BureauBERTo: adapting UmBERTo to the Italian bureaucratic language

Serena Auriemma[1,*], Mauro Madeddu[1], Martina Miliani[2,1], Alessandro Bondielli[3,1], Lucia C. Passaro[3] and Alessandro Lenci[1]

[1] *Dipartimento di Filologia, Letteratura e Linguistica, Università di Pisa, Via Santa Maria 36, Pisa, 56126, Italy*

[2] *Università per Stranieri di Siena, Piazzale Carlo Rosselli 27/28, Siena, 53100, Italy*

[3] *Dipartimento di Informatica, Università di Pisa, Largo B. Pontecorvo 3 Pisa, 56127, Italy*

Abstract

In this work, we introduce BureauBERTo, the first transformer-based language model adapted to the Italian Public Administration (PA) and technical-bureaucratic domains. We further pre-trained the general-purpose Italian model UmBERTo on a corpus of PA, banking, and insurance documents, and we expanded UmBERTo's vocabulary with domain-specific terms. We show that BureauBERTo benefitted from the adaptation by comparing it with UmBERTo in both an intrinsic and extrinsic evaluation. The intrinsic evaluation has been conducted through specific fill-mask experiments. The extrinsic one has been faced with a named entity recognition task on one of the sub-domains in BureauBERTo.

Keywords

NLP, Domain Adaptation, Transformers, Evaluation, Italian Bureaucratic Language, Public Administration

1. Introduction

The use of artificial intelligence (AI) in the context of Public Administration (PA) serves a dual purpose: increasing the efficiency of public entities by expediting data management processes and ensuring greater transparency, allowing citizens easier access and use of public documents. Since their first appearance in 2017 [1], transformer-based models have been leveraged in many ways to create models adapted to specific domains and effective in performing several downstream tasks [2, 3, 4, 5, 6]. Similarly, in the context of Italian PA, it could be advantageous to tailor a pre-trained model to such domain, as Italian administrative lingo differentiates semantically and syntactically from standard Italian. The Italian administrative lexicon is, indeed, characterized by extensive use of technicisms (e.g., *ravvedimento operoso, imponibile, capitolato*), some of which are directly derived from the legislative language, Latinisms (e.g., *una tantum; pro capite*), archaisms (e.g., *testè, quantunque*), neologisms (e.g., *esternalizzare* [*to entrust a task to an external body*]), and Anglicisms (e.g., *governance, front-office*). Texts are also rich of abbreviations, acronyms, legislative references, and formulaic or stereotypical expressions, such as *entro e non oltre, e successive modifiche ed integrazioni.* The presence of lengthy and syntactically complex sentences with recurring prepositional chains and subordinate clauses [7] also contributes to the peculiarity and difficulty of the administrative language.

It is important to note that these linguistic characteristics are not peculiar to Public Administration, but they also pertain to a broader group of domains with a massive use of bureaucratic language. In fact, we can find the same or similar characteristics in legal, banking, or insurance texts. This is the reason why, in the framework of the ABI2LE project (*ABility 2 LEarn*), we aim at exploiting transformer-based models and the possibilities offered by transfer-learning techniques, to develop a suite of NLP tools to automatically extract information from these domain-specific texts. In a previous work, [8] compared the performance of five generic transformer-based models on two main tasks in the administrative domain: a multi-label classification of PA documents and a PA-specialized Name Entity Recognition (NER) task, to identify the best-performing model to adapt to the PA domain. Domain adaptive pre-training (DAPT) has proven to be an effective technique to exploit off-the-shelf pre-trained models and obtain substantial gains in their performance on domain data simply by further training the model with domain texts [9]. Extending [8], we chose to additionally pre-train the general purpose model UmBERTo[1] on administrative, banking, and insurance corpora,[2] creating **BureauBERTo**, the first transformer-

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy

*Corresponding author.

serena.auriemma@phd.unipi.it (S. Auriemma)

0009-0006-6846-5826 (S. Auriemma); 0009-0002-7844-3963 (M. Madeddu); 0000-0003-1124-9955 (M. Miliani); 0000-0003-3426-6643 (A. Bondielli); 0000-0003-4934-5344 (L. C. Passaro); 0000-0001-5790-4308 (A. Lenci)

CEUR Workshop Proceedings (CEUR-WS.org)

[1] https://huggingface.co/Musixmatch/umberto-commoncrawl-cased-v1

[2] We excluded the legal domain since a model adapted to this domain for the Italian language already exists (see Sec. 2).

based model adapted to the Italian bureaucratic language (Section 3). Before training, we expanded the model vocabulary with about 8k new domain terms selected as the most frequent ones in the new corpus.

In this work, we address the following questions: (i) **What is the overlap among the vocabularies of our target technical-bureaucratic domains?** We estimated the overlap of the terms added to the BureauBERTo vocabulary occurring in texts related to the administrative, insurance, and banking domains (Section 3.3). (ii) **To what extent the vocabulary expansion is beneficial for the domain-adaptation of BureauBERTo? Did further pre-training affect the semantic representation of words?** We compared UmBERTo and BureauBERTo accuracy in the fill-mask task to assess the contribution of further pretraining (Section 4.1) (iii) **What are the advantages of employing a domain-specific vs. a generic model in a downstream task?** We evaluated BureauBERTo and UmBERTo performances on Named Entity Recognition (NER) in the administrative domain (Section 4.2).

2. Related Work

Pre-trained transformer-based models such as BERT [10] and its variants [11, 12, 13, 14] have achieved state-of-the-art performances in several NLP downstream tasks, many of which included in generic benchmark datasets such as GLUE [15] or SQUAD [16]. However, generic models tend to under-perform specialized ones when applied to domain-related texts and tasks. This has led to the development of models trained on domain-specific data, like the medical-scientific [2, 4] or legal fields [5, 17, 18, 19]. Following [5], who proposed the first legal domain-specific BERT further pre-trained on English legal documents, [17] created ITALIAN-LEGAL BERT by additionally pre-training the Italian BERT version on civil law corpora. Their domain-adapted model achieved better results on NER for the legal domain and the classification of sentences belonging to different sections of civil judgments. Another model fine-tuned on the Italian legal domain is LamBERTa [20], trained for retrieving the most pertinent civil code article to a given legal query. Italian legal texts share with administrative ones some linguistic features typical of the Italian bureaucratic language that contribute to making the language of the Italian Public Administration rather complex and artificial [21]. To improve the accessibility to public information in PA documents, [22] adapted a Neural Pairwise Ranking Model based on BERT architecture to assess the readability level of sentences extracted from Italian administrative texts.

Another peculiar aspect of the PA domain is that it comprises several sub-domains, each corresponding to a different sub-sector of the PA. For the Construction sector, [23] created ArchiBERTo, a multi-label sentence classifier to individuate the sentences corresponding to the criteria and quality objectives required by the public appointing party in the Design Guidance Document (*Documento di Indirizzo alla Progettazione*, DIP).

Despite the growing deployment of transformer-based models in the PA sector, a specific model for this domain is still missing. We, therefore, decided to create BureauBERTo, the first transformer model trained to understand Italian bureaucratic language.

3. BureauBERTo

We initialized our model starting from UmBERTo, which is the best generic model for handling administrative data [8]. UmBERTo is a cased Italian monolingual model based on RoBERTa [11]. It is trained using a SentencePiece tokenizer and Whole Word Masking on a large subset of the OSCAR corpus of approximately 70 GB of text. We additionally trained UmBERTo with a MLM objective (randomly masking 15% of the tokens), on a composite corpus containing PA, banking, and insurance documents. Further details on the training corpus and procedure are given in the next sections.

3.1. The Bureau Corpus

We constructed the pre-training corpus, henceforth the *Bureau Corpus*, by selecting administrative, banking and insurance documents. The corpus consists mostly of administrative acts of several Italian municipalities (65% of the whole corpus) collected from a *Solr* database as a part of the project SEMPLICE.[3] For the insurance and banking domains, documents were collected within the project ABI2LE by domain experts, who provided us with a collection of non-life insurance product information sheets and banking public communications, circulars, and provisions.

All documents were pre-processed by first removing line breaks, typical of PA and insurance texts layout. We then split documents into sentences using our customized version of the Italian spaCy tokenizer. We added a list of exceptions to the tokenization rules of the spaCy model containing acronyms and conventional abbreviations of the legal domain, released by [17] common to the PA domain, and other abbreviations that we gathered from bank and insurance texts. Sentences containing OCR errors, special characters, excessive punctuation, or written in foreign languages[4] were filtered out. In addition, we removed the whole document when sentences containing

[3] SEMantic instruments for PubLIc administrators and CitizEns: www.semplicepa.it

[4] https://github.com/saffsd/langid.py

Table 1
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%

errors were more than 40% of the sentences in the document. The final Bureau Corpus contains 35,293,226 sentences and approximately 1B tokens, for a total amount of 6.7 GB of plain text. Details about the Bureau Corpus composition are given in Table 1.

3.2. Domain-adaptive pre-training

Vocabulary expansion To allow the model to better capture the domain lexicons, we expanded the vocabulary of BureauBERTo with new domain-specific tokens. We extracted from the Bureau Corpus 8,305 representative words by applying the TF-IDF to the whole corpus. These terms were added to the original 32,000 tokens UmBERTo vocabulary, thus resulting in a domain-specific tokenizer with 40,305 tokens and an expansion of the model size from 110M to 117M parameters.

Model input format Following the "full sentences" approach in [11], we constructed the input dataset by applying the BureauBERTo tokenizer to contiguous sentences from one or more documents, using the separating special token after each sentence. Additionally, we shuffled the documents to alternate texts pertaining to the three sub-domains in the Bureau Corpus, and avoid effects akin to "catastrophic forgetting" [24].

Pre-training details The model was trained for 40 epochs, resulting in 17,400 steps with a batch size of ~8K[5] on a NVIDIA A100 GPU. We used a learning rate of 5e-5 along with an Adam optimizer (β1=0.9, β2 = 0.98) with weight decay of 0.1 and a 0.06 warm up steps ratio.

3.3. Lexical overlap among domains

To address the question of the lexical overlap among PA, insurance, and banking in-domain words, we computed the percentage of the 8,305 tokens extracted via TF-IDF from the Bureau Corpus, which belong to the three sub-domains. This analysis shows that 21.6% of the tokens are exclusive of the PA domain, while only 3.6% and 0.3% recur solely in banking and insurance documents, respectively. However, even though most of the domain words derive from the PA language (92%), approximately half of them also occur (at least ten times) in insurance (50.7%) and banking (52.3%) documents (see Figure 1).[6]

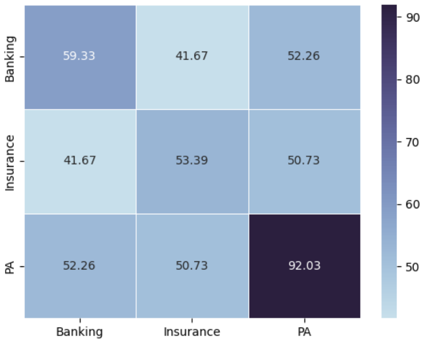

Figure 1: Percentage of words added to BureauBERTo vocabulary occurring at least 10 times in the sub-domains of the Bureau Corpus: PA, Insurance, and Banking.

Hence, the three domains share a rather significant portion of the lexicon, considering that this analysis does not take into account the common Italian vocabulary.

4. BureauBERTo Evaluation

We assessed the effectiveness of our domain adaptation with an intrinsic evaluation measuring the model accuracy in predicting top-k (where $k \in K = \{1, 3, 5, 10\}$) candidates for random and in-domain masked words (Section 4.1). We did not use Pseudo Log-Likelihood (PLL) as proposed in [26], because a meaningful comparison would require the use of the same tokenizer (i.e., the same vocabulary) for both models [27] that would prevent the adapted model from using the new terms added to its vocabulary.

We also performed an extrinsic evaluation by fine-tuning the model on a PA-specialized NER task. In both cases, we compared the performance of BureauBERTo with that of UmBERTo (Section 4.2).

4.1. Fill-mask evaluation

Datasets We evaluated BureauBERTo on the fill-mask task in each sub-domain in the Bureau Corpus. As for the PA one, we selected the ATTO corpus [8], a collection of 11,019 short administrative texts covering different PA

[5] Following [25], we used the "gradient accumulation" technique to have a batch size not bound by the size of GPU memory.

[6] See Appendix A for the vocabulary overlap between all the sub-domain corpora used for adaptation and fill-mask evaluation.

topics (e.g., Environment, Construction, Urban Planning, Education, etc.). For the banking domain we selected a group of 1,262 documents from the dump received by domain experts within the project ABI2LE. These documents are similar to those included in the Bureau Corpus (e.g., public circulars, communications, etc). For the insurance domain, we tested the model on a sample of 319 information sheets concerning life insurance products.

Experimental settings In the first fill-masking evaluation, we used a pre-tokenizer[7] that split the input into whole words according to white spaces and punctuation. We randomly masked one word per sentence, choosing words composed of at least two alphabetic characters. We only scored sentences with more than five and at most 100 of such words. In the second evaluation, we masked only domain-specific words, chosen from three manually created lists of about 100-200 terms related to the three technical-bureaucratic domains. All sentences that contained at least one of these words were selected to be scored for this task. When a sentence included more than one of those domain-specific terms, the word to mask was chosen randomly.

Results and discussion Table 2 shows that 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, and this suggests that the domain adaptation has been effective. The largest gains of BureauBERTo over UmBERTo are on the ATTO corpus belonging to the PA domain (+15,3% for the top-1 candidates of in-domain masked word) and on the insurance dataset (+18,9%).[8] Furthermore, the high performance in the insurance domain, despite the fact that only about 12% of the training documents came from this domain, suggests that transfer learning took place from the PA domain, which covers the largest portion of the training corpus.

4.2. PA specialized NER

Datasets We fine-tuned BureauBERTo on the PA corpus in [28], which contains 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). Following [28] and [8], we also evaluated the model on 25 documents from different municipalities to test the model behavior in dealing with different ways of indicating entities and different writing styles.

Table 2
UmBERTo (UmB.) and BureauBERTo (BB) results in the fill-mask task. The percentages refer to how many times the masked word is predicted within the first k candidates. On the left, the results when a random word (Random) is masked; on the right, when an in-domain term belonging to the vocabulary of both models (In-dom+in-voc.) is masked.

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%

Experimental settings We fine-tuned BureauBERTo using the same PA corpus train, validation, and test split as [8], to make our results comparable. We, therefore, employed as baseline the results obtained on the same datasets by UmBERTo [8] and by INFORMed PA, a PA-specialized model implemented by [28] based on the Stanford NER with a CRF as learning algorithm. BureauBERTo was fine-tuned for 5 epochs with a learning rate of 2e-5 and a batch size of 4. Sentences were tokenized and then truncated at 512 tokens. The training was executed on a NVIDIA A100 GPU.

Results and discussion Table 3 shows that transformer-based models always perform better than INFORMed PA. BureauBERTo and UmBERTo achieved similar overall results, but BureauBERTo obtained a significant improvement on the in-domain class ORG_{PA} (+4%).[9] Interestingly, this class has the highest formal variability [28]. As for ACT (+0.4%) and LAW (+0.9), the two models reached almost the same scores. A different scenario is pictured in Table 4. Here the models are tested over 25 documents published by 25 municipalities, and INFORMed PA reached higher overall results followed by BureauBERTo. These documents differ more in the way entities (LAW and ACT in particular) are expressed. We can hypothesize that shallow features like word shape and n-grams, have helped the model. However, it is important to notice that BureauBERTo performed better than both

[7] https://huggingface.co/docs/tokenizers/api/pre-tokenizers

[8] The Appendix B includes a few samples of masked sentences for a qualitative comparison of the predictions of the two models.

[9] The reported results in this section refer to F1 score.

Table 3
Performance comparison of UmBERTo, INFORMed PA, and BureauBERTo on the PA corpus.

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**

Table 4
Performance comparison of UmBERTo, INFORMed PA, and BureauBERTo on 25 documents dataset.

Model	Measure	ACT	LAW	LOC	ORG	ORG$_{PA}$	PER	MicAvg	MacAvg
UmBERTo	P	0.877	0.836	0.665	0.579	0.538	0.911	0.796	0.792
	R	0.906	0.936	0.770	0.760	0.677	0.918	0.870	0.859
	F1	0.890	0.883	0.714	0.657	0.600	0.915	0.831	0.822
INFORMed PA	P	0.975	0.949	0.799	0.802	0.871	0.914	0.914	0.885
	R	0.848	0.962	0.691	0.769	0.796	0.869	0.836	0.822
	F1	**0.907**	**0.955**	0.741	**0.785**	**0.832**	0.891	**0.873**	**0.852**
BureauBERTo	P	0.854	0.834	0.738	0.489	0.618	0.928	0.798	0.788
	R	0.918	0.923	0.799	0.752	0.871	0.951	0.899	0.889
	F1	0.884	0.876	**0.767**	0.593	0.723	**0.939**	0.846	0.832

INFORMed PA and BureauBERTo for LOC (+2.6%) and PER (+4.8%). This suggests that the domain adaptation did not provoke forgetting, since the adapted model is still able to generalize in recognizing general-purpose entities. To conclude, we assessed the benefits of domain adaptation of UmBERTo only in the PA sub-domain. Nevertheless, we expect to observe an additional improvement in its other sub-domains. Moreover, we expect a further improvement in the results in more complex tasks, possibly inspired by real-world scenarios, where it is even more evident the advantage offered by the additional vocabulary entries.

5. Conclusions and future work

In this paper we presented BureauBERTo, the first transformer-based model adapted to the Italian bureaucratic language. BureauBERTo was created by further pre-training UmBERTo on documents belonging to the PA, insurance, and banking domains. Coming back to our initial questions, we showed that: (i) *What is the overlap among the vocabularies of our target technical-bureaucratic domains?* the three domains share a significant portion of their lexicon; (ii) *To what extent the vocabulary expansion is beneficial for the domain-adaptation of BureauBERTo? Did further pre-training affect the semantic representation of words?* BureauBERTo benefited from the vocabulary extension since it performed better than UmBERTo in the fill-mask task. Moreover, it benefited from domain adaptation, which is evident by observing the higher performances in fill-masking for already-known terms; (iii) *What are the advantages of employing a domain-specific vs. a generic model in a downstream task?* In a PA-specialized NER task, BureauBERTo shows a gain in performance after the domain adaptation.

In future work, we plan to assess the benefits of domain adaptation in the other sub-domains and in other downstream tasks, specifically tailored to the examined technical-bureaucratic domains. To evaluate the performance of BureauBERTo in real-world scenario, we aim at exploiting it in tasks required for implementing the NLP tools provided by the ABI2LE project. Furthermore, we would like to test the model in tasks where general-purpose Italian transformer models were applied to bureaucratic texts, such as in readability [22] and in sentence classification [23], to compare the results achieved before and after the domain adaptation performed in BureauBERTo. Finally, we plan to 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.

Acknowledgments

This research has been funded by the Project "ABI2LE (Ability to Learning)", funded by Regione Toscana (POR Fesr 2014-2020).

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems (2017).

[2] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics (2020).

[3] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, M. McDermott, Publicly available clinical bert embeddings, in: Proceedings of the 2nd Clinical NLP Workshop, 2019.

[4] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, in: Proceedings of the 2019 Conference on EMNLP-IJCNLP, 2019.

[5] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, Legal-bert: The muppets straight out of law school, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 2898–2904.

[6] Y. Yang, M. C. S. Uy, A. Huang, Finbert: A pretrained language model for financial communications, arXiv preprint arXiv:2006.08097 (2020).

[7] M. A. Cortelazzo, Il linguaggio amministrativo. principi e pratiche di modernizzazione, 2021.

[8] S. Auriemma, M. Miliani, A. Bondielli, L. C. Passaro, A. Lenci, Evaluating pre-trained transformers on italian administrative texts (2022).

[9] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N. A. Smith, Don't stop pretraining: Adapt language models to domains and tasks, in: Proceedings of the 58th Annual Meeting of the ACL, 2020.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the NACL), 2019.

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[12] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).

[13] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, Advances in neural information processing systems (2019).

[14] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, O. Levy, Spanbert: Improving pre-training by representing and predicting spans, Transactions of the Association for Computational Linguistics (2020).

[15] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. Bowman, Glue: A multi-task benchmark and analysis platform for natural language understanding, in: Proceedings of the 2018 EMNLP Workshop BlackboxNLP, 2018.

[16] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100,000+ questions for machine comprehension of text, in: Proceedings of the 2016 Conference on EMNLP, 2016.

[17] D. Licari, G. Comandè, Italian-legal-bert: A pre-trained transformer language model for italian law (2022).

[18] S. Douka, H. Abdine, M. Vazirgiannis, R. El Hamdani, D. R. Amariles, Juribert: A masked-language model adaptation for french legal text, in: Proceedings of the NLLP Workshop 2021, 2021.

[19] A. Chriqui, I. Yahav, I. Bar-Siman-Tov, Legal hebert: A bert-based nlp model for hebrew legal, judicial and legislative texts, Judicial and Legislative Texts (June 27, 2022) (2022).

[20] A. Tagarelli, A. Simeri, Unsupervised law article mining based on deep pre-trained language representation models with application to the italian civil code, Artificial Intelligence and Law (2021).

[21] S. Lubello, Il linguaggio burocratico, Carocci, 2014.

[22] M. Miliani, S. Auriemma, F. Alva-Manchego, A. Lenci, Neural readability pairwise ranking for sentences in italian administrative language, in: Proceedings of the 2022 Conference of AACL-IJCNLP, 2022.

[23] M. Locatelli, L. C. Tagliabue, G. M. Di Giuda, et al., Archiberto: a hierarchization quality objectives nlp tool in the italian architecture, engineering and construction sector, in: Proceedings of 1st Workshop AIxPA (co-located with AIxIA 2022), 2022.

[24] A. Robins, Catastrophic forgetting, rehearsal and pseudorehearsal, Connection Science (1995).

[25] L. Gao, Y. Zhang, J. Han, J. Callan, Scaling deep contrastive learning batch size under memory limited setup, in: Proceedings of the 6th the Workshop RepL4NLP, 2021.

[26] J. Salazar, D. Liang, T. Q. Nguyen, K. Kirchhoff, Masked language model scoring, in: Proceedings of the 58th Annual Meeting of ACL, 2020.

[27] C. Huyen, Evaluation metrics for language modeling, The Gradient (2019).

[28] L. C. Passaro, A. Lenci, A. Gabbolini, Informed pa: A ner for the italian public administration domain, in: CLiC-it, 2017.

A. Vocabulary overlap

Figure A shows a heatmap representing the vocabulary overlap between sub-domain corpora used for BureauBERTo adaptation and fill-mask evaluation. This overlap is calculated over the 10k most frequent words of each corpus (excluding stopwords), as in Gururangan et al. [9].

We note that the lowest vocabulary overlap (22%) is between the evaluation data for the PA domain (i.e., the ATTO corpus), and the banking data used for the domain adaptation, with a value similar to the one reported by [9] between the Biomedical and Computer Science domains (21%). As expected, the highest overlaps (around 50%) are between the two PA corpora and between the two insurance corpora, which differ with respect to the described insurance products: data about non-life insurance products were used for the adaptation, whereas data on life insurance products for the evaluation. Surprisingly, we find a relatively low overlap (38%) between the two datasets related to the banking domain (i.e., the one used for the domain adaptation and the one used for the evaluation), and this might indicate a high variability among the documents in this domain.

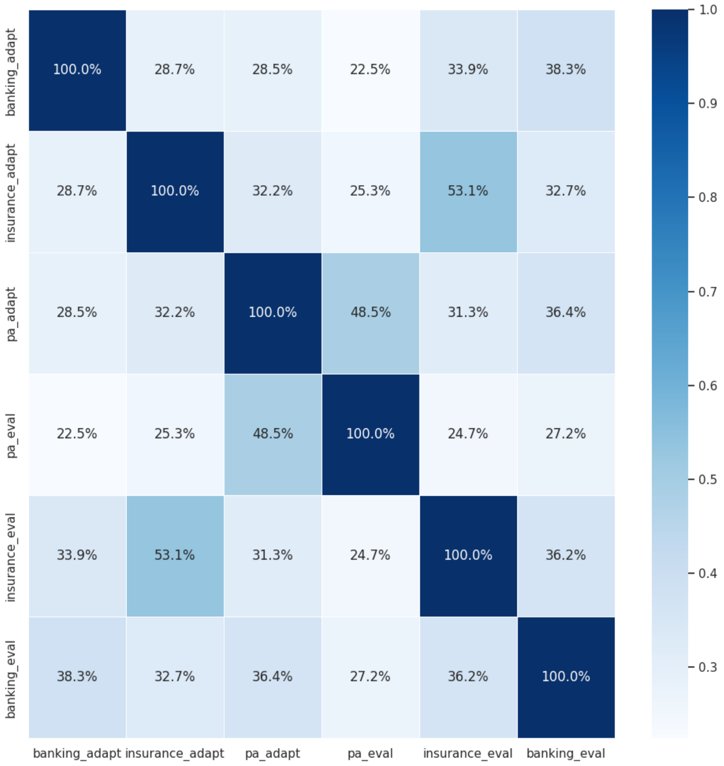

Figure 2: The heatmap shows the vocabulary overlap between corpora. This overlap is calculated over the 10k most frequent words of each corpora. The analyzed datasets were used for model adaptation and fill-mask evaluation for different sub-domains.

B. Examples of top-k candiates

We decided to report some examples of the results returned by UmBERTo and BureauBERTo in the fill-mask task. In particular, we wanted to observe more closely how the domain adaptation would affect the semantic knowledge related to in-domain words already present in UmBERTo vocabulary. We show five sentences where a word was masked and, for each of them, five candidates provided by both models. In Table 5 results obtained on the ATTO corpus, which belongs to the administrative domain, are reported. Table 6 shows the results achieved in the banking domain. Finally, Table 7 shows results regarding the insurance domain.

Table 5
Examples of candidates returned by UmBERTo and BureauBERTo in the fill-mask task, where in-domain terms belonging to the vocabulary of both models where masked. The data (we chose the ATTO corpus) and the terms belong here to the administrative domain.

Sentence	*k*	UmBERTo	BureauBERTo
regolarità ~~**contabile**~~ ai sensi e per gli effetti dell’art. 3 comma 1 lett. b del D.L. n. 174/2012	1	‘,’ (32.81%)	**‘contabile’** (87.02%)
	2	‘tecnica’ (18.00%)	‘formale’ (4.76%)
	3	‘giuridica’ (5.91%)	‘espositiva’ (2.82%)
	4	‘scientifica’ (5.44%)	‘temporale’ (1.93%)
	5	‘formale’ (5.34%)	‘complessiva’ (0.73%)
Dato ~~**atto**~~ che è stato esercitato il controllo preventivo di regolarità contabile ai sensi dell’articolo 147 bis del TUEL, si appone...	1	‘Considerato’ (63.99%)	**‘atto’** (99.99%)
	2	‘Dopo’ (16.69%)	‘Atto’ (0.01%)
	3	‘dopo’ (3.24%)	‘altresì’ (0.00%)
	4	‘Visto’ (3.14%)	‘inoltre’ (0.00%)
	5	‘-’ (3.00%)	‘conferma’ (0.00%)
Richiamata la ~~**determina**~~ n. 177 del 12.02.2021 con la quale è stato assunto impegno di	1	‘deliberazione’ (44.69%)	‘determinazione’ (72.66%)
	2	‘determinazione’ (34.51%)	**‘determina’** (16.18%)
	3	‘delibera’ (7.88%)	‘Determinazione’ (10.65%)
	4	**‘determina’** (4.78%)	‘DD’ (0.17%)
	5	‘nota’ (3.46%)	‘propria’ (0.07%)
Per quanto sopra, esprime parere favorevole ai medesimi e sulla relativa ~~**delibera**~~ della Giunta Comune, raccomandando di apportare le necessarie variazioni al Documento Unico di Programmazione (DUP).	1	‘deliberazione’ (44.96%)	‘proposta’ (85.22%)
	2	‘proposta’ (20.05%)	‘relazione’ (5.36%)
	3	**‘delibera’** (14.51%)	‘deliberazione’ (1.38%)
	4	‘relazione’ (7.00%)	‘competenza’ (0.92%)
	5	‘risoluzione’ (1.52%)	‘iniziativa’ (0.79%)

Table 6

Examples of candidates returned by UmBERTo and BureauBERTo in the fill-mask task, where in-domain terms belonging to the vocabulary of both models where masked. The data and the terms belong here to the banking domain.

Sentence	*k*	UmBERTo	BureauBERTo
Alle società di gestione e alle imprese d ~~investimento~~ extracomunitarie tali previsioni si applicano a condizione che...	1	‘gestione’ (24.25%)	**‘investimento’** (68.31%)
	2	‘capitali’ (23.99%)	‘assicurazione’ (21.54%)
	3	‘partecipazioni’ (12.69%)	‘gestione’ (3.35%)
	4	**‘investimento’** (8.22%)	‘assicurazioni’ (2.04%)
	5	‘capitale’ (6.63%)	‘servizi’ (1.41%)
...Si ipotizzi che l’intermediario C (intermediario standardizzato) abbia ~~erogato~~ nel mese di agosto dell’Anno T-2 un mutuo per un importo di...	1	‘stipulato’ (33.94%)	‘stipulato’ (47.81%)
	2	‘acceso’ (15.10%)	**‘erogato’** (16.55%)
	3	‘sottoscritto’ (11.53%)	‘contratto’ (8.44%)
	4	‘concesso’ (6.70%)	‘sottoscritto’ (6.69%)
	5	‘contratto’ (6.48%)	‘concesso’ (6.55%)
030) Sottogruppo: ~~Banca~~ d’Italia (cod.300);	1	‘Fratelli’ (40.45%)	**‘Banca’** (99.76%)
	2	**‘Banca’** (32.36%)	‘Leggi’ (0.05%)
	3	‘Unità’ (6.16%)	‘Consiglio’ (0.03%)
	4	‘Consiglio’ (2.17%)	‘banca’ (0.02%)
	5	‘Regno’ (1.30%)	‘Banco’ (0.01%)
Si tratta di misure che impongono, ad esempio, restrizioni sulla durata massima dei ~~finanziamenti~~ o limiti al piano di ammortamento...	1	**‘finanziamenti’** (29.26%)	‘prestiti’ (77.96%)
	2	‘prestiti’ (26.56%)	‘mutui’ (11.02%)
	3	‘mutui’ (20.70%)	**‘finanziamenti’** (7.70%)
	4	‘contratti’ (11.46%)	‘debiti’ (1.06%)
	5	‘pagamenti’ (2.03%)	‘titoli’ (0.84%)

Table 7

Examples of candidates returned by UmBERTo and BureauBERTo in the fill-mask task, where in-domain terms belonging to the vocabulary of both models where masked. The data and the terms belong here to the insurance domain, concerning life insurance products.

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%)
...la Copertura Assicurativa viene rideterminata e la Compagnia restituisce all’Impresa la parte di ~~premio~~ pagato relativa alla Copertura non più operante.	1	‘importo’ (28.82%)	‘Premio’ (64.00%)
	2	‘quanto’ (14.14%)	**‘premio’** (35.95%)
	3	‘capitale’ (14.00%)	‘importo’ (0.02%)
	4	**‘premio’** (11.50%)	‘rischio’ (0.00%)
	5	‘prezzo’ (6.13%)	‘quanto’ (0.00%)
Gli sviluppi delle prestazioni rivalutate e del valore di ~~riscatto~~ di seguito riportati sono calcolati sulla base	1	‘mercato’ (32.84%)	**‘riscatto’** (92.77%)
	2	‘riferimento’ (14.85%)	‘mercato’ (2.45%)
	3	‘liquidazione’ (5.18%)	‘importo’ (0.02%)
	4	‘rimborso’ (3.28%)	‘liquidazione’ (0.82%)
	5	‘rivalutazione’ (2.68%)	‘default’ (0.22%)
B) Rivalutazione del ~~capitale~~ assicurato La Misura di Rivalutazione, se positiva, viene attribuita, al Capitale Assicurato, a partire dal 1° gennaio	1	**‘capitale’** (58.71%)	‘Capitale’ (62.89%)
	2	‘Capitale’ (40.65%)	**‘capitale’ (36.57%)**
	3	‘patrimonio’ (0.22%)	‘valore’ (0.15%)
	4	‘Patrimonio’ (0.07%)	‘patrimonio’ (0.15%)
	5	‘rischio’ (0.04%)	‘Valore’ (0.10%)