CompAn – A Tool for Quantitative Comparison of Corpus Annotation

Vlasta Ohlídalová

Lexical Computing vlasta.ohlidalova@sketchengine.eu

RASLAN 2022 December 9, 2022

Presentation Overview

- 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

POS Tagging

- 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

5/12

POS tagging evaluation

- 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

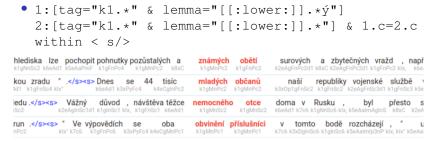


Figure: A few lines showing incorrectly annotated tokens in DESAM.

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	k1glnSc1	k4	o	8
2	69	k1gMnSc1	k1glnSc1	o	00
3	59	kF	k4	0	8
4	45	k1glnSc4	k4	0	0
5	39	k7c6	k7c4	8	00

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	o	00
2	14	pondělí	k1gNnSc6	k1gNnSc4	0	0
3	8	na	k7c6	k7c4	00	0
4	7	top	kA	k5nSp2	o	8
5	7	to	k3gNnSc1	k3gNnSc4	00	80

Figure: The example output of the tool when comparing words

Citing References

Charniak, E., Hendrickson, C., Jacobson, N., Perkowitz, 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