NERMuD at EVALITA 2023: Overview of the Named-Entities Recognition on Multi-Domain Documents Task

Alessio Palmero Aprosio[1,*,†], Teresa Paccosi[1,2,†]

[1]Fondazione Bruno Kessler, Via Sommarive 18, I-38121 Trento, Italy
[2]Dipartimento di Psicologia e Scienze Cognitive, Università di Trento, Corso Bettini 84, I-38068 Rovereto (TN), Italy

Abstract

In this paper, we describe NERMuD, a Named-Entities Recognition (NER) shared task presented at the EVALITA 2023 evaluation campaign. NERMuD is organized into two different sub-tasks: a domain-agnostic classification and a domain-specific one. We display the evaluation of the system presented by the only task participant, ExtremITA. ExtremITA proposes a unified approach for all the tasks of EVALITA 2023, and it addresses in our case only the domain-agnostic sub-task. We present an updated version of KIND, the dataset distributed for the training of the system. We then provide the baselines proposed, the results of the evaluation, and a brief discussion.

Keywords

Shared Task, Named-Entity Recognition, NERMuD 2023, EVALITA 2023

1. Introduction and Motivation

Named-entity recognition (NER) is one of the most common and important task in the field of Natural Language Processing (NLP). It involves identifying and classifying mentions of entities in texts and it is widely used in applications such as text understanding [1], information retrieval [2], knowledge base construction [3], and the protection of personal data [4]. These entities can belong to a set of predefined categories, with people, locations, and organizations being the most common ones.

Manually annotated data play a crucial role in training and evaluating NER systems, similar to other NLP tasks. Systems trained on datasets from specific domains often do not perform well when applied to different types of texts [5].

NER has been addressed in almost all languages, indicating a significant interest in the topic [6]. It is an important task in its own right, as it can be used to process large archival collections. While NER is considered a solved task, some studies have shown that there is always room for improvement depending on factors such as labels, languages, and topics [7]. It is worth noting that, despite the great number of studies on this topic, datasets and tasks for NER often focus on news and, more recently, social media, as seen in initiatives like I-CAB ([8]), NEEL-IT 2016 [9] and NER 2011 [10].

The rest of this article is structured as follows. Section 2 describes the task, and Section 3 gives an overview of the dataset provided. In Section 4 we portray the baseline and the evaluation metric, while in Section 5 we describe the work of the participant ExtremITA. In the end, Section 6 contains a brief discussion, while in Section 7 we draw some conclusions.

2. Task description

In this Section, we describe NERMuD, a task presented at EVALITA 2023 [11] that involves the extraction and classification of named entities – including persons, organizations, and locations – from documents in various domains.

NERMuD 2023 includes two different sub-tasks:

- **Domain-agnostic classification (DAC).** Participants are required to select and classify entities into three categories (person, organization, location) from different types of texts (news, fiction, political speeches) using a single general model.
- **Domain-specific classification (DSC).** Participants are required to make use of a different model for each of the above types, trying to increase the accuracy of every considered type.

The two classification tasks can be addressed in several ways. For example, using deep learning techniques or by adding external data such as gazetteers.

Participants are required to submit their runs and to provide a technical report that should include a brief description of their approach, focusing on the adopted algorithms, models, and resources, a summary of their experiments, and an analysis of the obtained results.

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
*Corresponding author.
†These authors contributed equally.
aprosio@fbk.eu (A. Palmero Aprosio); tpaccosi@fbk.eu (T. Paccosi)
0000-0002-1484-0882 (A. Palmero Aprosio); 0009-0009-2348-7556 (T. Paccosi)

Each participant can submit up to 3 runs for each subtask.

The runs should be contained in a TSV file with fields delimited by a tab and it should follow the same format of the training dataset. No missing data are allowed: a label should be predicted for each token in the test set.

3. Available dataset

The corpus that can be used for training is the Kessler Italian Named-entities Dataset (KIND) [12], presented in 2021 at the Language Resources and Evaluation Conference (LREC). KIND is available and freely downloadable on Github.[1]

The original dataset comprises over one million tokens and includes annotations for three entity classes: person, location, and organization. The majority of the dataset, approximately 600K tokens, features manual gold annotations across three distinct domains: news, literature, and political discourses. This specific subset can be used as the training data for the NERMuD 2023 task, which focuses on Named Entity Recognition and Multi-domain Classification.

All the texts used for the annotation are publicly available, under a license that allows both research and commercial use. In particular, the texts used for the NERMuD task come from:

- Wikinews (WN) as a source providing news texts from the last few decades;
- Some Italian fiction books (FIC) in the public domain, freely accessible for use;
- Writings and speeches from the Italian politician Alcide De Gasperi (ADG), a collection of texts including the works and speeches of Alcide De Gasperi, the Italian politician.

Since the dataset is already publicly released and available, a new set of data has been annotated and shared using the same guidelines (available on the KIND repository on Github).

The dataset has been collected in full compliance with ethical standards, ensuring that it aligns with the terms of use of the sources and that respects the intellectual property and privacy rights of the original authors of the texts.

Table 1 displays an overview of the dataset.

In the next subsections, we provide a quick description of the domains included in the dataset. For more information about the creation of the dataset, the text processing, and the annotation guidelines please refer to [12].

[1] https://github.com/dhfbk/KIND

3.1. Wikinews (WN)

Wikinews is a multi-language free-content project of collaborative journalism. The Italian chapter contains more than 11,000 news articles,[2] released under the Creative Commons Attribution 2.5 License.[3]

In building the dataset, we randomly choose 1,198 articles evenly distributed in the last 20 years, for a total of 364,816 tokens.

3.2. Literature (FIC)

For the annotation of fiction literature, we have included 86 book chapters from a collection of 11 publicly available Italian-authored books. This annotated dataset comprises a total of 219,638 tokens. While the majority of the selected books are novels, we have also included a mix of epistles and biographies. The plain texts come from the Liber Liber website.[4]

In particular, we select: *Il giorno delle Mésules* (Ettore Castiglioni, 1993, 12,853 tokens), *L'amante di Cesare* (Augusto De Angelis, 1936, 13,464 tokens), *Canne al vento* (Grazia Deledda, 1913, 13,945 tokens), *1861-1911 - Cinquant'anni di vita nazionale ricordati ai fanciulli* (Guido Fabiani, 1911, 10,801 tokens), *Lettere dal carcere* (Antonio Gramsci, 1947, 10,655), *Anarchismo e democrazia* (Errico Malatesta, 1974, 11,557 tokens), *L'amore negato* (Maria Messina, 1928, 31,115 tokens), *La luna e i falò* (Cesare Pavese, 1950, 10,705 tokens), *La coscienza di Zeno* (Italo Svevo, 1923, 56,364 tokens), *Le cose più grandi di lui* (Luciano Zuccoli, 1922, 20,989 tokens), *L'occhio del lago* (Tullio Giordana, 1899, 27,190 tokens).

We prioritized selecting texts in the public domain that are as recent as possible (considering that, under the current legislation, copyright expires 70 years after the death of the author). This choice was made to ensure that the model trained on this data would be well-suited for applying to novels written in recent years. By focusing on more contemporary texts, the language used in these novels is expected to be more similar to the language used in present-day novels. Additionally, for the test data, we specifically chose works by the author Tullio Giordana. His works are then not included in the train or the dev sets, to not have a model possibly biased in terms of style.

3.3. Alcide De Gasperi's Writings (ADG)

Finally, we annotate 173 documents (164,537 tokens) from the corpus described in [13], spanning 50 years of European history. The corpus is composed of a comprehensive collection of Alcide De Gasperi's public documents, 2,762

[2] https://it.wikinews.org/wiki/Speciale:Statistiche
[3] https://creativecommons.org/licenses/by/2.5/
[4] https://www.liberliber.it/

Table 1
Overview of the dataset

Dataset	Tokens	Train				Dev				Test			
		Total	LOC	PER	ORG	Total	LOC	PER	ORG	Total	LOC	PER	ORG
WikiNews	364,816	249,077	6,862	8,928	7,593	59,220	1,711	1,802	1,823	56,519	1,310	2,322	1,992
Fiction	219,638	170,942	733	3,439	182	21,506	463	636	284	27,190	37	443	1
De Gasperi	164,537	123,504	1,046	1,129	2,396	27,128	274	253	533	13,905	107	226	326
Total	748,991	543,523	8,641	13,496	10,171	107,854	2,448	2,691	2,640	97,614	1,454	2,991	2,319

in total, written or transcribed between 1901 and 1954, and it is available for consultation on the *Alcide Digitale* website.[5]

4. Baseline and Evaluation

During the definition of the task, we proposed two baselines: an old-style Conditional Random Field [14], and a plain BERT [15] implementation. These options represent the most effective algorithms that can be implemented without the use of GPUs, as well as the simplest algorithms that can be performed using transformers. Both implementations of the baselines can be found on Github.[6]

The CRF model is based on the classifier available in scikit-learn out-of-the-box. In addition to standard features extracted from the text, including vector information from fastText models [16], we also used a set of gazetteers (list of persons, organizations and locations) collected from the Italian Wikipedia using some of the classes contained in DBpedia [17]: Person, Organization, and Place, respectively.

The BERT NER classification model is inspired by the blog post of Tobias Sterbak,[7] using BertForTokenClassification[8] from Hugging Face.

Final results will be calculated in terms of macro-average F_1. The evaluation script is released in the KIND official Github project.[9]

5. Participants

The task has only one participant, the "ExtremITA" group [18], who participated in all the tasks presented at EVALITA 2023 with two unified multi-task learning approaches.

The purpose of ExtremITA is to investigate how the adoption of a Large Language Model can be taken to its extreme consequences by proposing a single model capable of tackling a wide array of heterogeneous tasks (among them, NERMuD).

The authors tested two different models:

extremIT5 - An Encoder-Decoder model based on IT5 [19] consisting of approximately 110 million parameters. This model is trained by concatenating the name of the task and the input sentence/paragraph in the input texts, each representing an example from a generic EVALITA task. Its purpose is to generate a piece of text that solves the target task. For NERMuD, in particular, the list of expected Named Entities is reported as a sequence of text spans, each associated with the corresponding entity type (in the form [⟨ENTITY_TYPE⟩] ⟨TEXT_SPAN_THAT_EVOKES_ENTITY⟩).

extremITLLaMA - An instruction-tuned Decoder-only model, built upon the LLaMA foundational models [20], with a total of 7k million parameters. The initial model was trained using the LoRA technique [21] on Italian translations of Alpaca [22] instruction data. The adapters are then merged into the original model. A final fine-tuning phase using LLaMA is then performed. For each example from EVALITA, an input text is paired with a manually crafted question that simulates an instruction to be solved, representing the specific task. The natural language instruction used in NERMuD is *"Scrivi le menzioni di entità nel testo, indicandone il tipo: [PER] (persona), [LOC] (luogo), [ORG] (organizzazione)."*[10].

In both cases, NERMuD was transformed into a sequence-to-sequence task from its original token classification format.

6. Discussion

Table 2 displays all the results on the different sets available for test. It is worth noting that the ExtremITA's unified approach is very similar to the baseline.

[5] https://alcidedigitale.fbk.eu/
[6] https://github.com/dhfbk/bert-ner
[7] https://bit.ly/ner-bert
[8] https://bit.ly/BertForTokenClassification
[9] https://github.com/dhfbk/KIND

[10] Write the entities' mentions in the text, indicating their type: [PER] (person), [LOC] (location), and [ORG] (organization)

Table 2
Evaluation of the two baselines and of the two ExtremITA runs.

Algorithm	Test	PER			LOC			ORG			Micro			Macro		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Baseline (CRF)	ADG	0.93	0.85	0.89	0.84	0.88	0.86	0.79	0.57	0.66	0.85	0.72	0.78	0.85	0.77	0.80
Baseline (CRF)	WN	0.95	0.95	0.95	0.82	0.76	0.79	0.83	0.77	0.80	0.88	0.84	0.86	0.87	0.83	0.85
Baseline (CRF)	FIC	0.92	0.69	0.79	0.19	0.54	0.28	–	–	–	0.74	0.67	0.70	0.37	0.41	0.35
Baseline (CRF)	ALL	0.95	0.90	0.92	0.78	0.77	0.77	0.83	0.74	0.78	0.87	0.82	0.84	0.85	0.80	0.83
Baseline (BERT)	ADG	0.93	0.92	0.93	0.80	0.90	0.85	0.83	0.67	0.74	0.86	0.79	0.83	0.85	0.83	0.84
Baseline (BERT)	WN	0.96	0.98	0.97	0.89	0.87	0.88	0.87	0.88	0.87	0.91	0.92	0.91	0.91	0.91	0.91
Baseline (BERT)	FIC	0.95	0.82	0.88	0.51	0.76	0.61	–	–	–	0.90	0.81	0.85	0.49	0.52	0.50
Baseline (BERT)	ALL	0.95	0.95	0.95	0.87	0.87	0.87	0.87	0.85	0.86	0.91	0.90	0.90	0.90	0.89	0.89
ExtremIT5	ADG	0.89	0.87	0.88	0.78	0.85	0.81	0.75	0.58	0.65	0.81	0.72	0.76	0.80	0.77	0.78
ExtremIT5	WN	0.96	0.83	0.89	0.89	0.79	0.84	0.80	0.77	0.79	0.88	0.80	0.84	0.88	0.80	0.84
ExtremIT5	FIC	0.96	0.87	0.91	0.62	0.76	0.68	–	–	–	0.92	0.86	0.89	0.53	0.54	0.53
ExtremIT5	ALL	0.95	0.84	0.89	0.87	0.80	0.83	0.80	0.74	0.77	0.88	0.79	0.84	0.87	0.79	0.83
Extr.ITLLaMA	ADG	0.94	0.87	0.91	0.77	0.87	0.82	0.82	0.56	0.66	0.85	0.72	0.78	0.84	0.77	0.79
Extr.ITLLaMA	WN	0.96	0.96	0.96	0.91	0.84	0.87	0.86	0.83	0.84	0.91	0.88	0.90	0.91	0.87	0.89
Extr.ITLLaMA	FIC	0.97	0.82	0.89	0.81	0.81	0.81	–	–	–	0.96	0.82	0.88	0.59	0.54	0.57
Extr.ITLLaMA	ALL	0.96	0.93	0.95	0.89	0.84	0.86	0.86	0.79	0.82	0.91	0.86	0.89	0.90	0.85	0.88

The evaluation of ORG entities for the fiction domain is missing, as none of the classifiers were able to correctly identify the only ORG entity present in the test set (the work "Borsa" in the sentence "Ha avuto disgrazie alla Borsa"). Overall, the BERT baseline outperforms ExtremITA in most runs, with the exception of LOC extraction in fictional texts, where ExtremITLLaMA performs better. This difference in performance can likely be attributed to the textual data used to train the models.

In general, it is possible to notice that the best ExtremITA run overcomes almost always the classification in terms of precision.

7. Conclusions

In this paper we described the first evaluation task for multi-domain named-entity recognition in Italian texts. The task evaluated the performance of participant systems in terms of extracting entities that refers to persons, organizations, and location. The texts used for the tasks cover three different domains: news, political speeches, fiction.

Unfortunately, the task attracted only one participant, ExtremITA, who however presented an interesting and very innovative multi-task approach, probably the first one dealing with so many different tasks in Italian. Although in general the results of ExtremITA do not overcome the two strong baselines proposed (CRF w/ gazetteers, and BERT), the difference in terms of F_1 is very small, demonstrating a promising future for that kind of approaches.

As an outcome of the task, a new version of the KIND dataset is released, increasing its size with respect to the previous version.

References

[1] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, Q. Liu, ERNIE: Enhanced language representation with informative entities, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1441–1451. URL: https://aclanthology.org/P19-1139. doi:10.18653/v1/P19-1139.

[2] J. Guo, G. Xu, X. Cheng, H. Li, Named entity recognition in query, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09, Association for Computing Machinery, New York, NY, USA, 2009, p. 267–274. URL: https://doi.org/10.1145/1571941.1571989. doi:10.1145/1571941.1571989.

[3] O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, A. Yates, Unsupervised named-entity extraction from the web: An experimental study, Artificial Intelligence 165 (2005) 91–134. URL: https://www.sciencedirect.com/science/article/pii/S0004370205000366. doi:https://doi.org/10.1016/j.artint.2005.03.001.

[4] T. Paccosi, A. Palmero Aprosio, Redit: A tool and

dataset for extraction of personal data in documents of the public administration domain, in: Proceedings of CLiC-it 2021 Italian Conference on Computational Linguistics, 2022.

[5] T. Poibeau, L. Kosseim, Proper Name Extraction from Non-Journalistic Texts, Brill, Leiden, The Netherlands, 2001, pp. 144 – 157. URL: https://brill.com/view/book/edcoll/9789004333901/B9789004333901-s011.xml. doi:https://doi.org/10.1163/9789004333901_011.

[6] A. Goyal, V. Gupta, M. Kumar, Recent named entity recognition and classification techniques: A systematic review, Computer Science Review 29 (2018) 21–43. URL: https://www.sciencedirect.com/science/article/pii/S1574013717302782. doi:https://doi.org/10.1016/j.cosrev.2018.06.001.

[7] M. Marrero, J. Urbano, S. Sánchez-Cuadrado, J. Morato, J. M. Gómez-Berbís, Named entity recognition: Fallacies, challenges and opportunities, Computer Standards Interfaces 35 (2013) 482–489. URL: https://www.sciencedirect.com/science/article/pii/S0920548912001080. doi:https://doi.org/10.1016/j.csi.2012.09.004.

[8] B. Magnini, E. Pianta, C. Girardi, M. Negri, L. Romano, M. Speranza, V. Bartalesi Lenzi, R. Sprugnoli, I-CAB: the Italian content annotation bank, in: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), European Language Resources Association (ELRA), Genoa, Italy, 2006. URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/518_pdf.pdf.

[9] P. Basile, A. Caputo, A. Gentile, G. Rizzo, Overview of the evalita 2016 named entity recognition and linking in italian tweets (neel-it) task, 2016.

[10] V. Bartalesi Lenzi, M. Speranza, R. Sprugnoli, Named entity recognition on transcribed broadcast news at evalita 2011, in: B. Magnini, F. Cutugno, M. Falcone, E. Pianta (Eds.), Evaluation of Natural Language and Speech Tools for Italian, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 86–97.

[11] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, EVALITA 2023: Overview of the 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[12] T. Paccosi, A. Palmero Aprosio, KIND: an Italian multi-domain dataset for named entity recognition, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 501–507. URL: https://aclanthology.org/2022.lrec-1.52.

[13] S. Tonelli, R. Sprugnoli, G. Moretti, Prendo la parola in questo consesso mondiale: A multi-genre 20th century corpus in the political domain, in: CLiC-it, 2019.

[14] J. D. Lafferty, A. McCallum, F. C. N. Pereira, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, in: Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, pp. 282–289. URL: http://dl.acm.org/citation.cfm?id=645530.655813.

[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[16] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146.

[17] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, Dbpedia: A nucleus for a web of open data, in: K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, P. Cudré-Mauroux (Eds.), The Semantic Web, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 722–735.

[18] C. D. Hromei, D. Croce, V. Basile, B. Roberto, ExtremITA@EVALITA2023: Multi-Task Sustainable Scaling to Large Language Models as its Extreme, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[19] G. Sarti, M. Nissim, IT5: Large-scale Text-to-text Pretraining for Italian Language Understanding and Generation, ArXiv preprint 2203.03759 (2022). URL: https://arxiv.org/abs/2203.03759.

[20] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021).

[22] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, https://github.com/tatsu-lab/stanford_alpaca, 2023.