Part-of-speech taggers' performance reduces on novel data sets

Minerva Suvanto
Applied AI Group, Dept. M2, Chalmers University of Technology
minerva.suvanto@chalmers.se

Abstract

Part-of-speech (POS) tagging, a task in natural language processing (NLP), is commonly considered a *solved problem,* with reported accuracies of various methods reaching around 0.97. We show that when applied to a new test data set [1], standard taggers exhibit reduced performance. We approach fixing the issue by generating a set of corrective rules from completely unambiguous sentences [2].

Part-of-speech tagging

The task is to label a *text sequence* with parts-of-speech.

She	*had*	*to*	*duck*	*the*	*bird*	*flying*	*towards*	*her*
PRON	VERB	PREP	NOUN	DET	NOUN	VERB	PREP	PRON

Commonly reported accuracies reach 97% and the human error rate is 3% → is POS tagging a *solved problem?*

New data sets

Data set I [1] contains sentences where a single word is tagged, making it suitable for testing tagger performance.

Data set II [2] contains fully tagged, unambiguous sentences, and is suitable for testing and training.

Table 1: Accuracies of standard taggers on two common POS data sets and our two new data sets. Tagger accuracies drop on our data sets in comparison to the common data sets.

Tagger	Brown	PTB	Data set I	Data set II
Brill	0.966±0.001	0.903±0.001	0.473	0.516
Hunpos	0.919±0.001	0.987±0.000	0.482	0.521
Stanford	0.940±0.002	0.967±0.000	0.673	0.836
Perceptron	0.917±0.001	0.974±0.000	0.561	0.481
BiLSTM-CRF	**0.949±0.001**	**0.972±0.000**	**0.868**	**0.946**

Note! The lower performance of the first four taggers is not too surprising, as they have been used in the process for collecting the data sets.
The accuracy of the **final tagger**, however, shows a true drop in performance.

References

1. Wahde, M., Suvanto, M., & Della Vedova, M. L. (2024). A Challenging Data Set for Evaluating Part-of-Speech Taggers. In *ICAART (2)* (pp. 79-86).
2. Suvanto, M., Wahde, M., & Della Vedova, M. L. (2025), Part-of-speech Taggers Make Errors on Unambiguous Sentences. To appear in Lecture Notes in Artificial Intelligence (LNAI)

Where do the taggers fail?

One or more taggers failed on these sentences!

*Just **text** me when you get here*
→**NOUN**
TEMPORAL DRIFT

MORPHOLOGICAL SIMILARIES

*That **cursed** bird keeps stealing my milk!*
→**VERB**

*He was hired to **bus** tables*
→**NOUN**
SIMILAR CONTEXT

SURROUNDING TAGS

*He walked down the **lit** corridor*
DET →**NOUN** NOUN

NO GRAMMAR KNOWLEDGE

*The glue **sets** in five minutes*
DET NOUN →**NOUN** ADP NUM NOUN

Corrective approach

The word is tagged incorrectly!

She	*then*	*helloes*	*me*	*back*
PRON	ADV	**NOUN**	PRON	ADV
-2	-1	0	1	2

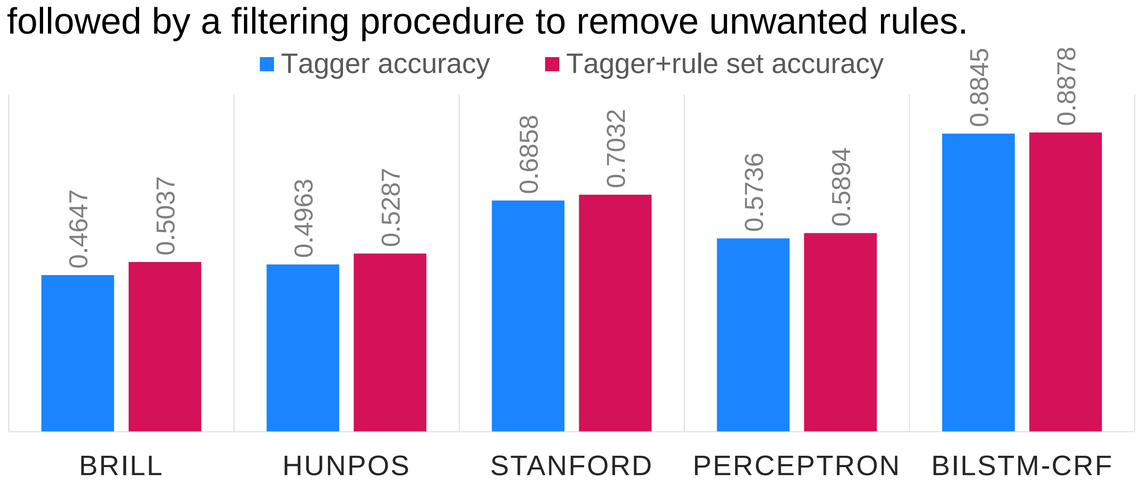
These are word positions relative to the target token.

IF POS(0) =NOUN ✓ IF POS(-1) =ADV ✓ IF POS(-2) =PRON ✓ IF POS(-3) =BOS*

*Beginning of a sentence

The rule corrects the tag! ← THEN ASSIGN POS(0)=VERB

Table 2: Accuracies on a subset of data set I before and after applying a set of rules of the type above. Rules are generated from data set II, followed by a filtering procedure to remove unwanted rules.



Improvements are small; future work is to enhance the rule set by sorting, filtering and merging the rules.

Conclusion? POS tagging is not as solved as claimed.

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM