Leveraging LLMs to Improve Human Annotation Efficiency with INCEpTION

Luís Filipe Cunha[1,3,6(✉)], Nana Yu[2,3], Purificação Silvano[2,3], Ricardo Campos[1,4,5], and Alípio Jorge[1,3]

[1] INESC TEC, Porto, Portugal
{luis.f.cunha,ricardo.campos,alipio.jorge}@inesctec.pt
[2] CLUP, Porto, Portugal
robertananayu@hotmail.com, msilvano@letras.up.pt
[3] University of Porto, Porto, Portugal
[4] University of Beira Interior, Covilhã, Portugal
[5] Ci2 - Smart Cities Research Center - Polytechnic Institute of Tomar, Tomar, Portugal
[6] University of Minho, Braga, Portugal

Abstract. Manual text annotation is a complex and time-consuming task. However, recent advancements demonstrate that such a task can be accelerated with automated pre-annotation. In this paper, we present a methodology to improve the efficiency of manual text annotation by leveraging LLMs for text pre-annotation. For this purpose, we train a BERT model for a token classification task and integrate it into the INCEpTION annotation tool to generate span-level suggestions for human annotators. To assess the usefulness of our approach, we conducted an experiment where an experienced linguist annotated plain text both with and without our model's pre-annotations. Our results show that the model-assisted approach reduces annotation time by nearly 23%.

Keywords: Annotation of Corpora · Language Models · INCEpTION

1 Introduction

Over the last few years, Large Language Models (LLMs) have assumed a central role in data annotation tasks, often replacing humans to reduce costs and time. [4]. While this approach offers certain benefits, LLMs frequently produce inaccurate annotations, creating a need for additional oversight and quality control [5,9]. One possibility, as referred by Pangakis in his work [8] is to validate the annotations generated by LLMs against human-annotated labels, thus highlighting the importance of human-annotated data. Generating gold-standard annotated corpora remains, however, a highly complex and time-consuming task [1,2].

In this work, we present a methodology to accelerate manual text annotation by leveraging LLMs to pre-annotate text with span-level suggestions in the context of a token classification task. This approach allows the human annotator

C. Hauff et al. (Eds.): ECIR 2025, LNCS 15576, pp. 53–58, 2025.
https://doi.org/10.1007/978-3-031-88720-8_10

to shift its focus from performing manual annotations from scratch to reviewing and refining the model's output. To accomplish this, we fine-tuned a BERT model on the token classification task using the LUSA corpus [7]. This model was then integrated into INCEpTION [6], a state-of-the-art text annotation tool. This integration allows our model to pre-annotate the raw text, which can then be revised in INCEpTION by the human annotator. Although demonstrated on a specific annotation framework and task, this methodology can be easily adapted for other annotation schemas within token-level tasks. To evaluate our methodology, we resorted to an expert linguist who annotated a set of documents, with and without our model's assistance. We then compared the time taken to complete the annotation task in each scenario to assess the model's impact on efficiency. The results obtained show that our approach accelerates the annotation task by 23%.

Our contributions are two-fold. First, we trained a token classification model that can be integrated with INCEpTION for automated text pre-annotation. Second, we extended INCEpTION recommender library, allowing integration with any Hugging Face token classification model.

Our demo is available at https://nabu.dcc.fc.up.pt/inception/login.html with the following credentials: Username: `demo`; Password: `demouser`. A walk-through video of the demo is also available at https://vimeo.com/1023744554.

2 Methodology

In this paper, we trained an LLM to pre-annotate text as a means to improve the efficiency of manual data annotation. To showcase our methodology, we will particularize with the pre-annotation of a series of Portuguese texts.

To this regard, we fine-tuned a BERT model [11] for the token classification task using the LUSA corpus [7,10], a collection of Portuguese news articles manually annotated by a team of expert linguists, focused on the narrative extraction task. This dataset includes span-level annotations with the following labels: *Time*, *Event*, *Spatial Relation*, and *Participant*, which were used to train a model capable of identifying these labels. The resulting model is publicly available[1] and has been integrated into INCEpTION using for this a Python library[2], which we extended to support models from Hugging Face Hub. Our contribution has been merged into the original library, significantly broadening the range of models that can be incorporated into INCEpTION, since Hugging Face Hub is one of the largest repositories for LLMs.

Figure 1 provides an overview of the implemented methodology and its workflow. The raw documents are first uploaded into INCEpTION, which then connects to the external recommender server. This server hosts our BERT model, which generates the document's pre-annotations. INCEpTION then presents these to the human annotator as suggestions. The human expert can then easily validate these suggestions by accepting, rejecting, modifying or extending them

[1] https://huggingface.co/lfcc/lusa_events.
[2] https://github.com/inception-project/inception-external-recommender.

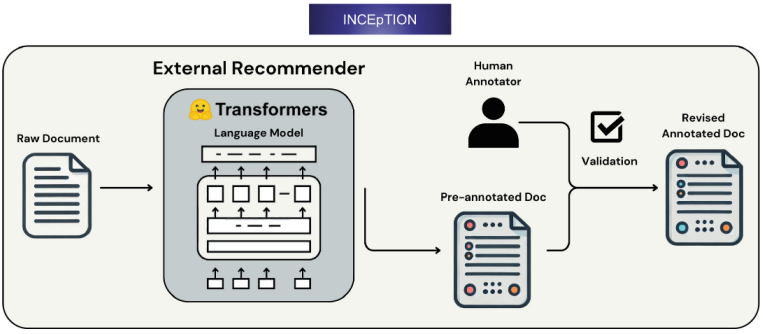

Fig. 1. Overview of the annotation workflow.

as needed. Figure 2 shows the INCEpTION annotation interface, with model suggestions highlighted in grey and the accepted entities in their respective colours. In this figure, the annotation process is still ongoing, with some pre-annotations yet to be validated by the linguist.

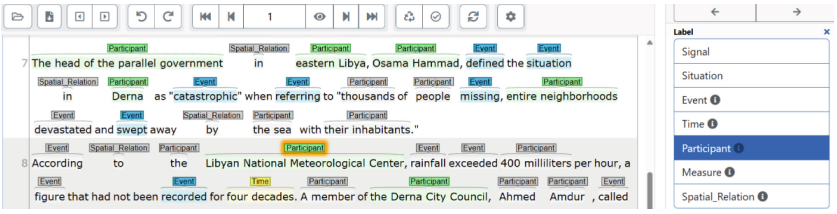

Fig. 2. INCEpTION annotation interface displaying model-generated pre-annotations.

For the purposes of this demonstration, we also trained an English model. To accomplish this, we translated the LUSA dataset into English using Google Translate and the LinguaAligner library [3][3], to align the annotations with the translated text. A linguist validated the translation. The resulting dataset was then used to train the English model to be integrated into our demo. To showcase our approach's versatility, we also used an additional token classification model[4] trained on the CoNLL-2003 [12] dataset, as shown in the demo video.

3 Evaluation and Results

To evaluate our trained pre-annotations models (Portuguese and English), we assess their effectiveness on the LUSA dataset. Then, we apply them to a real-

[3] https://github.com/lfcc1/LinguAligner.
[4] https://huggingface.co/dslim/bert-base-NER.

world data annotation scenario and discuss the results. The LUSA corpus consists of 117 documents. Of these, 98 were used for training both models and 19 for validation.

Table 1. Evaluation results for the Portuguese and English BERT Models

Label Type	Portuguese LUSA Model				English LUSA Model			
	Precision	Recall	F1	Sup	Precision	Recall	F1	Sup
Event	85.94	86.99	86.46	492	71.14	79.19	74.95	442
Participant	80.32	82.71	81.50	538	63.28	68.43	65.75	491
Spatial Relation	38.37	62.26	47.48	53	18.07	60.00	27.78	50
Time	70.59	76.19	73.28	63	65.33	80.33	72.06	61
Overall	79.10	83.25	**81.12**	1,146	60.52	73.28	**66.29**	1,044

Table 1 summarizes the effectiveness of our models on the token classification task by label type, with the Portuguese and English models achieving average micro F1-scores of 81.12% and 66.29%, respectively, the latter reflecting expected performance differences due to training on translated data. Regarding the label types, both models struggled to extract Spatial Relation entities, likely due to the smaller number of examples combined with a wider variety of cases.

To evaluate the practical impact of our models on data annotation, an expert linguist manually annotated 20 documents. Of these, 10 documents were pre-annotated using our Portuguese model, while the remaining 10 were annotated from scratch. The time taken to annotate each document was recorded. For a fair comparison, documents were paired based on a similar number of tokens.

Table 2. Annotation Time Measurement

With Pre-annotation												
Doc Id	120	123	126	104	125	122	106	145	142	117	Total	Avg
#Tokens	364	355	340	334	331	203	219	221	259	280	2906	290.6
Time (min)	25	30	26	24	26	23	23	20	21	23	241	**24.1**
Without Pre-annotation												
Doc Id	150	159	182	113	177	130	143	168	163	185	Total	Avg
#Tokens	365	356	340	333	331	203	219	222	258	280	2907	290.7
Time (min)	36	35	34	35	33	29	28	27	28	27	312	**31.2**

Table 2 presents the time measurements for the manual annotation process. By comparing the average annotation time measurements for documents with pre-annotations (24.1 min) and those without pre-annotations (31.2 min), we

conclude that using our model's recommendations allowed the expert to annotate the documents 22.76% faster, saving an average of 7.1 min per document, which contained an average of 291 tokens. This demonstrates that our model significantly improved the efficiency of the annotation process.

In future work, we aim to improve our methods to enable the pre-annotation of relations between entities, assisting with more complex annotation tasks.

Acknowledgments. This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020. DOI https://doi.org/10.54499/LA/P/0063/2020. Luís Filipe Cunha thanks the Fundação para a Ciência e Tecnologia (FCT), Portugal for the Ph.D. Grant (2024.04202.BD). The authors also would like to acknowledge project StorySense, with reference 2022.0931 2.PTDC (DOI https://doi.org/10.54499/2022.09312.PTDC). The authors also thank Richard Eckart de Castilho from the Technical University of Darmstadt, Germany, for his valuable support during the development of this work.

References

1. Baledent, A., Mathet, Y., Widlöcher, A., Couronne, C., Manguin, J.L.: Validity, agreement, consensuality and annotated data quality. In: Calzolari, N., et al. (eds.) Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 2940–2948. European Language Resources Association, Marseille, France, June 2022. https://aclanthology.org/2022.lrec-1.315
2. Beck, J.: Quality aspects of annotated data. AStA Wirtschafts- und Sozialstatistisches Archiv **17**(3), 331–353 (2023). https://doi.org/10.1007/s11943-023-00332-y
3. Cunha, L.F., Silvano, P.a., Campos, R., Jorge, A.: Ace-2005-pt: Corpus for event extraction in Portuguese. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, pp. 661–666. Association for Computing Machinery, New York (2024). https://doi.org/10.1145/3626772.3657872
4. Gilardi, F., Alizadeh, M., Kubli, M.: Chatgpt outperforms crowd workers for text-annotation tasks. Proc. Natl. Acad. Sci. **120**(30), e2305016120 (2023). https://doi.org/10.1073/pnas.2305016120
5. Huang, F., Kwak, H., An, J.: Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In: Companion Proceedings of the ACM Web Conference 2023, WWW '23 Companion, , pp. 294–297. Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3543873.3587368
6. Klie, J.C., Bugert, M., Boullosa, B., Eckart de Castilho, R., Gurevych, I.: The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In: Zhao, D. (ed.) Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations. pp. 5–9. Association for Computational Linguistics, Santa Fe, New Mexico, August 2018. https://aclanthology.org/C18-2002

7. Nunes, S., et al.: Text2Story lusa: a dataset for narrative analysis in European Portuguese news articles. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 15773–15782. ELRA and ICCL, Torino, Italia, May 2024. https://aclanthology.org/2024.lrec-main.1370
8. Pangakis, N., Wolken, S., Fasching, N.: Automated annotation with generative AI requires validation. CoRR abs/2306.00176 (2023). https://doi.org/10.48550/ARXIV.2306.00176
9. Reiss, M.V.: Testing the reliability of chatgpt for text annotation and classification: a cautionary remark (2023). https://arxiv.org/abs/2304.11085
10. Silvano, P., et al.: Text2story lusa annotated corpus [data set] (2023). https://doi.org/10.25747/ESFS-1P16
11. Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: pretrained BERT models for Brazilian Portuguese. In: 9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear) (2020)
12. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, pp. 142–147 (2003). https://aclanthology.org/W03-0419