Comparative Analysis of Automatic POS Taggers Applied to German Learner Texts

Irina Kotiurova, Evgeny Maksimov, Polina Trenina, Andrey Solnyshkov
Petrozavodsk State University
Petrozavodsk, Russia
koturova@petrsu.ru, beaverstail31@gmail.com, trenina_p@mail.ru, solnyshkov.a48@gmail.com

***Abstract*—The article presents the analysis of testing and comparison of five part-of-speech taggers: CoreNLP, spaCy, TextBlob, RFTagger and TreeTagger, based on the texts from the annotated learner corpus of Petrozavodsk State University (PACT, Petrozavodsk Annotated Corpus of Texts). The conclusions were drawn about the frequency of errors in the part-of-speech identification, the tagging quality, weaknesses and strengths of each tagger.**

I. Introduction

In today's world everything is automated including the processes of thinking. Linguistics is also moving along the path of processing and analyzing big data, leaving the search and research material collecting made manually in the past. Corpus linguistics, which appeared just a few decades ago, has become one of the most promising, rapidly developing areas of science, opening up enormous opportunities for researchers and ordinary users.

The importance of Corpus technologies in linguistics increases fast, developing both extensively, in breadth, i.e. spreading to ever wider spheres, including pedagogy and prediction processes, and intensively, in depth, i.e. improving the quality and analysis capabilities of corpus data. The first corpus was Brown Corpus, being created in 1961 it contained 500 text fragments of 2 thousand words each [1]. In fact, the first linguistic corpora appeared in the USA, therefore the most extensive and most functional ones present English-language materials. However, the opened up prospects were appreciated in other countries that led to the creation of corpora in other languages showing high performance in the volume and language representation. For example, this was Russian National Corpus for the Russian language and Cosmas II for the German language. The effectiveness of their use has encouraged the appearance of numerous linguistic corpora, both large and small.

When starting to create a linguistic corpus, developers have to solve many conceptually important issues, one of them is the question of choosing a part of speech tagger (POS-tagger). The task of POS annotation includes giving a tag with the corresponding name of the part of speech to each token in the text [2]. At present there are various taggers available showing themselves successfully to a different extent depending on the language and types of text [3], [4], [5], [6].

One of the types of linguistic corpora is the learner corpus, its teaching and research possibilities are difficult to overestimate [7]. At Petrozavodsk State University the work on creating a corpus of learner texts (PACT - Petrozavodsk annotated learner corpus) in German is being in progress, for which it was necessary to choose one of five well-known taggers: TreeTagger, RFTagger, spaCy, CoreNLP, TextBlob. All these tools are applicable to the German language; however, learner texts have their own characteristics, primarily associated with a large number of errors. For instance, the noun spelling with a lowercase letter instead of a capital letter as well as mistakes in the grammatical forms may be critical for some POS taggers [8].

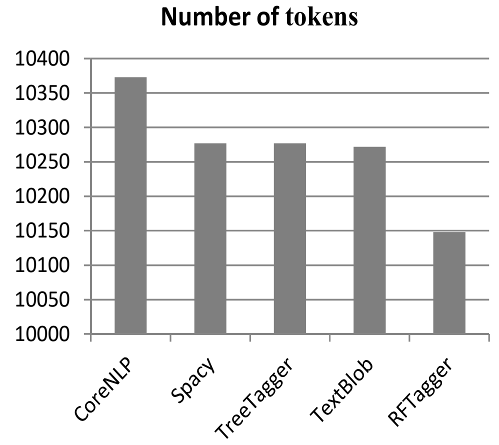
The rest of the paper is organized as follows. Section II describes the methodology, which we used for choosing the automatic POS-tagger for our learner corpus in German.
Section III shows experimental facts on the tagging by each of the five tools and make specific findings about the quality of these taggers applied to German learner texts.. Section IV summarizes the results of our experimental study.

II. Methodology

It was necessary to conduct a study, where TreeTagger, RFTagger, spaCy (version 2.3.0), CoreNLP and TextBlob were applied to 35 identical student texts with a total volume of about 10,000 tokens in order to compare the quality of the tagging. The total tokens number is named approximately not by chance. These taggers carrying out the part-of-speech identification of lexical units have a tokenization function as well, which precedes the definition of the part of speech of each selected token. At this stage significant differences are already observed.

So, spaCy divided the studied volume of student texts by 10277 tokens, RFTagger - by 10148 tokens, TextBlob - by 10272, TreeTagger - by 10277, and CoreNLP - by 10373.

The difference in the number of tokens in the same array of texts having been passed through different taggers is primarily due to the fact that CoreNLP assigns mistakenly part of speech labels to punctuation marks, such as whether they are verbs or nouns. A deeper analysis of the reasons for this discrepancy is a subject for an independent study. The number of part-of-speech tags as well as their designations differ in the analyzed instruments The CoreNLP and spaCy use the Universal POS tagset [9], the TextBlob uses the Penn TreeBank Tagset [10], and the TreeTagger and RFTagger use STTS ("Stuttgart-Tübingen Tagset") [11]. The STTS tagset was designed specifically for the German language, so it is supposed to have performed better.



III. ANALYSIS

Thus, having passed the same 35 texts through five different taggers, we got the following picture of the mistakes they made:

TextBlob

Out of 10272 tokens in TextBlob wrong are 1201 part-of-speech tags. Error rate is 11.69%.

- TextBlob uses a set of tags designed specifically for the English language. It contains tags that are not needed for tagging German texts (for example, VBP - verb, 3sg pres, VBZ - verb, non-3sg pres, VBG - verb, gerund). On the other hand, it lacks tags that are important for German (for example, to annotate possessive pronouns or separable verbs prefixes).

- TextBlob makes critically many mistakes related to non-distinction between singular and plural nouns. It is important to note that only TextBlob generally has different tags for singular and plural nouns.

- TextBlob does not have a special tag for tagging the separable prefix, so it regularly tags them incorrectly: most often as a particle.

- A lot of errors (44 errors) are associated with the assignment of the NNP - Proper Noun tag to a wide range of words and even punctuation marks, for example, such as Lieblingsfilm, ahnte, bewusst Spielzeug, Schlösser, offener and ".

SpaCy

SpaCy split the same 35 texts into 10277 tokens, i.e. much the same as TextBlob, but made 1022 errors in a part-of-speech tagging. Error rate is 9,94 %.

- Critically many errors in spaCy are associated with the definition of the lemmas *werden*, *sein* and *haben* (in various grammatical forms) as auxiliary verbs when they act as full-valued finite verbs.

- SpaCy assigns a verb tag to many adjectives with the ending *–en*.

- SpaCy regularly makes mistakes when identifying complex nouns, ordinal numbers and the particle *zu*.

- About a third of all errors in spaCy part-of-speech tagging is associated with incorrect identification of adjectives and verbs as nouns.

- SpaCy does not mark numbers, it does not matter whether we are talking about cardinal or ordinal numbers, which are followed by a period in a written German text (for example, 300 is a cardinal number, but 300. is an ordinal number).

CoreNLP

CoreNLP out of 10373 tokens mistakenly tagged 610. Percentage of error - 5.88%

- Like spaCy, this tagger very often taggs the full-valued verbs *werden*, *sein* and *haben* (in different grammatical forms) as auxiliary verbs.

- CoreNLP confuses some punctuation marks, tagging, for example, quotation marks with different parts of speech: PROPN, NUM, ADP, NOUN, etc.

- Similar to spaCy and TextBlob described above, the CoreNLP does not have a special tag for tagging separable prefixes, marking them as ADP - Adposition.

TreeTagger

523 out of 10277 tokens in TreeTagger are tagged wrong. Error rate - 5.08%

- A great number of TreeTagger errors appear in the words containing umlauts and c-cet (ä, ö, ü, ß). TreeTagger does not recognize these characters and, as a result, makes a lot of mistakes in determining the part of speech of the corresponding word.

- Many errors in the performance of the TreeTagger tagging are associated with the definition of the lemmas *werden, sein* and *haben* in all cases as auxiliary. Only TextBlob does not have this error, since there is no special tag for an auxiliary verb.

RFTagger

394 POS-tags out of 10148 tokens were recognized as errors. Percentage of error is 3,88%.

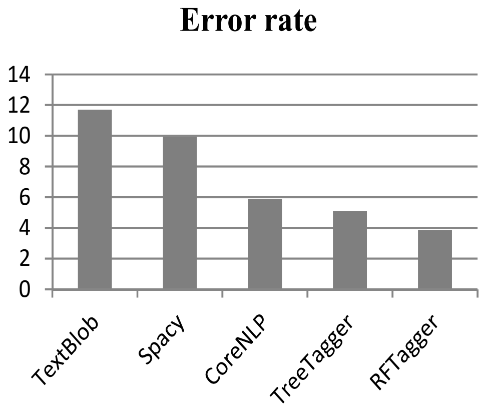
- Though RFTagger uses the STTS tag set, which has a special tag for the auxiliary verb VAFIN, it did not appear in the tagged texts. The auxiliary verbs werden, sein, haben are tagged, as well as the full-valued verbs werden, sein, haben, with the tag VVFIN (finite Verben, voll) or VVINF (Infinitiv, voll) correspondingly to the grammatical form.

- Like TreeTagger, RFTagger has problems with the words containing umlauts and ß.

- Like other markers, RFTagger does not recognize proper names (NE: Eigennamen), tagging them as common names (NN: normale Nomina)

However, the percentage of errors in tagging using RFTagger is lower than in other cases.

Thus, as you see in the diagram below, the highest percentage of errors in the tagging was shown by TextBlob, and the lowest - by RFTagger.



On the whole, 5 tools - TreeTagger, RFTagger, spaCy, CoreNLP, TextBlob - have shown themselves in different ways in tagging the same texts. Summary statistics can be seen in the table below.

TABLE I. SUMMARY STATISTICS OF TAGGING WITH 5 TAGGERS

	Number of tokens	Number of mistakes	Error rate
CoreNLP	10373	610	5,88 %
Spacy	10277	1022	9.94 %
TreeTagger	10277	523	5,08 %
TextBlob	10272	1201	11,69 %
RFTagger	10148	394	3.88 %

Thus, it should be noted that different taggers revealed different tendencies in the inaccuracy of certain elements. Among the most significant are the following:

- An obvious tendency is the incorrect part-of-speech tagging of elements containing umlauts (ä, ö, ü) and ß using RFTagger and TreeTagger.

- Many errors in all taggers are associated with the replacement of a common noun with a proper name and vice versa.

- All taggers make many mistakes of various kinds when identify proper names. For example, the name *Will* was defined by most taggers as a (modal) verb, and the element *von* in the proper noun *Alexander von Humbold* was defined as a preposition.

- In all taggers, the largest number of errors is associated with the incorrect verbs interpretation.

IV. Conclusion

Summing up the study, it can be argued that the focus of the STTS tag set on the German language distinguishes the TreeTagger and RFTagger among their counterparts in a positive way, which is statistically confirmed. In most cases, spelling mistakes and misprints of students do not affect the correctness of the identification of parts of speech by the taggers. The most successful for POS-tagging of German-language student texts is RFTagger, since the percentage of errors in its tagging is the lowest compared to other tools - 3.88%. Errors in many cases can be brought into accordance and, with the help of additional edits in the work of the tagger, they can be minimized.

V. Acknowledgement

This work is implemented within collaboration with the Artificial Intelligence Center of Petrozavodsk State University (PetrSU).

References

[1] O. Kholkovskaia, *Role of the Brown Corpus in the History of Corpus Linguistics*, POSTER, May, 2017.

[2] A. Díaz-Negrillo, D. Meurers, S. Valera, and H. Wunsch "Towards interlanguage POS annotation for effective learner corpora in SLA and FLT", *Language Forum,* Vol. 36, No 1-2., 2010. pp. 139-154.

[3] G. Soumitra, B.K. Mishra, , "Parts-of-Speech Tagging in NLP: Utility, Types, and Some Popular POS Taggers", *Natural Language Processing in Artificial Intelligence*, 2020, pp. 131.

[4] A.Maytham, A.Ramsay, "Improved POS-Tagging for Arabic by Combining Diverse Taggers", *IFIP Advances in Information and Communication Technology*, 2012, pp. 107–116.

[5] I. Rehbein, "Fine-grained pos tagging of german tweets", *In Language Processing and Knowledge in the Web*, Springer, 2013, pp. 162–175.

[6] G. Wisniewski, F. Yvon, "How Bad are PoS Tagger in Cross-Corpora Settings? Evaluating Annotation Divergence in the UD Project", *2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics*, Minneapolis, Minnesota, United States, Jun. 2019, pp. 218 – 227.

[7] I.A. Kotyurova, "Sozdanie korpusov uchebnykh tekstov kak razvivayushheesya napravlenie korpusnoj lingvistiki", *International Scientific Journal*, №5, 2020, pp. 100-109.

[8] Van Rooy, B. and Schäfer, L., "The effect of learner errors on POS tag errors during automatic POS tagging". *Southern African Linguistics and Applied Language Studies*, 20, 2009, pp. 325-335.

[9] Universal POS tagset, Web: https://www.sketchengine.eu/tagsets/universal-pos-tags/

[10] Penn TreeBank Tagset, Web: https://www.sketchengine.eu/penn-treebank-tagset/

[11] Stuttgart-Tübingen Tagset STTS, Web: https://www.sketchengine.eu/german-stts-part-of-speech-tagset/