UNIVERSITÀ DI PISA

BureauBERTo: Adapting UmBERTo to the Italian bureaucratic language

Serena Auriemma, Mauro Madeddu, Martina Miliani, Alessandro Bondielli, Lucia C. Passaro, Alessandro Lenci

Ital-IA 2023, 3rd National Conference on Artificial Intelligence

Workshop: AI per la Pubblica Amministrazione

May 29th, 2023 - CNR, Pisa

Introduction

- ➢ Transformer-based language models have advanced the SoTA in NLP, but are sensitive to domain shifts.
- ➢ Many domain-adapted LMs have been developed for the biomedical and scientific domains, as BioBERT (Lee et al., 2020) and SciBERT (Beltagy et al.,2019); and for the legal and financial ones, like LegalBert (Chalkidis et al., 2020) and FinBert (Araci, 2019)
- ➢ For Italian, released LMs related to the bureaucratic sector are Italian-Legal-BERT, ArchiBERTo, LamBERTa.
- ➢ Despite the growing deployment of transformer-based models in similar domains, a specific pre-trained model for the bureaucratic language is still missing.

Our main contributions

- We introduce BureauBERTo, the first transformer-based model adapted to the Italian bureaucratic language.
- And we address the following research questions:

1. *What is the overlap among the vocabularies of our target technical-bureaucratic domains?*
2. *To what extent the vocabulary expansion is beneficial for the domain-adaptation of BureauBERTo? Does further pre-training affect the semantic representation of words?*
3. *What are the advantages of employing a domain-specific vs. a generic model in a downstream task?*

BureauBERTo

- We initialized our model starting from UmBERTo, which is the best generic model for handling administrative data (Auriemma et al. 2022).
- We performed domain adaptation mostly following the same hyperparameters of RoBERTa .

Epochs	40
steps	17400
batch size	~8k
learning rate	5e-5
Adam	$\beta1=0.9$, $\beta2 = 0.98$
weight decay	0.1

BureauBERTo

- We expanded the model vocabulary with 8,305 new domain terms selected by applying the TF-IDF on a composite corpus containing PA, banking, and insurance documents (i.e., the **Bureau Corpus**).
- We reached a total vocabulary size of 40,305 token types, increasing the model size from 110M to 117M params.
- Finally, we trained UmBERTo with a MLM objective randomly masking 15% of the tokens.

The Bureau Corpus

- We constructed the Bureau Corpus by selecting:
  - administrative acts of several Italian municipalities;
  - banking public communications, circulars, and provisions;
  - a collection of non-life insurance product information documents.

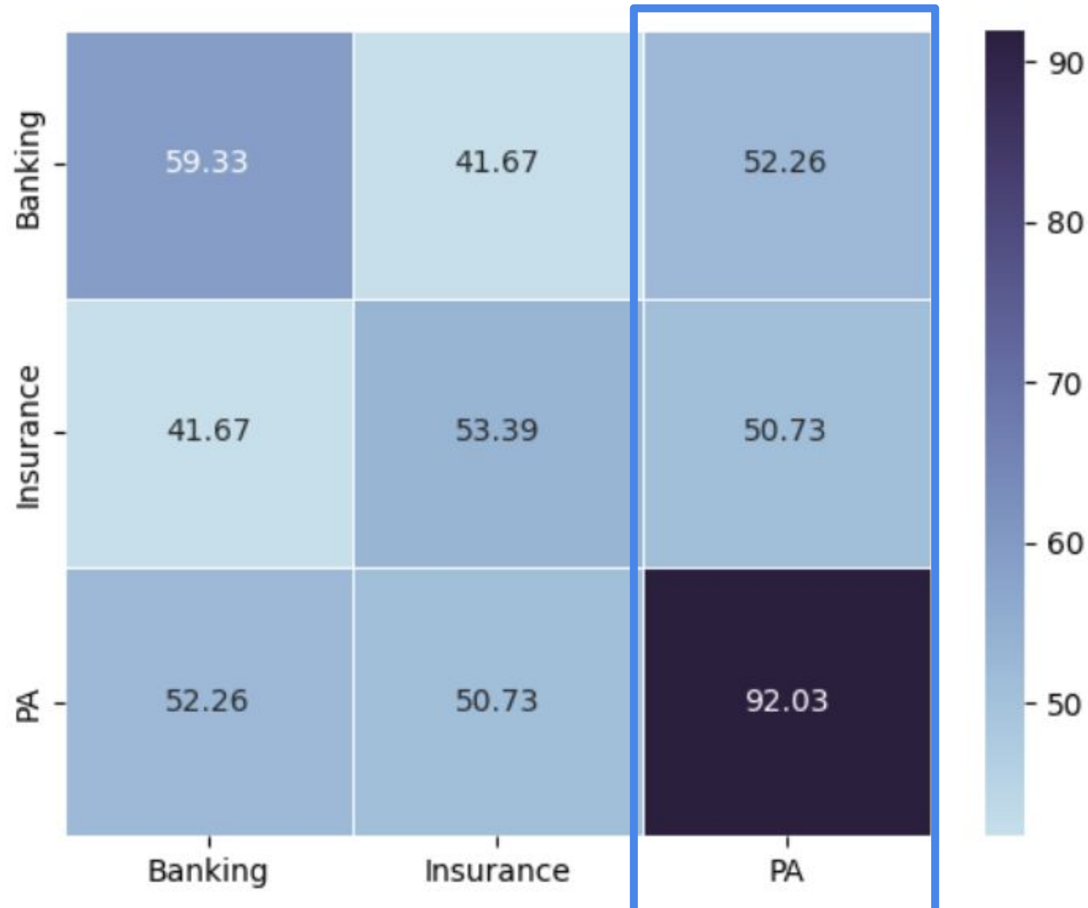
Dataset size, number of sentences, and percentage of each domain data (in terms of sentences) in the Bureau Corpus.

Domain	Size	N.sents	% of domain data
PA	4.3 GB	23,176,626	65.7%
Banking	1.8 GB	7,835,289	22.2%
Insurance	674 MB	4,281,311	12.1%
Bureau Corpus	**6.7 GB**	35,293,226	100%

Research Questions

1. *What is the overlap among the vocabularies of our target technical-bureaucratic domains?*

- 21.6% of the tokens are exclusive of the PA domain
- 3.6% of the tokens occur only in banking documents
- 0.3 % of the tokens appear only in insurance texts.



Research Questions

2. *To what extent the vocabulary expansion is beneficial for the domain-adaptation of BureauBERTo? Does further pre-training affect the semantic representation of words?*

- ➢ We assessed the effectiveness of our domain adaptation with a fill-mask task.
- ➢ We measure the model accuracy in predicting:
  - top-k (where $k \in K$ = {1, 3, 5, 10}) candidates
  - for random and domain specific words

Sentence	k	UmBERTo	BureauBERTo
...determina la cessazione della presente ~~copertura~~ assicurativa ed il rimborso del Premio pagato da parte della Compagnia all'Impresa...	1	'garanzia' (47.76%)	**'copertura'** (94.00%)
	2	'polizza' (25.88%)	'garanzia' (4.06%)
	3	**'copertura'** (14.48%)	'polizza' (0.47%)
	4	'Convenzione' (1.82%)	'Convenzione' (0.38%)
	5	'convenzione' (1.51%)	'prestazione' (0.23%)

Research Questions

2. *To what extent the vocabulary expansion is beneficial for the domain-adaptation of BureauBERTo? Does further pre-training affect the semantic representation of words?*

- BureauBERTo improves over UmBERTo in both masked word prediction tasks across all datasets.
- The gap between the two models widens when masking only in domain words

Domain	k	Random		In-dom.+in-voc.	
		UmB.	BB	UmB.	BB
PA - ATTO	1	29.81%	39.74%	46.16 %	61.46
	3	39.94%	50.70%	67.48%	83.16%
	5	43.32%	53.75%	72.82%	86.09%
	10	47.21%	57.49%	78.76%	88.93%
Banking	1	30.51%	36.33%	52.82%	58.27%
	3	42.58%	48.99%	69.78%	74.72%
	5	47.07%	53.62%	75.75%	80.34%
	10	52.29%	58.97%	81.82%	86.11%
Insurance	1	28.62%	41.68%	43.61%	62.51%
	3	40.42%	53.78%	60.02%	77.72%
	5	44.70%	57.59%	66.60%	81.94%
	10	49.79%	62.21%	74.08%	87.12%

Research Questions

3. *What are the advantages of employing a domain-specific vs. a generic model in a downstream task?*

- We fine-tuned BureauBERTo in a PA-specialized **NER task** to compare its performance with those of the generic transformer model UmBERTo (Auriemma et al., 2022) and INFORMed PA, a PA-specialized Stanford NER (Passaro et al., 2017).
- All models were trained on the INFORMed PA corpus, a collection of 460 documents from the *Albo Pretorio Nazionale,* annotated with:
  - standard NER entities (i.e., person, locations, and organizations)
  - and in-domain classes: LAW (national legislation), ACT (PA acts), and ORG*PA* (PA organizations, like city hall's offices).

Research Questions

3. *What are the advantages of employing a domain-specific vs. a generic model in a downstream task?*

➢ Results

- BureauBERTo obtained a significant improvement on the in-domain class ORGPA (+4%)
- Other domain-specific entities are better recognized by BureauBERTo
  - Slightly better performance for ACT, LAW and PER

Model	Measure	ACT	LAW	LOC	ORG	ORG$_{PA}$	PER	MicAvg	MacAvg
UmBERTo	P	0.916	0.846	0.808	0.795	0.785	0.908	0.858	0.872
	R	0.942	0.877	0.841	0.838	0.828	0.900	0.890	0.899
	F1	0.928	0.861	**0.824**	**0.816**	0.806	0.904	0.873	0.885
INFORMed PA	P	0.788	0.827	0.702	0.709	0.616	0.837	-	0.74
	R	0.891	0.842	0.740	0.689	0.777	0.878	-	0.803
	F1	0.836	0.834	0.720	0.698	0.686	0.857	-	0.772
BureauBERTo	P	0.915	0.863	0.761	0.776	0.790	0.915	0.850	0.868
	R	0.951	0.877	0.805	0.859	0.912	0.927	0.899	0.914
	F1	**0.932**	**0.870**	0.783	**0.816**	**0.846**	**0.921**	**0.874**	**0.890**

Conclusions and future work

- The experiments suggest that generalizing PA, Insurance, and Banking domains to the "bureaucratic" one is effective for the transfer.
- Additional aspects of the adaptation need to be studied more in depth (e.g., downstream tasks for all the sub-domains).
- In the future, we plan to:
    - perform additional experiments for additional downstream tasks
    - challenge our model to solve tasks on a different, albeit close domain, such as the legal one. This will assess the transfer-learning capabilities of BureauBERTo to other bureaucratic domains.