Big is beautiful
Bootstrapping a PoS tagger for Swedish

Eva Forsbom
Uppsala University/GSLT
evafo@stp.lingfil.uu.se

GSLT retreat
Gullmarsstrand January 28, 2006

Abstract

A statistical part-of-speech tagger trained on a one-million word Swedish corpus with validated tags was used to tag two considerably larger untagged corpora (≈ 78 and 20 million words, respectively) to bootstrap new, improved, tagger models. The new taggers all showed better accuracy both for seen and unseen words, and the best tagger had 97.02% overall accuracy evaluated on the original corpus (using 10-fold cross-validation).

What's the problem?

In applications which rely on a part-of-speech (PoS) tagger for pre-processing, any tagging error will lower the accuracy of subsequent modules. For example, a module reducing wordforms to their baseform given a PoS tag, would mess up a proper noun ending with a frequent nominal inflectional suffix if the word was erroneously tagged as a common noun, such as

```
Franzen     NCUSN@DS     Franz+en
```

When training data is limited, ensembles of taggers, or bootstrapping techniques, could be used to find or remove noise from the data, and thereby get better precision.

If a tagger does not make systematic errors, but is only making mistakes due to lack of training data, it might also be possible to boost performance by using the tagger to tag considerably much more data, even if not validated, and train a new tagger model on that data.

A little goes a long way

In our experiment, we used the statistical TnT tagger [1], which can be tuned with various parameters for training and tagging. The following parameters can be used for training:

default different suffix tries for capitalised and non-capitalised words.

case-labelled different suffix tries + extra capitalisation marker on tags.

case-folded same suffix tries for capitalised and non-capitalised words.

Tagging parameters are the following (last is default):

n-gram 1-, 2-, 3-gram.

smoothing replace 0 by constant, add constant to all frequences, linear deleted interpolation.

unknowns none allowed, use dummy, combine statistics from all words, combine statistics on singletons.

suffix length 1-10.

As a baseline, we used a model trained on the one million word balanced Stockholm-Umeå corpus (SUC) [7], which has manually validated tags. It was trained on 90% and evaluated on 10% of the corpus using 10-fold cross-validation, where each corpus text was randomly assigned to one of the folds. Default settings were used for both training and tagging. The accuracy (see Table 1) is better than that obtained in two other studies on the same data, probably due to the sampling method: sentences from the texts [3] and text chunks [4], but still worse than for English and German [1].

Model	Known		Unknown			Total	
	acc.	σ	acc.	σ	(%)	acc.	σ
SUC_{def}	96.13	0.15	85.84	0.93	7.9	95.32	0.16
SUC_{lab}	96.22	0.12	86.13	1.06	7.9	95.43	0.14
SUC_{fol}	95.96	0.14	83.78	1.06	7.2	95.09	0.18
$\text{SUC}_{Megyesi}$	95.50	n/a	82.29	n/a	14.8	93.55	n/a
SUC_{Nivre}	95.66	n/a	38.05	n/a	7.8	91.46	n/a
NEGRA_{Brants}	97.7	0.23	89.0	0.72	11.9	96.7	0.29
WSJ_{Brants}	97.0	0.15	85.5	0.69	2.9	96.7	0.15

Table 1: Accuracy for SUC models on SUC.

In order to improve the results, we tried for each tunable parameter in turn all possible values (excluding weights for smoothing): 60 combinations for each model. Six taggers from the case-labelled model increased the overall accuracy to 95.44% (combinations of 6-, 7-, or 8- character suffix tries, and handling of unknown words using all words or singletons). All taggers from the case-labelled model had better or equal accuracy than their correspondants from the default model (≈ 0.1 points), and all taggers from the default model had better or equal accuracy than those from the case-folded model (≈ 0.2 points). Taggers using trigrams performed better than those using bigrams (≈ 0.3 points). The best taggers used linear interpolation smoothing, but otherwise the performance difference for smoothing are small and inconclusive, as is also the case for the handling of unknown words.

The best optimisation choice would therefore be to use case-labelling, and to look closer into suffix length, smoothing and unknown word handling, and, of course, to get more training data.

The more, the merrier?

To get more training data, we used the 77,6 million word Scarrie corpus [2], bootstrapped with the $\text{SUC}_{Megyesi}$ tagger [5]. The estimated tagging accuracy is 95.9%,

computed from a 4000 word validated sample, and the ratio of unknown words is 7.3% [5].

The bootstrapped corpus was used to train new models. The case-labelled model was also evaluated for the suggested optimisable tagging parameters on SUC (the best are shown in Table 2).

Model	Known		Unknown			Total	
	acc.	σ	acc.	σ	(%)	acc.	σ
Scarrie$^{def}_{SUC_{Megyesi}}$	96.52	n/a	84.61	n/a	2.7	96.20	n/a
Scarrie$^{lab}_{SUC_{Megyesi}}$	96.66	n/a	84.67	n/a	2.7	96.34	n/a
Scarrie$^{fol}_{SUC_{Megyesi}}$	96.18	n/a	80.66	n/a	2.3	95.82	n/a
Scarrie$^{lab.a8u2d2}_{SUC_{Megyesi}}$	96.88	0.13	84.10	1.02	2.7	96.53	0.15
Scarrie$^{lab.a8u3d2}_{SUC_{Megyesi}}$	96.88	0.13	84.11	1.01	2.7	96.53	0.15
Scarrie$^{lab}_{SUC_{lab}}$	96.91	n/a	85.48	n/a	2.7	96.60	n/a
Scarrie$^{lab.a8u2d2}_{SUC_{lab}}$	97.13	0.12	85.51	1.17	2.7	96.82	0.11
Scarrie$^{lab.a8u3d2}_{SUC_{lab}}$	97.13	0.12	85.50	1.17	2.7	96.82	0.11

Table 2: Accuracy for bootstrapped Scarrie models (a*n*=suffix length, u2=unknown words from all words, u3=unknown words from singletons, d2=additive smoothing, d3=linear interpolation smoothing).

We also retagged the Scarrie corpus with the SUC_{lab} model and the suggested optimisable tagging parameters. The best taggers are shown in Table 2. The major difference of the optimised taggers and the default tagger is that the best optimised taggers use additive smoothing. Suffix length and unknown word handling only contribute ≈ 0.01 points to the improvement.

Less is more?

The Scarrie$^{lab.a8u2d2}_{SUC_{lab}}$ and Scarrie$^{lab.a8u3d2}_{SUC_{lab}}$ models are comparable to the English and German taggers, but the accuracy has a price in storage space and tagging time (cf. Table 5). Therefore, we also tried a medium-sized corpus, Parole [6], of ≈ 20 million statistically tagged words (expected accuracy n/a) to see if it might do just as well. Taggers trained on the original tags did not do as well as Parole$_{SUC_{Megyesi}}$ taggers, which in turn did not do as well as Parole$_{SUC_{lab}}$ (see Table 3). Parole$_{SUC_{lab}}$ is also comparable to the English and German taggers.

Model	Known		Unknown			Total	
	acc.	σ	acc.	σ	(%)	acc.	σ
$\text{Parole}_{orig}^{def}$	96.13	n/a	76.05	n/a	3.5	95.44	n/a
$\text{Parole}_{orig}^{lab}$	96.23	n/a	75.88	n/a	3.5	95.53	n/a
$\text{Parole}_{orig}^{fol}$	95.94	n/a	74.48	n/a	3.1	95.28	n/a
$\text{Parole}_{orig}^{lab.a6u2d2}$	96.37	0.15	76.83	1.18	3.5	95.69	0.15
$\text{Parole}_{orig}^{lab.a6u2d2}$	96.37	0.15	76.83	1.18	3.5	95.69	0.15
$\text{Parole}_{orig}^{lab.a7u2d2}$	96.37	0.15	76.76	1.28	3.5	95.69	0.15
$\text{Parole}_{orig}^{lab.a7u3d2}$	96.37	0.15	76.73	1.28	3.5	95.69	0.15
$\text{Parole}_{SUC_{Megyesi}}^{lab}$	96.76	n/a	86.56	n/a	3.5	96.41	n/a
$\text{Parole}_{SUC_{Megyesi}}^{lab.a8u2d2}$	96.88	0.11	86.54	1.15	3.5	96.53	0.13
$\text{Parole}_{SUC_{Megyesi}}^{lab.a8u3d2}$	96.88	0.11	86.53	1.18	3.5	96.53	0.13
$\text{Parole}_{SUC_{lab}}^{lab}$	96.97	n/a	87.16	n/a	3.5	96.63	n/a
$\text{Parole}_{SUC_{lab}}^{lab.a7u2d2}$	97.11	0.11	87.09	1.23	3.5	96.77	0.12
$\text{Parole}_{SUC_{lab}}^{lab.a7u3d2}$	97.11	0.11	87.10	1.24	3.5	96.77	0.12
$\text{Parole}_{SUC_{lab}}^{lab.a8u2d2}$	97.11	0.11	87.26	1.22	3.5	96.77	0.12
$\text{Parole}_{SUC_{lab}}^{lab.a8u3d2}$	97.11	0.11	87.26	1.23	3.5	96.77	0.12

Table 3: Accuracy for bootstrapped Parole models.

United we are strong?

Is 77 million the limit, or is it possible to squeeze a bit more accuracy out of these data? We tried to join the two corpora and re-tag with SUC_{lab} and the same optimisable parameters as before, and got an overall accuracy of 97.02% for the best taggers (see Table 4).

Model	Known		Unknown			Total	
	acc.	σ	acc.	σ	(%)	acc.	σ
$\text{Scarrie+Parole}_{SUC_{lab}}^{lab}$	97.07	0.11	87.25	0.11	2.2	96.86	0.12
$\text{Scarrie+Parole}_{SUC_{lab}}^{lab.a8u2d2}$	97.22	0.10	87.86	1.26	2.2	97.02	0.10
$\text{Scarrie+Parole}_{SUC_{lab}}^{lab.a8u3d2}$	97.22	0.10	87.91	1.26	2.2	97.02	0.10
$\text{Scarrie+Parole}_{SUC_{lab}}^{lab.a10u2d2}$	97.22	0.10	87.99	1.25	2.2	97.02	0.10
$\text{Scarrie+Parole}_{SUC_{lab}}^{lab.a10u3d2}$	97.22	0.10	88.03	1.25	2.2	97.02	0.10

Table 4: Accuracy for bootstrapped Scarrie+Parole models.

Safety in numbers

We have shown that a PoS tagger trained on a small, but accurately tagged, corpus and used to bootstrap a tagger by training it on a considerably larger, automatically tagged, corpus, can result in a better tagger that overcomes some of the mistakes of the original tagger, just by the sheer number of occurrences of phenomena in the larger corpus.

Model	Memory	Loading	Tagging
SUC_{def}	23m	1.22s	7.34s
SUC_{lab}	24m	1.25s	8.17s
SUC_{fol}	21m	1.00s	6.73s
$Parole_{def}$	126m	16.71s	15.76s
$Parole_{lab}$	131m	17.13s	14.74s
$Parole_{fol}$	109m	12.75s	13.16s
$Scarrie_{def}$	255m	47.67s	17.46s
$Scarrie_{lab}$	263m	47.77s	19.66s
$Scarrie_{fol}$	228m	38.74s	16.46s

Table 5: Memory and time usage for loading all models and tagging SUC.

References

[1] Thorsten Brants. TnT - a statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP-2000)*, Seattle, Washington, 2000.

[2] Bengt Dahlqvist. 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. Department of Linguistics, Uppsala University, 1999.

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

[4] Joakim Nivre. Sparse data and smoothing in statistical part-of-speech tagging. *Journal of Quantitative Linguistics*, 7(1):1–17, 2000.

[5] Kristina Ohlander. Lingvistisk annotering av tidningstext och extraktion av verbala konstruktionsegenskaper [Linguistic annotation of newspaper texts and extraction of verb valency features]. Master's thesis, Department of Linguistics and Philology, Uppsala University, 2005.

[6] PAROLE. Parole. `http://scrooge.spraakdata.gu.se/lb/parole/`. Göteborg University. Department of Swedish and Språkbanken.

[7] SUC. Stockholm-Umeå corpus. Version 2.0. Stockholm University, Department of Linguistics and Umeå University, Department of Linguistics, 2002.