A Study on Improving Corpus Creation by Pair Annotation

Ines Siebigteroth
FH Aachen, University
Jülich, Germany
siebigteroth@fh-aachen.de

Bodo Kraft
FH Aachen, University
Jülich, Germany
kraft@fh-aachen.de

Oliver Schmidts
FH Aachen, University
Jülich, Germany
schmidts@fh-aachen.de

Albert Zündorf
University of Kassel
Kassel, Germany
zuendorf@uni-kassel.de

Abstract

One of the most expensive steps in the development process of a natural language processing application is the annotation of corpora. We provide a study to evaluate the pair annotation approach for corpus development. This approach applies the process of pair programming, a well-established practice in the field of software engineering, to the annotation process.

We verify pair annotation by comparing two groups: One group consists of pair annotators, while the other group is made up of single annotators as a reference. Every group annotates a set of question and answer documents gathered from Stack Overflow. Evaluating the quality of annotations we apply the kappa measure of inter-annotator agreement between the annotated documents and a previously defined gold standard.

The results show that pair annotation can serve as an approach to improve initial skill training and quality of corpus creation.

2012 ACM Subject Classification Computing methodologies → Natural language processing; Software and its engineering → Pair programming

Keywords and phrases Corpus Creation, NLP, Pair Annotation, Pair Programming

Category Poster track

Acknowledgements We would like to express our gratitude to the student group of FH Aachen. Without their volunteering participation this study would not have been possible.

Corpus creation is a complex task, depending on domain and language used. For example, the precision of Part-of-Speech (POS) drops on domain specific text if they were not trained on this domain before [6]. Even a native speaker or linguist may have trouble correctly annotating domain specific language. Despite recent approaches to (semi-)automate the annotation process [9, 10, 11, 12], building reliable, domain specific corpora requires extensive manual work.

Kent Beck introduced the idea of pair programming [3], which is a standard technique today, to the software engineering community as optimization approach to the software development process. Demirşahin et al. [5] proposed pair annotation by adopting pair

LDK 2019 - Posters Track.
Editors: Thierry Declerck and John P. McCrae

programming to the process of corpus annotation to accelerate the annotation process while reducing errors and resolving disagreements faster. However, they mentioned, that their approach needs further evaluation, due to their small group of individuals and project specificity of their approach. We conduct this new study, evaluating the impact of pair annotation on corpus creation of non-native speakers.

Closely related to the advantages and disadvantages of pair programming we expected that the inter-annotator agreement between the gold standard and annotated documents by the pairs should be higher than documents annotated by single individuals, since the navigator checks the work of the driver and they discuss possible corrections immediately. These discussions may lead to a better understanding of the tags, improving results in future annotations. This quality improvement compensates the higher costs for two annotators in comparison to the costs for a single annotator.

Furthermore, we expected that the pairs would be more motivated and focused towards their task. As a result, higher quality of annotations and an enhanced learning curve with regard to increasing knowledge about the tags should be visible through Kappa statistics.

Gathering question and answer documents from the platform Stack Overflow [2], we selected the programming language Java as specific technical domain. Regarding to the international community of Stack Overflow the documents were written in English. We choose five documents, each consisting of 100 words, and apply the process of text segmentation to them. To evaluate the inter-annotator agreement we annotated a gold standard for each of the five documents.

As visualized in Figure 1, we divided a group of sixteen computer science students into two groups: one group consisting of five pair annotation teams and a reference group consisting of six single annotators. In order to ensure a realistic composition, especially without taking personal preferences into account, we took care of the group division and not the students themselves.

The skill of the students was on a similar level, because all have sufficient knowledge of the chosen domain. Furthermore, all were not familiar with any process of corpus annotation and they are neither native speakers nor linguists.

Annotators particularly used the Quick Pad Tagger (QPT) from NLPf [12] as a tool for the POS tagging process. For this study we chose the universal POS tag set[1], the default POS tag set from the NLPf framework [12], to simplify the annotation process for the experiment participants, and thus they are not native speakers or linguists.

The groups annotated the documents within five rounds, whereby each round lasted 30 minutes and only one document was annotated in each round. The defined time for each round was sufficient for each document. To recover, the students had a lunch break of 45 minutes after round three and a 15 minute break after each other round.

We interviewed them during the breaks and after the last round for evaluation regarding their experience to analyze their motivation and concentration over time. In addition, we conduct these interviews to investigate their learning experience during the study.

After the students had annotated all documents, we used the framework DKPro Statistics [7] to determine the inter-annotator agreement between the previously defined gold standard and the students' result by applying the implementation of the free-marginal kappa statistics as described by Brennan [4] and Randolph [8]. This kappa definition can be used when the annotators are not advised to classify a defined number of tokens to each category. Regarding to the concept of this study, the latter are used to analyze the results, where the categories correspond to the POS tags. We extended the framework by inter-annotator agreement of each category.

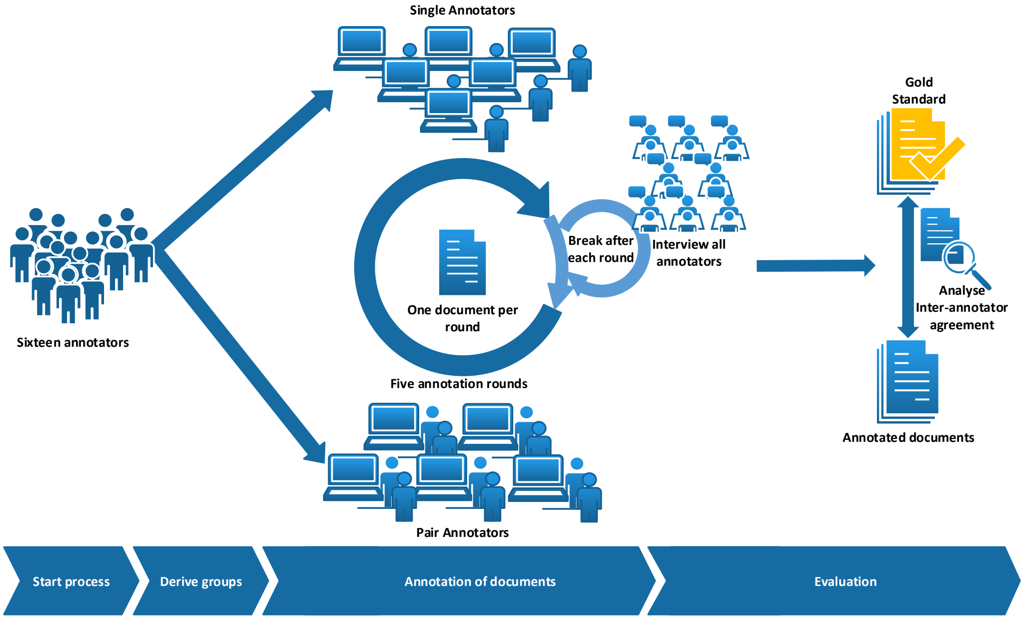


Figure 1 Visualization of annotation process

The formula for calculating the agreement of each category is shown in Equation 1.

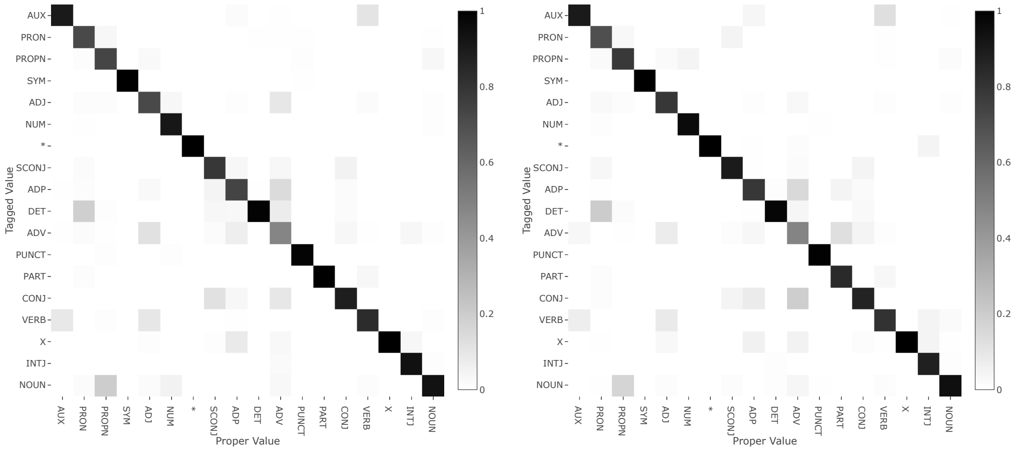
$$\kappa_i = \frac{P_{a_i} - P_e}{1 - P_e} \tag{1}$$

For each category i, the actual percentage of agreement P_{a_i} is determined for this category. The expected percentage of agreement P_e is 89% for every category. Furthermore, the average inter-annotator agreement for each group was calculated. Based on the determined kappa values, we evaluated the quality of the annotated documents and identified significant errors caused by confusion of POS tags respectively categories.

The kappa values for both groups, singles and pairs, indicated a high agreement and good quality overall with significant values always larger than 0.7. Furthermore, the values increased over time, which we interpreted as a learning curve. However, while even both groups achieved high kappa values, we recognized, that the pairs produced minimal better results. In addition, the variation of the kappa values of the single annotators was at least one and a half times as high as the variation of the pairs. The variation of the inter-annotator agreement of single annotators showed, that randomly chosen single annotators may have a different level of annotation skill, leading to unstable training data.

Based on the analysis of the variation, we supposed that the pairs are able to balance different skill levels leading to more stable results of higher quality. We interpreted this as an indication that working as a pair leads to higher quality training data. Furthermore, we interpreted this as an indication that the pairs are able to balance different skill levels leading to more stable results of higher quality than randomly chosen single annotators who may have a different level of annotation skill, leading to unstable training data.

Analyzing the confusion matrix as visualized in Figure 2, we observed that the diagonal of the matrix was particularly pronounced. We interpreted that most of the POS tags were set correctly. Nevertheless, the annotators confused the category pairs AUX and VERB, PRON and NOUN, CONJ and SCONJ between 20% and 40% of the time. Even more significant is



(a) Single

(b) Pair

Figure 2 Confusion matrix of categories per group

the analysis of the category pair ADJ and ADV. The inter-annotator agreement is located between 70% for adjectives and 50% for adverbs. Like the categories mentioned before the annotators confused them 20% of the time. Furthermore, they considered them as other categories.

We explained that by the fact, that the annotators were neither native speakers nor linguists, why it was difficult for them to determine the correct category in some cases. In the future we approve to give an introduction on English grammar to the annotators to improve the annotation of POS tags.

Nonetheless, we observed that the pairs were able to compensate the lack of grammar skills better by collaborating in a team. Regarding to this, we interpreted it, that it is easier for them to deal with missing knowledge. According to this, we recommend the usage of pair annotation to improve the skill on difficult tasks of the corpus annotation process.

In the interviews both groups of annotators described their motivation during the experiment without any difference, against our assumptions. However, Annotators mentioned they lost motivation, because of the repetitive annotation tasks after achieving a skill ceiling annotating the documents. They felt unable to improve further.

Regarding to the time of concentration and the experience of skill-improvement, we observed a difference between the groups. The pairs mentioned that they were able to concentrate better than the singles. Furthermore, they mentioned a steeper learning curve. Over all rounds the pairs reported their learning experience as higher as the single annotators. This self-evaluation fit to the higher and faster increasing inter-annotator agreement of the pairs, which confirms that the pairs achieve their skill maximum faster.

In the future the approach needs further verification with a larger group of pair and single annotators and more documents, with an introduction on grammar to the annotators to reduce the confusions that showed up during this study. Furthermore, we will investigate how the quality of the annotated corpus increases when we break up pairs and build new pairs consisting of single annotators with skill deficits and a better skilled individual from a broke up pair. Additionally, we will investigate the optimization of multi-staged corpus annotation processes by pair annotation, where Linguists will classify the POS tags at first and a group with domain knowledge will annotate the named entity relations (NER) next.

References

1 DKPro Core™ Type System Reference, January 2019. URL: http://dkpro.github.io/dkpro-core/releases/1.10.0/docs/typesystem-reference.html.

2 Stack Overflow - Where Developers Learn, Share, & Build Careers, January 2019. URL: https://stackoverflow.com/.

3 Kent Beck. *Extreme Programming Explained: Embrace Change*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2000.

4 Robert L. Brennan and Dale J. Prediger. Coefficient Kappa: Some Uses, Misuses, and Alternatives. *Educational and Psychological Measurement*, 41(3):687–699, October 1981. URL: http://journals.sagepub.com/doi/10.1177/001316448104100307, doi:10.1177/001316448104100307.

5 Işın Demirşahin, Ihsan Yalçınkaya, and Deniz Zeyrek. Pair Annotation: Adaption of Pair Programming to Corpus Annotation. In *Proceedings of the Sixth Linguistic Annotation Workshop*, LAW VI '12, pages 31–39, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL: http://dl.acm.org/citation.cfm?id=2392747.2392754.

6 Eugenie Giesbrecht and Stefan Evert. Is part-of-speech tagging a solved task? An evaluation of POS taggers for the German web as corpus. In *Proceedings of the fifth Web as Corpus workshop*, pages 27–35, 2009.

7 Christian M. Meyer, Margot Mieskes, Christian Stab, and Iryna Gurevych. DKPro Agreement: An Open-Source Java Library for Measuring Inter-Rater Agreement. In *Proceedings of the 25th International Conference on Computational Linguistics: System Demonstrations (COLING)*, pages 105–109, Dublin, Ireland, August 2014. URL: http://aclweb.org/anthology/C14-2023.

8 Justus Randolph. Free-Marginal Multirater Kappa (multirater κfree): An Alternative to Fleiss Fixed-Marginal Multirater Kappa. In *Advances in Data Analysis and Classification*, volume 4, Joensuu, 2005. URL: https://www.researchgate.net/publication/224890485_Free-Marginal_Multirater_Kappa_multirater_kfree_An_Alternative_to_Fleiss_Fixed-Marginal_Multirater_Kappa.

9 Marc Schreiber, Kai Barkschat, and Bodo Kraft. Using Continuous Integration to Organize and Monitor the Annotation Process of Domain Specific Corpora. In *5th International Conference on Information and Communication Systems (ICICS)*, pages 1–6, Irbid, Jordanien, 2014. Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/IACS.2014.6841958.

10 Marc Schreiber, Kai Barkschat, Bodo Kraft, and Albert Zundorf. Quick Pad Tagger : An Efficient Graphical User Interface for Building Annotated Corpora with Multiple Annotation Layers. In *Computer Science & Information Technology (CS & IT)*, pages 131–143. Academy & Industry Research Collaboration Center (AIRCC), February 2015. URL: http://www.airccj.org/CSCP/vol5/csit53513.pdf, doi:10.5121/csit.2015.50413.

11 Marc Schreiber, Bodo Kraft, and Albert Zündorf. Cost-efficient Quality Assurance of Natural Language Processing Tools through Continuous Monitoring with Continuous Integration. In *3rd International Workshop on Software Engineering Research and Industrial Practice*, pages 46 – 52, Austin, Texas, May 2016. URL: https://www.researchgate.net/publication/303330376_Cost-efficient_Quality_Assurance_of_Natural_Language_Processing_Tools_through_Continuous_Monitoring_with_Continuous_Integration, doi:10.1145/2897022.2897029.

12 Marc Schreiber, Bodo Kraft, and Albert Zündorf. NLP Lean Programming Framework: Developing NLP Applications More Effectively. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 1–5, New Orleans, Louisiana, 2018. Association for Computational Linguistics. URL: http://aclweb.org/anthology/N18-5001, doi:10.18653/v1/N18-5001.