Re-Annotating a Named Entity Recognition Literary Dataset

Arthur Amalvy, Vincent Labatut and Richard Dufour

May 11, 2023

1 Introduction

Dekker, Kuhn, and Erp 2019 introduced a NER dataset of approximately one chapter of forty novels, consisting only of `PER` entities. We corrected a number of errors and inconsistencies in this dataset, and added annotations for `LOC` and `ORG` entities. All annotations and corrections were done by a single annotator. This document describes our correction and annotation process.

2 Dataset issues

The following issues were found [1]:

Annotation errors obvious annotation errors have been encountered:

- False negatives where a known character is not annotated as an entity (e.g. `‘‘d’Artagnan’’` was sometimes annotated as `O`)
- False positives where some tokens were incorrectly labeled as an entity (e.g. `‘‘Quai de Ferraille’’` was annotated `O B-PER O` in *The Three Musketeers*)

Annotation inconsistencies inconsistencies were found across the dataset novels. One prominent exemple is titles: depending on the novel, they would either be considered as part of the entity they are attached to or not (e.g. `‘‘Mr. Frodo’’` was found annotated as either `O B-PER` or `B-PER I-PER`)

Encoding issues Some characters, such as accented ones, were incorrectly represented (e.g. `d’Artagnan` was represented as `dâ€™Artagnan`).

[1] All issues examples use the BIO tagging scheme, while the original dataset use the IO scheme.

Tokenization issues They were a few tokenization issues. In the fantasy novel *The Wheel of Time*, characters with apostrophed names were tokenized in an unique way, where an apostrophe would be considered a full token (e.g. `‘‘Rand al’Thor’’` would be tokenized as `[Rand, al, ’, Thor]`). This rule was inconsistent with the rest of the dataset, and contrasted with the popular CoNLL-2003 dataset (Tjong Kim Sang and De Meulder 2003).

Quoting issues Quoting was handled inconsistently in the original dataset :

- Most books used double quotes to represent character dialogues and simple quotes for other purposes (irony, written text...), but some used the reverse convention.
- Quote pairs were usually started by backquotes (`` ` `` or `` ` ``) and ended by simple quotes (`''` or `'`). This has the advantage that recursive quoting (i.e. quotes inside of quotes) can be correctly identified. However, the convention was sometimes incorrectly applied : some quotes started with simple quotes, and some ended with backquotes.
- Some starting quotes were not matched with an end quote. This can happen when the original book's presentation prevent any ambiguity in knowing the end of the quote (for example, when the quote ends a paragraph). However, the ambiguity is present in the dataset since the book presentation is not preserved by the CoNLL format.

3 Method

3.1 Annotation Guidelines

To fix annotation errors and inconsistencies in the annotation of `PER` entities, and annotate `LOC` and `ORG` entities, we adopted the following guidelines:

- Titles preceding names were made part of entities (`Mr Bennet` and `Lord Eddard` were annotated as `B-PER I-PER`)
- Names referring to families (`Proudfoots` or `Bolgers` in *The Fellowship of the Ring*) were annotated as persons.
- Ethnonyms and demonyms (such as `Chyurda` in *Assassin's Apprentice*) were annotated as `O`.
- Nicknames made evident by typography (for examples with capitalisation or dashes, such as `You-Know-Who` in *Harry Potter*) were annotated as entities.
- Discontinuous entities were annotated discontinuously (e.g. `Mr and Mrs Bennet` was annotated `B-PER O B-PER I-PER`).

- Capitalised common names (such as `the Director` in `Brave New World`, `the Mouse` in *Alice in Wonderland* or `the Messenger` in *The Painted Man*) were annotated according to the context. If the name was referring to a *function*, we annotated it as `O`. Otherwise, it was treated as a person.
- Nicknames following characters names were annotated as part of the entity (`Gandalf the Grey` was annotated `B-PER I-PER I-PER`). A notable exception is when the character is designated by including an organization or a location : in those cases, organizations and locations were annotated separately (`Lord Stark of Winterfell` was annotated `B-PER I-PER O B-LOC`, and `Lord Eddard of House Stark` was annotated `B-PER I-PER O B-ORG I-ORG`).
- Geopolical entites that are also locations were annotated as `ORG` depending on context (`France declared war over Germany` was annotated `B-ORG O O O B-ORG` since only geopolitical entities can take action).

3.2 `PER` Annotations Correction

We followed a semi-automatic error correction process, consisting of several steps applied in order:

False Negatives Rule We noticed a lot of false negatives in the existing annotation (e.g. `Sherlock Holmes` would be annotated `O O`). To fix those issues, we retrieved a list of characters names for all books. For each of those names, we also generated a list of alternative names using simple rules (e.g. {``Mr Sherlock Holmes''} => {``Mr Sherlock Holmes'', ``Mr Sherlock'', ``Mr Holmes'', ``Holmes'', ``Sherlock''}). Starting with longer names, when a mention of one of those names was not annotated as a person, we fixed the annotation. Due to its high observed precision, this step was performed automatically, with no action from the annotator.

False Positives Rules We asked the annotator the correct the following cases if needed:

- When a series of tokens was annotated as a person, but was not found in the list of characters previously established.
- When a series of tokens was annotated as a person, but all of those tokens were lowercase.

Bert Assisted Correction To catch the remaining errors, we finetuned BERT (Devlin et al. 2019) on a modified version of the CoNLL-2003 dataset [2]. We then asked the annotator to correct the annotation of spans of text if the prediction from BERT was different from the annotation.

[2] Modifications consisted of the inclusions of titles such as `Mr` or `Lord` as part of the entities, to remain consistent with our annotation guidelines

Manual Correction In last resort, we manually fixed the remaining errors we could find.

3.3 Quoting

To fix quoting issues, we applied the following guidelines :

- Character utterances should be enclosed in double quotes, starting with `` and ending with ''.
- Other use of quoting (such as, but not limited to: irony, written text, nicknames...) should be enclosed in single quotes, starting with ` and ending with '.
- A closing quote should be added when an utterance is not closed.

We once again adopted a semi-automatic method :

1. We executed a program that prompted the user for each single quote (since in some books, single quotes were used in place of double quotes). For each single quote, the user was able to choose if that quote should be replaced or not by an appropriate double quote, and if this choice for the current pattern (composed of the previous and the current token) should be automatically applied when seeing the pattern in the future.
2. We inspected quotes with less than 5 characters, or with more than 100 characters. Short quotes were often not part of dialogues, and therefore were inspected since a lot of them were enclosed in double rather than simple quotes. Long quotes were inspected since there were a high chance these could spuriously appear when an ending quote was missing.
3. We inspected quotes that had an uneven number of quotation marks, since these were malformatted.
4. We performed final corrections manually.

3.4 Other Fixes

Tokenization and encoding issues were fixed manually. The dataset was also converted from the IO NER tagging scheme to the BIO scheme.

4 Metrics

NOTE: metrics computation was done before we implemented the quoting fixs and annotated `LOC` and `ORG` entities. Some books now can't be compared since annotations are not aligned anymore (we added or removed some tokens).

To highlight the difference between our dataset and the version from Dekker, Kuhn, and Erp 2019, we computed the mean precision, recall and F1 score of their annotations compared with ours. Results can be found in table 1.

Precision	Recall	F1
93.93	75.02	82.51

Table 1: Mean precision, recall and F1 score of the annotations from Dekker, Kuhn, and Erp 2019 compared with ours

References

Dekker, N., T. Kuhn, and M. van Erp (2019). "Evaluating named entity recognition tools for extracting social networks from novels". In: *PeerJ Computer Science* 5, e189. DOI: 10.7717/peerj-cs.189.

Devlin, J. et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv* cs.CL, p. 1810.04805. URL: https://arxiv.org/abs/1810.04805.

Tjong Kim Sang, E. F. and F. De Meulder (2003). "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition". In: *7th Conference on Natural Language Learning*, pp. 142–147. URL: https://aclanthology.org/W03-0419.