BERT meets d'Artagnan: Data Augmentation for Robust Character Detection in Novels

Arthur Amalvy and **Vincent Labatut**
Laboratoire Informatique d'Avignon (LIA), Avignon Université, France
arthur.amalvy@univ-avignon.fr
vincent.labatut@univ-avignon.fr

Richard Dufour
Laboratoire des Sciences du Numérique de Nantes (LS2N), Nantes Université, France
richard.dufour@univ-nantes.fr

1 Introduction

Character detection in novels is the task of detecting which characters are present and where they appear in a novel. It is a useful task, as it can be used to automatically extract information from literary works: as an example, it is a necessary step of character network extraction (Labatut and Bost, 2019), which is a subject of interest in digital humanities. Despite the apparent simplicity of the task, it requires solving several natural language processing (NLP) problems, such as Named Entity Recognition (NER) and disambiguation. While NER is a well-studied task in the NLP community (Li et al., 2020; Yadav and Bethard, 2018), most datasets used to train and evaluate models do not cover the literary domain. Popular datasets, such as CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003), Ontonotes v5 (Weischedel et al., 2011) or WikiGold (Balasuriya et al., 2009), contain mainly newswire and web-based texts. The LitBank dataset (Bamman et al., 2019) focuses on literary texts, but is concerned with nested NER, an arguably harder task. Because of this difference in coverage, most NER models see their performance decrease on literary texts. Dekker et al. (2019) performed an evaluation of several NER tools on approximately one chapter of forty novels. They noted that performance varied wildly based on the considered novel, and that models were plagued by recurring issues, such as the detection of apostrophed names (`"D'Artagnan"`) or word names (`"Mercy"`). We show that a BERT-based model (Devlin et al., 2019) trained on an out-of-domain dataset is able to mitigate those issues. To tackle the remaining ones, we propose a simple data-augmentation scheme, that can be used to adapt an out-of-domain corpus to help models perform more robust character detection.

2 Method

2.1 Dataset

To evaluate our performance, we use the dataset of Dekker et al. (2019), consisting of approximately one annotated chapter of forty novels. Those novels are separated into two groups: a group of "old" classic works, and a group of "new" works, mostly consisting of novels from the fantasy genre.

We found that the dataset suffered from various errors and inconsistencies. Therefore, we established a set of guidelines that we consistently applied using a semi-automatic system to re-annotate the dataset.

2.2 Training

We finetune a pretrained BERT-base model with a token classification head for 2 epochs (Devlin et al., 2019) on the CoNLL-2003 NER corpus (Tjong Kim Sang and De Meulder, 2003). As in Dekker et al. (2019), we feed our model each sentence, as well as the previous and next ones for context.

2.3 Data Augmentation

Data augmentation has seen an increase of interest in NLP recently (Feng et al., 2021), and we hope that it can be used to adapt a corpus from a domain to another: this would allow leveraging the high number of NER corpus available to train NER systems in the literary domain or other low-resources

domains.

To test this hypothesis, we devise a scheme to adapt CoNLL-2003 to the literary domain. We generate new synthetic examples by taking existing samples from the corpus and replacing, label-wise, person entities by other person entities randomly sampled from a predetermined list. We experiment with two different lists, tailored to tackle problems encountered by Dekker et al. (2019):

- A list of 493 word names, that we obtain by taking the intersection of a set of English first names and the set of English common nouns retrieved using WordNet (Miller, 1995).
- A list of 1,324 fantasy names, consisting of characters from the game "The Elder Scrolls III: Morrowind". We chose this game because we needed a big-enough list of characters with typical fantasy names that was not from the studied novels.

3 Results

We evaluate our models on our re-annotated version of the corpus from Dekker et al. (2019). We compute precision, recall and F1-score as in the CoNLL-2003 shared task (Tjong Kim Sang and De Meulder, 2003). Since Dekker et al. (2019) provides the output for the four systems they tested, we can compare our results with theirs even though our datasets are different. Table 1 shows the mean F1 of each NER system on our corrected dataset. BERT (Devlin et al., 2019) significantly outperformed previous tools.

3.1 Discussion

In this section, we delve further into issues described by Dekker et al. (2019), and show the benefits that can be gained using our proposed data augmentation technique:

Apostrophed Names contrarily to models tested by Dekker et al. (2019), BERT is able to correctly detect most apostrophed names, such as `"D'Artagnan"`. However, some complex fantasy names are harder to detect. This is the case in *The Wheel of Time*, with characters such as `"Rand al'Thor"`. However, training BERT on our fantasy augmented version of the CoNLL-2003 reduced the number of those recall errors from 7 to 1, and improved the score by 2.66 F1 points for this novel.

Word Names Dekker et al. (2019) gives *The Black Company* as an example of a novel where most characters are designated using a code name (`"Mercy"`, `"Croaker"`), which resulted in poor performances for the systems they evaluated. While BERT trained on the original CoNLL-2003 detected most mentions of those characters, a few hard cases remained (`"Dancing"`, `"Silent"`). By using the CoNLL-2003 corpus augmented using our word names list, the model was able to better detect these names. Such recall errors went from a number of 10 to 6, and we gained 3.93 recall points. Precision, however, decreased by 4.07 points. While we argue that recall is a more interesting metric than precision because it is easier to filter wrongly detected characters than to retrieve non-detected ones, further work is required to alleviate this effect.

4 Conclusion

Our results show that a pretrained transformer-based model such as BERT is able to mitigate issues plaguing more classical NER models on literary texts. While issues remain, our presented data augmentation scheme shows promise in tackling them, and might also prove useful to adapt popular NER corpus to other domains as well.

References

R. Agerri and G. Rigau. 2016. Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence*, 238:63–82.

D. Balasuriya, N. Ringland, J. Nothman, T. Murphy, and J. R. Curran. 2009. Named entity recognition in Wikipedia. In *Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 10–18.

D. Bamman, S. Popat, and S. Shen. 2019. An annotated dataset of literary entities. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 2138–2144.

D. Bamman, T. Underwood, and N. A. Smith. 2014. A bayesian mixed effects model of literary character. In *52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 370–379.

N. Dekker, T. Kuhn, and M. van Erp. 2019. Evaluating named entity recognition tools for extracting social networks from novels. *PeerJ Computer Science*, 5:e189.

BookNLP (2014)	Illinois (2009)	IXA (2016)	Stanford CoreNLP (2005)	BERT (2019)
75.83	64.97	59.17	62.81	**91.02**

Table 1: Mean F1 of different NER systems on our modified version of the dataset from Dekker et al. (2019)

J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, cs.CL:1810.04805.

S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988.

J. R. Finkel, T. Grenager, and C. Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *43rd Annual Meeting of the Association for Computational Linguistics*, pages 363–370.

V. Labatut and X. Bost. 2019. Extraction and analysis of fictional character networks : A survey. *ACM Computing Surveys*, 52:89.

J. Li, A. Sun, J. Han, and C. Li. 2020. A survey on deep learning for named entity recognition. *arXiv*, cs.CL:1812.09449.

G. A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

L. Ratinov and D. Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics.

E. F. Tjong Kim Sang and F. De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *7th Conference on Natural Language Learning*, pages 142–147.

R. Weischedel, E. Hovy, M. Marcus, M. Palmer, R. Belvin, S. Pradhan, L. Ramshaw, and N. Xue. 2011. OntoNotes: A large training corpus for enhanced processing.

V. Yadav and S. Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *27th International Conference on Computational Linguistics*, pages 2145–2158.