Assessing Transfer Learning and automatically annotated data in the development of Named Entity Recognizers for new domains.

Emanuel Matos[1], Mário Rodrigues[2], António Teixeira[1]

[1]IEETA, DETI, University of Aveiro, Aveiro, Portugal
[2]IEETA, ESTGA, University of Aveiro, Aveiro, Portugal
{easm,mjfr,ajst}@ua.pt

Introduction

The identification of entities is a key step in many natural language processing (NLP) tasks and consequently named entity recognition (NER) is a task important in several contexts, text genres, and languages. The approaches to NER include sta- tistical models based on handcrafted templates that are instantiated on final rules using training data and, more recently, deep learning models that do not require designing the rules.

Proposal

The process consists of two parts:
1. Cross-domain application of a BERT-based NER model previously trained with automatically annotated data (top part of the Figure);
2. Fine-tuningofthisbasemodeltoanewdomainresorting to automatically annotated data for that domain.

For this initial exploration of cross-domain BERT-base NERs, the previously explored domain of Tourism was selected and complemented with the Legal domain.

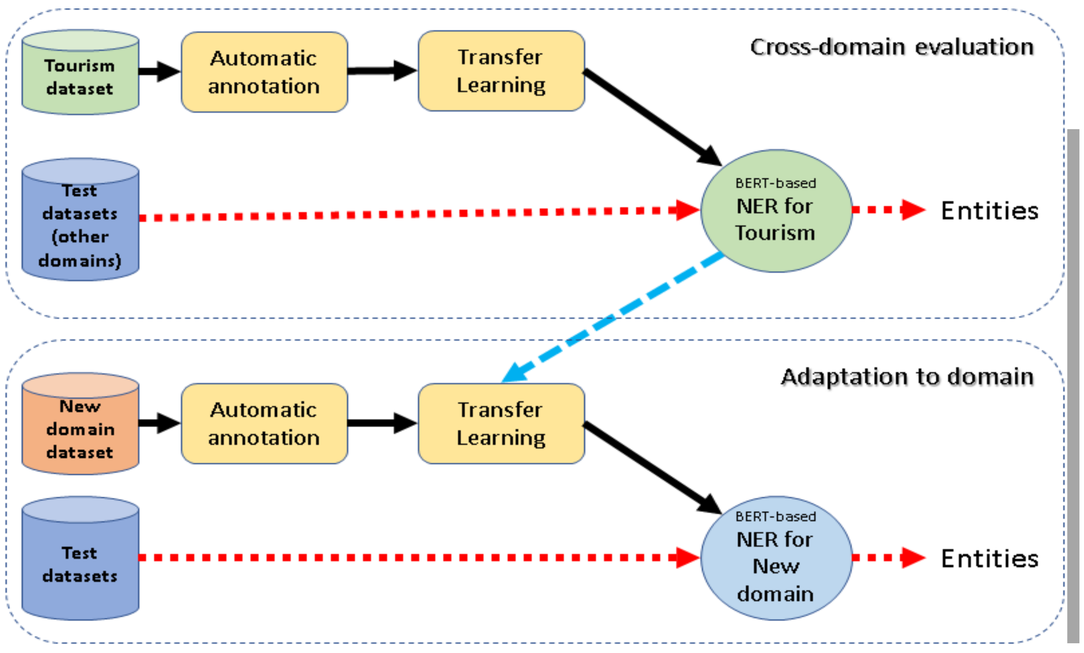


Fig. 1. Transfer Learning Idea

Results

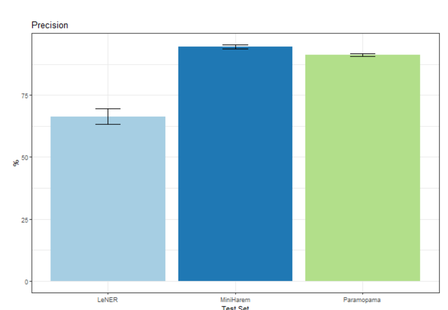
- Globalresults

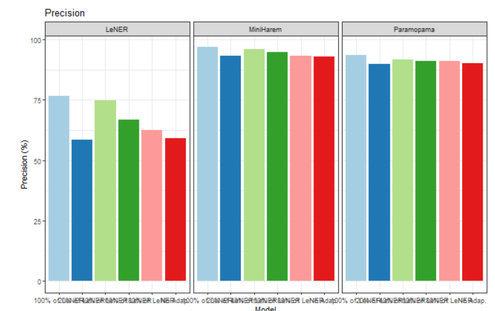
The results for the selected metrics as function of the NER model and test-set are presented in Table. For easier understanding of the results, part of the information in the Table is presented graphically in Figures.
Table and plots show that:

• The results obtained for LeNER legal dataset are worst than the ones obtained for the two other test sets, being particularly worst in terms of recall;
• Results for Paramopama are close to the one obtained for MiniHAREM, confirming for a new dataset the results reported in figures

• Despite showing some reduced contribute to improve Precision, the additional training (fine-tuning) of the model with LeNER train dataset resulted in degradation of Recall, for all 3 test sets. The best Recall is obtained without additional training for the 3 datasets, being the best overall score obtained with MiniHAREM.

Domain Adaptation	Paramopama			LeNER (test set)			MiniHAREM		
	Prec.	Rec.	F1	Prec	Rec	F1	Prec.	Rec.	F1
No Adaptation	90.33	44.53	59.65	59.04	20.17	30.07	92.83	53.98	68.26
20 % of LeNER	89.75	39.10	54.47	58.52	10.38	17.63	93.25	50.58	65.59
40 % of LeNER	91.67	16.51	27.98	74.87	3.85	7.32	96.12	33.92	50.14
60 % of LeNER	91.25	30.72	45.97	66.67	4.20	7.90	94.93	38.44	54.72
80 % of LeNER	91.00	33.36	48.82	62.56	7.74	13.78	93.17	49.24	64.43
100 % of LeNER	93.68	22.06	35.71	76.79	1.59	3.12	96.87	24.49	39.10





Related Work

Named Entity Recognition, as many tasks in NLP, has been ap- proached in two very different ways: systems based in rules and lists (the so-called Gazetteers) and data-driven systems base in machine learning. Machine Learning approaches to NER can be more flexible but they depend on the existence of adequate datasets for the target domain.
Through the years several machine learning methods have been applied to NER, being good results obtained with, for example, Support Vector Machines (SVM), Conditional Ran- dom Field (CRF) or Neural networks (NN). To address this limita- tion, the authors, in [9], developed NER systems for PT based on BERT using automatically annotated data. They adopted Transfer Learning, finetuning pretrained BERT models with an automatically annotated dataset for the Tourism domain, based in Wikivoyage texts. Best F1 obtained was 64.9 %

Datasets

- **Wikivoyage** Tourism dataset – resulting from automatic an- notation of texts from Wikivoyage.
- **LeNER** Legaltextsdataset– Consisting of 60 texts from LeNER dataset of legal documents in Brazilian Portuguese.
- **Paramopama** datasets – Extends the PtBR version of WikiNER corpus, revising incorrect assigned tags in order to improve corpus quality, also extend the corpus size and provide proper evaluation[24]. This dataset has a total of 240,755 words tagged as part of an entity, considering 4 types (PERSON, LOCATION, ORGANIZATION and TIME).
- **HAREM** datasets – Two HAREM datasets were used, the First HAREM and the MiniHAREM, both having manually annotated entities.

Conclusions

Addressing the problem of development of NERs for new do- mains without annotated datasets, this paper explores the cross- domain potential of application of BERT-based NER models with and without tuning with in-domain additional training. As expected, models had a much higher difficulty in recog- nizing entities of classes not present in the training dataset(s). Nevertheless, the Recall of approx. 10% for both LEGISLATION and JURISPRUDENCE in LeNER without any in- domain training is an interesting result, with potential to be ex- plored in bootstrap methods.
Additional training with a small amount of in-domain data (automatically annotated) was not very useful. Despite some improvement in Precision (e.g., for PERSON in LeNER), it contributed to degradation of Recall.

References

E. Matos, M. Rodrigues, P. Miguel, and A. Teixeira, “Towards Automatic Creation of Annotations to Foster Development of Named Entity Recognizers,” in 10th Symposium on Languages, Applications and Technologies (SLATE 2021), ser. Open Access Series in Informatics (OASIcs), R. Queiro´s, M. Pinto, A. Simo˜es, F. Portela, and M. J. a. Pereira, Eds., vol. 94. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum fu¨r Informatik, 2021, pp. 11:1–11:14. [Online]. Available: https://drops.dagstuhl.de/opus/volltexte/2021/14428

ACKNOWLEDGEMENTS

This research was supported by IEETA Research Unit, funded by National Funds through the FCT - Foundation for Science and Technology, in the context of the project UIDB/00127/2020.