the baseform (including word ID and any segmentation), part-of-speech, and inflectional pattern ID, as in the following examples:

```
"spårlöst"
  spårlös(44004) ADVERB inflections=1
"industrikoncernen"
  industri(22508)_koncern(25765) NOUN
                              inflections=3
```

The texts used for evaluation were tokenised at major punctuation characters and white space (752 tokens in total), but not normalised (i.e. case-folded) before mapping. There were 398 normalised types, but 403 different analyses, in at least one analyser: 4 types differed in analysis due to capitalisation, and 3 types due to homonymity. Accuracy was therefore counted on the basis of the 403 "types".

As we were comparing morphological analysers implemented for different application purposes, and using slightly differing levels of analysis, we had to be a bit lenient in our evaluation. The pronoun *det* ('it', neuter), for example, could be treated as either an inflection of *den* (uter) or as a baseform. Other examples are adverbs (*spårlöst*) derived from adjectives (*spårlös*), which could be treated either as neuter inflectional forms of the adjective, derivations, or as baseforms. When there was no possible difference in analysis level, we therefore counted an analysis as correct if the analysis of an inflected word was correct, and as incorrect if an analysis of a non-inflected word was incorrect. Where there was a possible difference in analysis level, we counted an analysis as acceptable if the analysis of an inflected word was one of the possible variants, and as unacceptable if the analysis of a possibly non-inflected word was none of the possible variants. As can be seen in Table 3, our models did a bit better (99.3%) than SWETWOL (98.3%), which in turn did better than Lexware (97.5%), but the differences are small.

However, Lexware was somewhat penalised since it uses its own tagger and some of the errors were tagging errors. Some words were also segmented correctly, but as Lexware uses baseforms also for the parts, it was not possible to derive the baseform from the parts, it was counted as an error, e.g. *20-|öre|valör* for *20-öresvalörerna* instead of the correct *20-öresvalör* 'value of 20 öre', i.e. the glueing *s* is missing. For some reason, probably because it was not in the lexicon, Lexware also messed up *O-listenoterade* 'listed on the O list (stock exchange)' while it did a perfect job on the almost identical *A-listenoterade*.

Case errors were not counted as errors here as normalisation was not the object of evaluation. SWETWOL always does lower-case conversion, but keeps record of initial capitals, so it is possible to restore it if necessary. Our Swedish models are case-normalised, so the input should really be normalised beforehand, so that any normalised forms that are in the models can be recognised. For example, the normalised wordform of *Aktier* 'stocks' is in the models, and correctly analysed if normalised beforehand, but incorrectly analysed if not normalised.

The BaseFull model also made a mistake on *rånarna* 'the robbers', which it analysed as *rån* 'robbery', and not the correct *rånare*. The word was not in the model, and *arna* is a common inflectional ending for definite plural nouns.

5 Concluding remarks

In this paper, we investigated the performance of a wordform-baseform mapper (or reverse), using various subsets of a Swedish frequency-ranked vocabulary pool as input models. We wanted to find out what kind of information, and how much, a good model should contain. The hypothesis was that a smaller model would fit two assumptions: 1) that irregular wordforms either are among the most top-frequent words in a vocabulary, or so rarely used that

[6] Nationalencyklopedins ordbok 'Swedish National Encyclopedia Wordlist'.

[7] http://www.nla.se/lexware/

	Analysis	Max	Full Full	Full Short	Base Full	Base Short	SWETWOL	Lexware
Non-inflected	Incorrect	198	0	1	0	1	2	2
	Unacceptable	44	0	0	0	0	0	1
Subtotal	Accurate	242	242	241	242	241	240	239
Inflected	Acceptable	13	13	13	13	13	13	13
	Correct	148	146	144	146	144	144	141
Subtotal	Accurate	161	159	157	159	157	157	154
Total	Accuracy (%)	100	99.3	98.8	99.3	98.8	98.3	97.5

Table 3: Results for baseform lookup for the 4 models, SWETWOL, and Lexware.

they are insignificant for a robust application, and 2) that rules for regular forms can be induced from a limited number of examples.

Eight subsets from the vocabulary pool were used as models, and were evaluated for both directions on a testset of mappings. Four of the models were also used for wordform-baseform mapping on 5 randomly selected texts from the Scarrie corpus.

For the corpus text evaluation, the smaller models performed as good as the larger ones, which indicates that our hypothesis is plausible.

Most models were useful for some situation, but no model was best for all situations, so when using the mapping application for either baseform or wordform mapping, one can choose a suitable setting for how many top-ranked alternatives should be returned and a suitable model, depending on the intended application of the baseform or wordform mapping, and requirements on accuracy, speed, and space limitations.

For wordform-baseform mapping, the best models also performed on par with state-of-the-art systems.

A demonstrator and the BaseModel package, with programs, models, and testset, are available from `http://stp.lingfil.uu.se/~evafo/resources/baseformmodels/`.

Acknowledgements

This research was funded by the Swedish and Nordic Graduate School of Language Technology. We also wish to thank the reviewers for their comments.

References

Thorsten Brants. 2000. TnT - a statistical part-of-speech tagger. In *Proc. of the Sixth Applied Natural Language Processing Conference*, Seattle, Washington.

Bengt Dahlqvist. 1999. A Swedish text corpus for generating dictionaries. In Anna Sågvall Hein, editor, *The SCARRIE Swedish Newspaper Corpus*, Working Papers in Computational Linguistics & Language Engineering 6. Dep. of Linguistics, Uppsala University.

Elżbieta Dura. 1998. *Parsing Words*. Data Linguistica 19. Göteborg University, Göteborg. PhD thesis.

Eva Ejerhed, Gunnel Källgren, and Benny Brodda. 1997. Stockholm-Umeå corpus version 1.0, SUC 1.0. Dep. of Linguistics, Stockholm University and Dep. of Linguistics, Umeå University.

Eva Ejerhed, Gunnel Källgren, and Benny Brodda. 2006. Stockholm-Umeå corpus version 2.0, SUC 2.0. Dep. of Linguistics, Stockholm University and Dep. of Linguistics, Umeå University.

Eva Forsbom. 2006. A Swedish base vocabulary pool. Presented at the Swedish Language Technology Conference, Göteborg.

W. Nelson Francis and Henry Kučera, 1979. *Manual of information to accompany a Standard Sample of Present-day Edited American English, for use with digital computers*. Providence, R.I. Original ed. 1964, revised 1971, revised and augmented 1979.

Fred Karlsson. 1992. SWETWOL: A comprehensive morphological analyzer for Swedish. *Nordic Journal of Linguistics*, 15(1):1–45.

Beáta Megyesi. 2002. *Data-Driven Syntactic Analysis. Methods and Applications for Swedish*. TRITA-TMH 2002:7. Inst. for Speech, Music and Hearing, Royal Institute of Technology, Stockholm. PhD thesis.

Martin F. Porter. 2001. Snowball: A language for stemming algorithms. `http://snowball.tartarus.org`.

Tom Westerberg. 2003. Den stora svenska ordlistan [The large Swedish dictionary]. Version 1.13.

Richard Wicentowski. 2002. *Modeling and Learning Multilingual Inflectional Morphology in a Minimally Supervised Framework*. Ph.D. thesis, John Hopkins University, Baltimore, Maryland.