Challenging specialized transformers on zero-shot classification

Serena Auriemma[1], Mauro Madeddu[1], Martina Miliani[1], Alessando Bondielli[1], Alessandro Lenci[1] and Lucia Passaro[3]

[1]Dipartimento di Filologia, Letteratura e Linguistica, Università di Pisa, Via Santa Maria, Pisa, 56126, Italy
[3]Dipartimento di Informatica, Università di Pisa, Largo B. Pontecorvo, 3 Pisa, 56127, Italy

Abstract

This paper investigates the feasibility of employing basic prompting systems for domain-specific language models. The study focuses on bureaucratic language and uses the recently introduced BureauBERTo model for experimentation. The experiments reveal that while further pre-trained models exhibit reduced robustness concerning general knowledge, they display greater adaptability in modeling domain-specific tasks, even under a zero-shot paradigm. This demonstrates the potential of leveraging simple prompting systems in specialized contexts, providing valuable insights both for research and industry.

Keywords

Domain Adaptation, Transformers, Prompting, Zero-shot, Italian Bureaucratic Language, Public Administration

1. Introduction

Pre-trained Language Models (PLMs) have had a significant impact on Natural Language Processing (NLP), and the pre-train and fine-tune paradigm has become the predominant approach for applying effective models on a wide variety of downstream tasks [1, 2, 3, *inter alia*].

However, one of the main concerns when working with PLMs is the paucity of annotated data, especially for specific domains, required to fine-tune the additional classification layer on top of these models for downstream tasks, such as classification. Recently, prompt-based tuning has started to affirm as a promising way to perform similar tasks, significantly reducing the need for annotated data. This approach has been proven to be very effective with Large Language Models (LLMs) [4]. However, it is often the case that LLMs are not available for low-resource languages, and that their performance drastically decreases when they are challenged on specific domains. Hence, we decided to test a domain-specific model, BureauBERTo [5], a LM further pre-trained on Italian bureaucratic texts (e.g., administrative acts, banking and insurance documents), in a zero-shot scenario exploiting the prompt-based tuning technique.

Since BureauBERTo has shown to be particularly accurate in the fill mask task [5],[1] where the model had to predict both random and in-domain masked words, we wanted to further inspect the domain lexical knowledge acquired by this model during the domain adaptation. We aimed at leveraging this knowledge to implement two classification tasks in the PA domain, modeled as prompt-based classification. Thus, we challenged the model to predict both the topics of PA texts, and the type of generic and PA-related named entities occurring in sentences extracted from administrative documents.

We conducted two prompting experiments for each task. We first adopted the Italian name of the classification classes as label words, then we associated in-domain terms to each class. We also compared BureauBERTo with an Italian generic PLM, UmBERTo (Section 3).

Our findings show that in a zero-shot classification scenario when the label words of each class are shallowly related to the content of the text or to the entity type fed to the model in the prompt template, both the generic and the domain-specialized models perform poorly in the classification task. However, when the classes are represented by multiple word labels semantically related to the text/entity to be classified, the PLMs improve their performance by a wide margin. This gaining is particularly evident in the domain-adapted model BureauBERTo, which outperformed UmBERTo in both prompt-document classification and prompt-entity typing tasks, suggesting that the domain linguistic knowledge acquired by this model during the additional pre-training phase could be particularly useful in a prompt-based tuning scenario where the model is much more reliant on its word knowledge, compared to when the same task is accomplished via

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
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); 0000-0001-5790-4308 (A. Lenci); 0000-0003-4934-5344 (L. Passaro)

[1]See Appendix B for the plot of the model results in the fill-mask task

fine-tuning.

2. Related work

PLMs have proven to be effective in NLP tasks related to specific domains, whether they were trained from scratch [6, 7], or further pre-trained on domain data [8, 9, 10] with a Masked Language Modeling (MLM). More recently, the MLM training objective has been leveraged to solve various NLP tasks reformulated as a sort of cloze task, allowing the PLM to directly solve it without any or with very few labelled examples. One of the first works in this direction was proposed by [11], who performed zero-shot learning using pre-trained LMs without fine-tuning on a dataset of training examples. Within similar conditions, but using the larger GPT-3, [4] achieved near state-of-the-art results for some SuperGLUE [12] tasks. [13] showed that competitive performance with those of GPT-3 can be achieved with much smaller models like the 220M parameters ALBERT, by performing some gradient-based fine-tuning of the model using the labeled examples on a cloze task. Since then, prompt-based learning has gained attention as a simple way to perform, among other tasks, zero-shot classification [14]. However, it's essential to note that the performance of prompt-based learning techniques scales with model size [15]. Consequently, general purpose Large Language Models (LLMs) with billions of parameters are typically used in prompt-learning experiments, even for specialized domains such as the legal one [16]. In contrast, for the biomedical and clinical domains, [17] showed that smaller specialized models like BioBERT [8] and Clinical BERT [18] outperform GPT-2 and T5 in a few-shot prompt based classification of medical texts. The authors hypothesize that the advantage of the BERT-based models is possibly due both to their domain adaptation and to their bidirectional MLM training objective, which is more similar to the prompt template format than those of auto-regressive and sequence-to-sequence models like GPT-2 and T5. [19] reported a similar finding even for the much larger GPT-3 over BioBERT. Nevertheless, these approaches are constrained by the model input size, which limits the length of the conditioning input context and can significantly affect performance [19].

Although prompt-based classification with specialized models has been explored for the medical and clinical domains, to the best of our knowledge, this is the first work that focuses on applying prompts to the Italian administrative language and in a zero-shot classification scenario. Additionally, a notable challenge in prompt-based approaches lies in their sensitivity to variations in prompt templates and verbalizers [20, 21, 22]. We conducted experiments using different verbalizers, i.e., a generic verbalizer and a custom verbalizer using domain-specific terms, to investigate how domain-related word labels affect the model's performance in different classification tasks.

3. Models

For our experiments, we decided to compare the performance of two PLMs, namely UmBERTo and BureauBERTo. UmBERTo[2] is a RoBERTa-based language model trained on the Italian section of the OSCAR corpus,[3] that has been shown to perform well on administrative data [23] compared to other generic PLMs of the same size (110M parameters). BureauBERTo[4] [5] is a domain-adapted model obtained by further pre-training UmBERTo on Italian PA, banking, and insurance documents.

4. Experimental settings

4.1. Prompting

Prompt-based classification requires a specific template to reformulate the original classification task as a cloze-task, where the text to be classified is fed to the model followed by a prompt sentence, such as "This `<text>` is about [MASK]". In this way, the model has to predict the probability that a certain word is filled in the "[MASK]" token. The mapping from the label candidate word to a specific class is gained through the *verbalizer* [13], which represents the original class names as a set of label words, greatly influencing the model performance in the task [24]. Hence, we decided to conduct our prompt-based classification experiments in two settings, using a standard and a custom *verbalizer* to better understand the correlation between the lexical knowledge of PLMs and the use of domain-related terms as the set of word labels in the prompt *verbalizer*.

The first verbalizer, i.e., the *base-verbalizer* simply uses the Italian name of the classification classes as label words (e.g., *Ambiente - "Environment"* is the label word for the class Ambiente - "Environment"), while the second verbalizer is a *manual verbalizer* that we constructed by adding some synonyms of the class name and some related PA terms as label words for each class, to better depict the document classes and the entity types (in this case the label words for the class Ambiente - "Environment" are: *ambiente - "environment", natura - "nature", territorio - "territory", flora - "flora",etc.*).

[2] https://github.com/musixmatchresearch/umberto
[3] https://oscar-corpus.com
[4] https://huggingface.co/colinglab/BureauBERTo

4.2. Datesets

We evaluate the models in two tasks on two different datasets. For the **prompt document classification**, we used a subset of the ATTO corpus [23], which is a collection of administrative documents annotated with labels denoting topics. We filtered this dataset keeping only those instances (2,811) that were annotated with a single topic label.

For the **prompt entity typing** task, we used the PA-corpus of [25], a collection of 460 PA-documents with token-level annotations of Named Entities denoting both general entities, such as persons, locations, organizations, and domain-specific entities, like legislative norms, acts, and PA-related organizations.

4.3. Evaluation metrics

We evaluated the performance of the models with common classification metrics.

4.4. Prompt entity typing

We modeled the NER task introduced by [25] as an entity typing task. Entity typing can be considered a subtask of NER and focuses on entity classification. In other words, systems assign a label to an already extracted entity. This task is often formulated to challenge systems at retrieving sub-categories organized in a hierarchical structure (e.g., an entity corresponding to a person may be specified as director, major, lawyer, etc.) As in [25], we asked models to identify only coarse-grained entities: generic ones, such as persons (PER), locations (LOC), and organizations (ORG); and related to the administrative domain: law references (LAW), administrative acts (ACT), and PA organizations (OPA).

We prompted the models by giving as input a sentence and an entity occurring in it, asking to predict the entity type in place of a masking token. The resulting template is: `<text>. In questa frase, <entity> è un esempio di <mask>.`[5]

As anticipated, we verbalized the entities in two ways. In the first experiment, we provided an Italian translation of the entity or a single word representing the entity class. In the second experiment, we expanded most of the label words by including synonyms and other terms related to the various classes.[6]

4.5. Prompt document classification

For the recognition of the topics in PA documents, we designed the following template to model the document classification task as a masked language modeling problem: `<text>.Questo documento parla di <mask>.`[7]

Thus, PLMs are challenged to infer the topic of the document by predicting the most appropriate label word to represent the masked token in the prompt, following the document text. Since the ATTO corpus contains only short documents of a maximum of 600 tokens, by setting the tokenizer's truncation at 512 tokens[8], we were able to feed the models the entire document in almost all cases. Like with the prompt entity typing, we perform the prompt-based classification twice. In the first experiment, we used the *base verbalizer*, where each class is linked to one or few label words that correspond to the names of the classes in the original annotation of the ATTO corpus. For the second experiment, we use the *manual verbalizer*, which contains, in addition to the label words of the *base verbalizer*, a collection of domain terms manually selected as PA representative topic labels for each class. The complete list of the label words used in both verbalizers is shown in Table 1.[9]

5. Results and discussion

Table 2 shows the results of prompting applied to the entity typing task.

In the first experiment, where a single class label is used (see Sec. 4.4), UmBERTo almost doubled the results obtained by BureauBERTo for F1 Micro (0.404 vs. 0.263) and Macro Average (0.335 vs. 0.201). Surprisingly, for a domain entity like ACT, BureauBERTo missed all the entities, whereas UmBERTo obtained a low but higher score (0.140). For the LAW entity, UmBERTo overpasses BureauBERTo, as well. We may suppose that this is due to the fact that UmBERTo was trained on Common-Crawl, which also contains legal and administrative texts in its Italian section. Very high results are obtained by UmBERTo for PER entities, reaching 0.827 in our zero-shot scenario. On the contrary, both models obtain very low results for LOC, OPA, and ORG. These two latter classes are very similar to each other: ORG refers to organizations in general, comprising firms and associations, whereas OPA can be considered as a subclass of ORG, and refers to organizations within the Public Administration, such as municipal departments. Such overlapping may impact on classification.

For what concerns the second experiment, we added to the prompt also highly distinctive words for each class. In this case, we notice a better ability of BureauBERTo to recognize domain-specific entities such as ACT, LAW,

[5] In English: `<text>. In this sentence, <entity> is an example of <mask>.`

[6] Both verbalizers for entity typing are in Appendix A.

[7] In English: `<text>. This document is about <mask>.`

[8] 512 is the maximum number of tokens that these Transformers models can receive as input.

[9] See Appendix A for the English translation of the label words for document classification.

Table 1
The Table shows the label words adopted in the experiments of prompt document classification.

Class	Basic Labels	+In-domain Lexicon
Ambiente	ambiente	ambiente, natura, territorio, flora, fauna, animali, clima, inquinamento, rifiuti, igiene, caccia, pesca, verde, ecologia, agricoltura, acque
Avvocatura	avvocatura	avvocatura, avvocati, giustizia, legale, ricorso, giudici, Tribunale, Corte, Appello, Assise, notifica, atti, Albo, Pretorio, protocollo
Bandi-Contratti	bandi, contratti	bandi, contratti, bando, contratto, gara, appalto, assunzione, liquidazione
Commercio-Attività-economiche	commercio, attività, economiche	commercio, economia, attività, economica, beni, commerciare, vendite, acquisti, commercianti, confesercenti
Cultura-Turismo-sport	cultura, turismo, sport	cultura, turismo, sport, culturale, turisti, musei, arte, cinema, vacanze, spettacolo, scuola, manifestazioni
Demografico	demografico	demografico, popolazione, abitanti, residenti, censimento, anagrafe, residenza, domicilio, cittadinanza, leva
Edilizia	edilizia	edilizia, costruzioni, cantiere, ristrutturazione, planimetrie, residenziale
Personale	personale	personale, risorse, umane, assunzioni, lavoro, part-time
Pubblica-Istruzione	istruzione	istruzione, istituto, scolatisco, scuola, insegnante, formazione, educazione
Servizi-Informativi	servizi, informazioni	servizi, informazioni, informativi
Servizio-Finanziario	finanza	finanza, euro, finanziario, contabilità, contabile, copertura, rimborsi, pagamenti, versamenti, bilancio, spese, sanzioni, multe, tributi, retribuzioni, emolumenti
Sociale	sociale	sociale, leva, militare, disabili, protezione, civile, invalidi
Urbanistica	urbanistica	urbanistica, trasporti, trasporto, traffico, circolazione, veicoli, viabilità, viaria

and OPA. However, despite the general improvement in recognizing such classes, we notice that it performs worse than UmBERTo for traditional entities. This experiment based on the comparison of general-purpose language models and domain-adapted ones has yielded compelling insights. Generally, both types of models demonstrate enhanced performance when enriched with domain-specific terms within their prompts. However, it is evident that the domain-adapted model outperforms the general-purpose model, exhibiting an improvement of more than twofold (0.516 vs 0.368 for Macro Average F1 score). This significant boost in performance suggests that the domain-adapted model is likely to be more attuned and proficient in leveraging domain-specific terminology.

Nevertheless, it is important to acknowledge that domain-specific terms may wield less influence over generic entities such as PER. With the in-domain lexicon added to the verbalizer, UmBERTo fails to recognize any PER entity. By looking at the confusion matrix for UmBERTo, we observed that the model identifies almost all the people's names as ORG entities. Thus, we carried out an ablation study by deleting the in-domain terms added for the PER entity class, i.e. *generalità - "particulars"* and *nominativo - "name"*.

The results in Table 3 show that the performance of UmBERTo increases not only for the PER entities but that the ablation improves the F1-score of the ORG class as well. Whereas UmBERTo reaches the highest performances for overall F1 Micro Avg, the deletion of in-domain lexicon from the verbalizer seems to penalize BureauBERTo in the recognition of PER entities. Following the trend observed in UmBERTo, the ablation impacts the model's ability to properly recognize the other classes. Despite this, the adapted model still obtained higher results on the in-domain entity classes: ACT, LAW, and OPA further solidifying the advantages of domain-adapted models in specialized contexts. Finally, it is worth noting that we observed a high variability of results according to different prompts and verbalizer configurations, as shown in the ablation study. In fact, deleting the in-domain lexicon related to one of the entity classes affected the performance achieved by the models on all the others, due to wrong classifications (e.g., people names confused with location addresses or company names). Therefore, future investigations into prompt

tuning are necessary and can lead to further interesting insights.

Regarding the prompt document classification experiments, whose results are summarized in table 4, we observed a similar trend. When only one or few word labels are used to represent a topic class, both the generic and the domain-specialized models obtained a rather low accuracy (0.22 vs. 0.09) and Macro Average F1 scores (0.16 vs. 0.06). In this case, UmBERTo outperformed BureauBERTo in almost all classes, with the exception of Cultura, Turismo e Sport - 'Culture, tourism, and sports', Demografico - 'Demographics', and Personale - 'Personnel'. Looking into the details of the scores obtained by UmBERTo in its most recognizable classes (Pubblica istruzione - 'Public Education', Edilizia - 'Constructions' and Urbanistica - 'Urban planning'), we speculate that the single-word labels used to define these classes provided a sufficient cue to enable the model to appropriately recognize these topics. This is in line with the fact that the UmBERTo pre-training corpus included texts extracted from Italian municipalities' web pages, which often refer to such topics.

On the other hand, in the second experiment, where we manually added to the prompt *verbalizer* a set of salient PA-related terms to depict the document topics at a finer-grained level, we observed a significant improvement in the overall performance of both models. The benefits of a custom-made set of domain-related terms are particularly evident for the specialized model BureuBERTo, which reached a better accuracy (0.60 vs. 0.54) and Weighted Average F1-score (0.57 vs. 0.51) than UmBERTo. It appears that the model adapted to the domain may possess heightened sensitivity, enabling it to effectively capitalize on the contextual cues offered by domain-specific terms. However, by performing a class-wise comparison between the two experimental settings, we observed that for some classes that shared a common domain lexicon, such as Pubblica Istruzione - 'Public Education' and Cultura, Turismo e Sport - 'Culture, Tourism, and Sports', or Servizi finanziari - 'Financial services' and Bandi e Contratti - 'Tenders and Contracts' the models' classification could have been influenced in favor of one of the two classes, due to their topic descriptor lexical overlap. These findings confirm the necessity of further inquiry into the effect of lexical specificity on prompt-based classifications, especially for domain-adapted models.

6. Conclusion and future work

In this paper, we propose a zero-shot prompt tuning classification approach for solving two tasks related to the Italian PA domain: the classification of documents according to their topic and the recognition of the entity types occurring in administrative sentences.

We compared the results obtained in these two tasks by the PA-specialized model BureauBERTo with those of the domain-agnostic model UmBERTo. Our findings show that by enriching with domain terms the set of word labels encoded in the prompt *verbalizer* both models demonstrated enhanced performances. Moreover, BureauBERTo exhibited an improvement over UmBERTo of +0.06 Weighted Average F1 score in the document classification (0.51 vs. 0.57) and of more than twofold in the entity typing task (0.516 vs. 0.368 for Macro Average F1 score), meaning that the domain adapted model is more proficient in leveraging domain-specific terminology.

These results underscore the importance of tailoring language models to specific domains to unlock their full potential and address the nuanced challenges posed by diverse subject matters. However, it is also worth mentioning that we noticed a high variability in the task results according to different prompting and different label words. In particular, when the label words adopted to depict a certain topic class are, within the domain context, semantically related to the label words of another class, the models' classification output seems to be biased in favor of one of the two classes.

In conclusion, our study underscores the critical need for a thorough exploration of prompt engineering, particularly in the context of the entity typing task. This imperative arises not only from the potential to augment the predictive capabilities of models, but also from the need to consolidate the knowledge related to general entity classes. Notably, the Public Administration (PA) domain exhibits distinctive characteristics, both in terms of referencing entity names within documents and employing domain-specific terminology. Notably, the identified patterns within the PA domain deviate from the broader, general-purpose Italian style, indicating the necessity for tailored, domain-specific prompt experimentation.

This investigative effort shed the linguistic intricacies that exert an impact on Transformer model performance. Our findings, as revealed in the ablation study on entity linking, emphasize the pivotal importance of delving into the interplay among different entity classes present in datasets. A nuanced analysis of how these classes interact and potentially overlap is indispensable for honing the model's ability to distinguish between them in a domain-specific context.

To conclude, this leads us to surmise as a future direction for our work a further inspection of how domain-adapted PLMs encode in their embedding the semantics of domain-related terms and how this information relates to their performance in prompt-based tasks.

Table 2
Performance comparison of UmBERTo and BureauBERTo on the entity typing task. We grouped together generic entities (LOC, ORG, PER) and domain-related entities (ACT, LAW, OPA). In the upper part of the table are the results of the first experiment, with a unique word as a label. In the bottom part, we report the results for the second experiment where we used multiple labels for each entity class. In bold the best results for each experiment. The best overall results are underlined.

Model	Measure	LOC	ORG	PER	ACT	LAW	OPA	MicAvg	MacAvg
		Basic Labels							
UmBERTo	P	0.7	0.181	0.836	0.4	0.455	0.818	0.462	0.565
	R	0.045	0.372	0.818	0.085	0.618	0.107	0.36	0.341
	F1	**0.085**	**0.244**	**0.827**	**0.140**	**0.524**	**0.189**	**0.404**	**0.335**
BureauBERTo	P	0.5	0.115	0.774	0	0.294	1	0.323	0.447
	R	0.013	0.223	0.526	0	0.447	0.024	0.221	0.205
	F1	0.025	0.152	0.626	0	0.355	0.047	0.263	0.201
		+In-domain Lexicon							
UmBERTo	P	0.767	0.11	0	0.63	0.6	0.364	0.421	0.412
	R	0.445	0.234	0	0.309	0.756	0.476	0.368	0.370
	F1	**0.563**	0.15	0	0.414	0.669	0.412	0.393	0.368
BureauBERTo	P	0.814	0.178	0.45	0.521	0.727	0.797	0.534	0.581
	R	0.368	0.245	0.555	0.404	0.756	0.607	0.492	0.489
	F1	0.507	**0.206**	**0.497**	**0.455**	**0.741**	**0.689**	**0.512**	**0.516**

Table 3
Ablation study conducted on BureauBERTo and UmBERTo on entity typing task. In bold are the best results for each entity class.

Model	Measure	LOC	ORG	PER	ACT	LAW	OPA	MicAvg	MacAvg
UmBERTo	P	0.792	0.294	0.482	0.634	0.728	0.536	0.569	0.578
	R	0.368	0.266	0.577	0.277	0.805	0.536	0.482	0.471
	F1	**0.502**	**0.279**	**0.525**	0.385	0.764	0.536	**0.522**	0.499
BureauBERTo	P	0.746	0.280	0.385	0.629	0.793	0.750	0.573	0.597
	R	0.303	0.223	0.453	0.415	0.780	0.571	0.456	0.458
	F1	0.431	0.249	0.416	**0.500**	**0.787**	**0.649**	0.508	**0.505**

Acknowledgments

This research has been funded by the Project "ABI2LE (Ability to Learning)", Regione Toscana (POR Fesr 2014-2020); by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the European Commission under the NextGeneration EU programme; and partially supported by: TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[2] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 328–339. URL: https://aclanthology.org/P18-1031. doi:10.18653/v1/P18-1031.

[3] 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 North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19

Table 4
Performance comparison of UmBERTo and BureauBERTo on the document classification task. On the left side of the table are the results of the first experiment, where we employed basic label words. On the right side are the results for the second experiment where we used multiple labels for each class. In bold the best results for each experiment. The best overall results are underlined. C-A-E refers to Commercio-Attività-Economiche, whereas C-T-S stands for Cultura-Turismo-Sport.

	Basic Labels						**+In-domain Lexicon**					
	P		R		F1		P		R		F1	
Class	UmB	BB	UmB	BB	UmB	BB	UmB	BB	UmB	BB	UmB	BB
Ambiente	1.00	0.91	0.04	0.03	**0.08**	0.07	0.79	0.86	0.37	0.38	0.51	**0.52**
Avvocatura	0.00	0.00	0.00	0.00	0.00	0.00	0.50	1.00	0.02	0.03	0.03	**0.07**
Bandi-contratti	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.11	0.36	0.31	0.09	**0.16**
C-A-E	0.00	0.00	0.12	0.00	**0.01**	0.00	0.50	1.00	0.38	0.25	**0.43**	0.40
C-T-S	0.62	0.44	0.07	0.12	0.13	**0.19**	0.61	0.41	0.21	0.40	0.31	**0.40**
Demografico	0.00	0.06	0.00	0.98	0.00	**0.12**	0.73	0.64	0.43	0.28	**0.54**	0.39
Edilizia	0.56	1.00	0.31	0.03	**0.40**	0.06	0.88	0.74	0.11	0.29	0.20	**0.41**
Personale	0.10	0.45	0.51	0.22	0.16	**0.29**	0.38	0.70	0.71	0.45	0.49	**0.55**
Pubblica-istruzione	0.67	0.00	0.49	0.00	**0.57**	0.00	0.25	0.14	0.03	0.07	0.05	**0.09**
Servizi-informativi	0.02	0.00	0.53	0.00	**0.03**	0.00	0.06	0.02	0.47	0.27	**0.10**	0.03
Servizio-finanziario	0.58	0.00	0.19	0.00	**0.28**	0.00	0.89	0.54	0.49	0.90	0.63	**0.67**
Sociale	0.20	0.00	0.03	0.00	**0.06**	0.00	1.00	0.00	0.02	0.00	**0.04**	0.00
Urbanistica	0.45	0.50	0.30	0.00	**0.36**	0.00	0.62	0.83	0.98	0.97	0.76	**0.89**
Accuracy	-	-	.	.	**0.22**	0.09	-	-	-	-	0.54	**0.60**
Macro Avg	0.32	0.26	0.20	0.11	**0.16**	0.06	0.56	0.54	0.35	0.35	0.32	**0.35**
Weighted Avg	0.45	0.37	0.22	0.09	**0.24**	0.05	0.69	0.65	0.54	0.60	0.51	**0.57**

-1423. doi:10.18653/v1/N19-1423.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[5] S. Auriemma, M. Madeddu, M. Miliani, A. Bondielli, L. C. Passaro, A. Lenci, BureauBERTo: adapting UmBERTo to the Italian bureaucratic language, in: F. Falchi, F. Giannotti, A. Monreale, C. Boldrini, S. Rinzivillo, S. Colantonio (Eds.), Proceedings of the Italia Intelligenza Artificiale - Thematic Workshops co-located with the 3rd CINI National Lab AIIS Conference on Artificial Intelligence (Ital IA 2023), volume 3486 of *CEUR Workshop Proceedings*, CEUR-WS.org, Pisa, Italy, 2023, pp. 240–248. URL: https://ceur-ws.org/Vol-3486/42.pdf.

[6] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, arXiv preprint arXiv:1903.10676 (2019).

[7] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Transactions on Computing for Healthcare (HEALTH) 3 (2021) 1–23.

[8] 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 36 (2020) 1234–1240.

[9] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, Legal-bert: The muppets straight out of law school, arXiv preprint arXiv:2010.02559 (2020).

[10] D. Licari, G. Comandè, Italian-legal-bert: A pretrained transformer language model for italian law, in: CEUR Workshop Proceedings (Ed.), The Knowledge Management for Law Workshop (KM4LAW), 2022.

[11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners (2019).

[12] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, S. Bowman, Superglue: A stickier benchmark for general-purpose language understanding systems, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 32, Curran

Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf.

[13] T. Schick, H. Schütze, It's not just size that matters: Small language models are also few-shot learners, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 2339–2352. URL: https://aclanthology.org/2021.naacl-main.185. doi:10.18653/v1/2021.naacl-main.185.

[14] R. Puri, B. Catanzaro, Zero-shot text classification with generative language models, Computing Research Repository (CoRR) abs/1912.10165 (2019). URL: http://arxiv.org/abs/1912.10165. arXiv:1912.10165.

[15] B. Lester, R. Al-Rfou, N. Constant, The power of scale for parameter-efficient prompt tuning, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 3045–3059.

[16] F. Yu, L. Quartey, F. Schilder, Legal prompting: Teaching a language model to think like a lawyer, arXiv preprint arXiv:2212.01326 (2022).

[17] S. Sivarajkumar, Y. Wang, Healthprompt: A zero-shot learning paradigm for clinical natural language processing., in: AMIA... Annual Symposium proceedings. AMIA Symposium, volume 2022, 2022, pp. 972–981.

[18] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, M. McDermott, Publicly available clinical bert embeddings, arXiv preprint arXiv:1904.03323 (2019).

[19] M. Moradi, K. Blagec, F. Haberl, M. Samwald, Gpt-3 models are poor few-shot learners in the biomedical domain, arXiv preprint arXiv:2109.02555 (2021).

[20] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2021.

[21] Y. Lu, M. Bartolo, A. Moore, S. Riedel, P. Stenetorp, Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 8086–8098.

[22] D. Trautmann, A. Petrova, F. Schilder, Legal prompt engineering for multilingual legal judgement prediction, arXiv preprint arXiv:2212.02199 (2022).

[23] S. Auriemma, M. Miliani, A. Bondielli, L. C. Passaro, A. Lenci, Evaluating pre-trained transformers on italian administrative texts, in: Proceedings of 1st Workshop AIxPA (co-located with AIxIA 2022), 2022.

[24] T. Gao, A. Fisch, D. Chen, Making pre-trained language models better few-shot learners, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 3816–3830. URL: https://aclanthology.org/2021.acl-long.295. doi:10.18653/v1/2021.acl-long.295.

[25] L. C. Passaro, A. Lenci, A. Gabbolini, Informed PA: A NER for the italian public administration domain, in: R. B. andMalvina Nissim, G. Satta (Eds.), Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), Rome, Italy, December 11-13, 2017, volume 2006 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2017. URL: https://ceur-ws.org/Vol-2006/paper048.pdf.

A. Label Words

Table 5 shows the verbalizer for entity typing. Table 6 contains the English version of the verbalizer adopted for the document classification (see Table 1 for the Italian version).

Table 5
The Table shows the label words adopted in the experiments of prompt entity typing.

Class	Basic Labels	+In-domain Lexicon
PER	persona	persona (person), generalità (particulars), nominativo (name)
LOC	luogo	luogo (place), località (locality)
ORG	organizzazione	organizzazione (organization), azienda (firm), società (corporation), associazione (association), compagnia (company)
LAW	legge	legge (law), norma (rule), decreto (decree), legislativo (legislative)
ACT	atto	atto (act), delibera (resolution), determina (decision), deliberazione (deliberation), regolamento (regulation)
OPA	ufficio	ufficio (office)

Table 6
The Table shows the label adopted in two experiments related to Document Classification. This is an English translation of Table 1. Although some Italian words are translated as multi words word labels can be represented as single words only.

Class	Basic Labels	+In-domain Lexicon
Environment	environment	environment, nature, land, flora, fauna, animals, climate, pollution, waste, hygiene, hunting, fishing, green, ecology, agriculture, water
Advocacy	advocacy	advocacy, attorneys, justice, legal, appeal, judges, courthouse, court, appello, assise, notification, acts, albo, pretorio, protocol
Tenders-Contracts	tenders, contracts	tenders, contracts, notice, contract, tender, hiring, liquidation
Trade-Economic-Activities	trade, economic, activities	trade, economy, business, economic, goods, trade, sales, purchases, merchants, confesercenti
Culture-Turism-Sport	culture, turism, sport	culture, tourism, sports, cultural, tourists, museums, art, cinema, vacations, entertainment, school, events
Demographic	demographic	demographics, population, inhabitants, residents, census, registry, residence, domicile, citizenship, conscription
Building	building	building, construction, yard, renovation, planimetry, residential
Personnel	personnel	personnel, resources, human, hiring, work, part-time
Education	education	education, institute, school, teacher, training, education
Information-Services	services, information	services, information, informative
Financial-Services	finance	finance, euro, financial, accounting, accountant, coverage, refunds, payments, disbursements, budget, expenses, penalties, fines, taxes, wages, emoluments
Welfare	welfare	welfare, conscription, military, disabled, protection, civilian, disability
Urban-Planning	urban planning	urban planning, transportation, transports, traffic, circulation, vehicles, roadway

B. Fill-mask results

Preliminary experiments on a fill-mask task (Fig.1) showed that BureauBERTo outperformed UmBERTo when predicting masked words on Public Administration documents [5]. This motivated us to evaluate BureauBERTo domain-specific knowledge in an unsupervised setting in prompt-based zero-shot classification tasks.

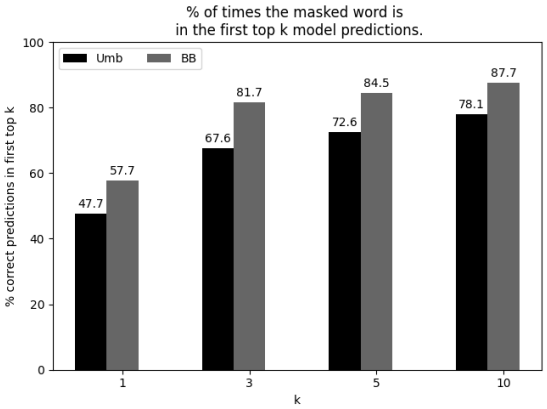

Figure 1: Results of a fill-mask task experiment in which [5] masked domain-specific words in sentences from the ATTO corpus (PA domain). Percentages indicate the number of times the masked word was in the model's top k predictions.