

CompAn – A Tool for Quantitative Comparison of Corpus Annotation

Vlasta Ohlidalová

Lexical Computing
vlasta.ohlidalova@sketchengine.eu

RASLAN 2022
December 9, 2022

Presentation Overview

- 1 POS Tagging
- 2 Is it not good enough yet?
- 3 POS tagging evaluation
- 4 Gold standard
inconsistency
correctness
- 5 A tool for quantitative comparison of corpus annotation
- 6 Referencing

- Simple task – 90% in English by simply choosing the most often variant[1]
- The current results are close to 100% (95–97%)

Is it not good enough yet?

- Accuracy is counted from all tokens, not only words,
- accuracy will vary considerably for different text types,
- sentence accuracy:
 - a tagger success rate of 97% would mean sentence accuracy 45.6%,
 - for 95% accuracy on sentence level, we would need token accuracy 99.6%

Text types accuracy

Genre	Accuracy
child infections (report)	98.25%
political speech (labor union)	97.52%
job market news	97.46%
news report (school district)	97.10%
scientific news/medicine	96.88%
history (Gold War) report	96.67%
story about Holy Paul	95.42%
biological exposition	94.23%
movie description	93.89%
IT news/Cebit	93.69%
news report (Archbishop)	91.97%
information about a conference	90.98%
Rolling Stones tour (forum)	88.01%

Table: Statistics of TreeTagger POS tagging accuracy on various texts in

- comparison of the tagger results to gold standard
- issues of this approach:
 - trained and evaluated on the same type of text,
 - correctness of the gold standard

Gold standard

- consistent and correct
 - inconsistent/non-existent standard – 28%,
 - wrong gold standard – 15.5%[4]
- inconsistency among annotators
- incorrect annotation
 - 1: [tag="k1.*" & lemma="[:lower:].*ý"]
 - 2: [tag="k1.*" & lemma="[:lower:].*"] & 1.c=2.c within < s/>

hlediska	ze	pochopit	pohnutky	pozůstalých	a	známých	obětí	surových	a	zbytečných	vražd	, např			
k1gNnSc2	k6eAd1	k5eAaPmF	k1gFnPc4	k1gMnPc2	k8xC	k1gMnPc2	k1gFnPc2	k2eAgFnPc2d1	k8xC	k2eAgFnPc2d1	k1gFnPc2	k1x, k6e			
kou	zradu	"	.	</s><s>	Dnes	se	44	tisíc	mladých	občanů	naš	republiky	vojenské	službě	'
ld1	k1gFnSc4	k1x*			k6eAd1	k3xPyFc4	k4xCgInPc2		k1gMnPc2	k1gMnPc2	k3xOp1gFnSc2	k1gFnSc2	k2eAgFnSc3d1	k1gFnSc3	k5e
ledu	.	</s><s>	Vážný	důvod	, návštěva	těžce	nemocného	otce	doma	v Rusku	, byl	přesto	s		
iSc2			k2eAgInSc1d1	k1gInSc1	k1x, k1gFnSc1	k6eAd1	k1gMnSc2	k1gMnSc2	k6eAd1	k7c6	k1gNnSc6	k1x, k5eAaImAgInS	k8xC	k2eA	
run	.	</s><s>	"	Ve	výpovědích	se	oba	obvinění	příslušníci	v tomto	bodě	rozcházejí	, "	u	
nPc2			k1x*	k7c6	k1gFnPc6	k3xPyFc4	k4xCgMnPc1	k1gMnPc1	k1gMnPc1	k7c6	k3xDgInSc6	k1gInSc6	k5eAaImP3nP	k1x, k5eAa	

Figure: A few lines showing incorrectly annotated tokens in DESAM.

A tool for quantitative comparison of corpus annotation

- web application with a Python backend,
- uses corpora indexed by Manatee ((No)Sketch Engine)[3, 5],
- does not evaluate, only compares (manual annotation needed).

CompAn when comparing attribute (POS tag)











	Freq	rftagger	rftagger_synt	Conc	Conc
1	112	k1gInSc1	k4		
2	69	k1gMnSc1	k1gInSc1		
3	59	kF	k4		
4	45	k1gInSc4	k4		
5	39	k7c6	k7c4		

Figure: The example output of the tool when comparing attribute value (tags in this case)

CompAn when comparing words












	Freq	Word	rftagger	rftagger_synt	Conc	Conc
1	26	v	k7c6	k7c4		
2	14	pondělí	k1gNnSc6	k1gNnSc4		
3	8	na	k7c6	k7c4		
4	7	top	kA	k5nSp2		
5	7	to	k3gNnSc1	k3gNnSc4		

Figure: The example output of the tool when comparing words

Citing References

-  Charniak, E., Hendrickson, C., Jacobson, N., Perkowski, M.: Equations for part-of-speech tagging. In: Proceedings of the Eleventh National Conference on Artificial Intelligence. p. 784–789 (1993), <https://www.aaai.org/Papers/AAAI/1993/AAAI93-117.pdf>
-  Giesbrecht, E.: Evaluation of POS Tagging for Web as Corpus. Master's thesis
-  Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V.: The Sketch Engine: ten years on. *Lexicography* pp. 7–36 (2014)
-  Manning, C.: Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? vol. 6608, pp. 171–189 (05 2011). https://doi.org/10.1007/978-3-642-19400-9_14
-  Rychlý, P.: Manatee/Bonito-A Modular Corpus Manager. In: RASLAN. pp. 65–70 (2007)

Thank you for your attention