Annotated corpora through time: a case study on NER

Yoann Dupont
Sorbonne Nouvelle / 8 avenue de Saint-Mandé, 75012 France
Lattice, UMR 8094 / 1 rue Maurice Arnoux, 92120 Montrouge
yoann.dupont@sorbonne-nouvelle.fr

1 Base process to version corpora

Annotated corpora are invaluable ressources that require a lot of human and organisational effort to complete. Once released, they often do not change anymore, even for corrections in the annotation. However, no corpus is ever perfect and some works aim to correct annotations in released corpora (Boudin and Hernandez, 2012; Reiss et al., 2020). Comparison with previous works make it hard to use these modifications as every comparison would need to be done twice (for old and new version). With this contribution we aim to spark a conversation on how we can overcome these limitations to provide more open and replicable studies with annotated. To this end, we will use the case study of Named Entity Recognition (NER).

Our proposition is to mimic the methodology already in use for software version control systems (VCS) to annotated corpora. We used semantic versioning[1] process with git[2] as a tool as a starting ground for this proposition. This way of versionning breaks versions into three main levels: major, minor and patch. While designed for software, we find that this way of versionning makes sense for annotated corpora as well, with some adaptations. Major stands for "incompatible API changes", which can be adapted to the following changes for corpora: changes in the annotation schema (tagset or scope).

Minor versions affect the corpus "in a backward compatible manner". We tend to consider a minor version is for adding or removing documents along their annotations, as they do not fundamentally change the essence of the annotation and are not simple corrections. Finally patches are corrections of the annotation of existing documents. Validating the whole annotation is required to realease of the corpus. While patches may be considered validated by default, this may not be true for new documents or changes in annotation schema. To this end, we recommend to capitalize on the branching possibilities of VCS systems: ongoing work should be in a development branch and have a main branch were only finalized versions should be present.

2 Case study on NER

We experimented this process on a Named Entity Recognition with the WiNER-fr corpus (Dupont, 2019). For this corpus, we used standoff annotations for two reasons. The first one is that it allows to preserve text without enforcing a tokenization, keeping this concerns separate. This is advantageous compared to "traditional" IOB format. This leads to the second reason: standoff annotation are used to minimalize changes. The tokenization process is subject to error, this could lead to non-minimal changes and make changes harder to read on a large scale. We used the BRAT format (Stenetorp et al., 2012) to represent annotations on a separate file from the original. This format represents "types" (here: entities) as a numeric identifier, a character span, a label and a string for validation on source text. Nesting of entities was handled by sorting entities top-down in the file. One problem that was encountered with the BRAT file format was the handling of identfiers, which were the index of the entity in the annotation list. When an entity was added or removed, this would modify all the remaining identifiers, leading to unnecessarily large diffs. We will explore how to prevent this issue by changing how the identifier is computed.

3 Limitations

These recommendations were provided for an NER case and may not be relevant for other tasks. They cover the neither the annotation process itself nor the creation of reproducible annotation tools.

[1]Official site: https://semver.org
[2]Official website: https://git-scm.com

References

Florian Boudin and Nicolas Hernandez. 2012. Détection et correction automatique d'erreurs d'annotation morpho-syntaxique du French TreeBank (detecting and correcting POS annotation in the French TreeBank) [in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 281–291, Grenoble, France. ATALA/AFCP.

Yoann Dupont. 2019. Un corpus libre, évolutif et versionné en entités nommées du français (a free, evolving and versioned french named entity recognition corpus). In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume II : Articles courts*, pages 437–446, Toulouse, France. ATALA.

Frederick Reiss, Hong Xu, Bryan Cutler, Karthik Muthuraman, and Zachary Eichenberger. 2020. Identifying incorrect labels in the CoNLL-2003 corpus. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 215–226, Online. Association for Computational Linguistics.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.