SLIMER-IT: Zero-Shot NER on Italian Language

Andrew Zamai[1,2], Leonardo Rigutini[2], Marco Maggini[1] and Andrea Zugarini[2,*]

[1]*Università degli Studi di Siena, Italy*

[2]*expert.ai, Siena, Italy*

Abstract

Traditional approaches to Named Entity Recognition (NER) frame the task into a BIO sequence labeling problem. Although these systems often excel in the downstream task at hand, they require extensive annotated data and struggle to generalize to out-of-distribution input domains and unseen entity types. On the contrary, Large Language Models (LLMs) have demonstrated strong zero-shot capabilities. While several works address Zero-Shot NER in English, little has been done in other languages. In this paper, we define an evaluation framework for Zero-Shot NER, applying it to the Italian language. Furthermore, we introduce SLIMER-IT, the Italian version of SLIMER, an instruction-tuning approach for zero-shot NER leveraging prompts enriched with definition and guidelines. Comparisons with other state-of-the-art models, demonstrate the superiority of SLIMER-IT on never-seen-before entity tags.

Keywords

Named Entity Recognition, Zero-Shot NER, Large Language Models, Instruction tuning

1. Introduction

Named Entity Recognition (NER) plays a fundamental role in Natural Language Processing (NLP), often being a key component in information extraction pipelines. The task involves identifying and categorizing entities in a given text according to a predefined set of labels. While *person*, *organization*, and *location* are the most common, applications of NER in certain fields may require the identification of domain-specific entities.

Manually annotated data has always been critical for the training of NER systems [1]. Traditional methods tackle NER as a token classification problem, where models are specialized on a narrow domain and a pre-defined labels set [2]. While achieving strong performance for the data distribution they were trained on, they require extensive human annotations relative to the downstream task at hand. Additionally, they lack generalization capabilities when it comes to addressing out-of-distribution input domains and/or unseen labels [1, 3, 4].

On the contrary, Large Language Models (LLMs) have recently demonstrated strong zero-shot capabilities. Models like GPT-3 can tackle NER via In-Context Learning [5, 6], with Instruction-Tuning further improving performance [7, 8, 9]. To this end, several models have been proposed to tackle zero-shot NER [10, 4, 3, 11, 12, 13]. In particular, SLIMER [13] proved to be particularly effective on unseen named entity types, by leveraging definitions and guidelines to steer the model generation.
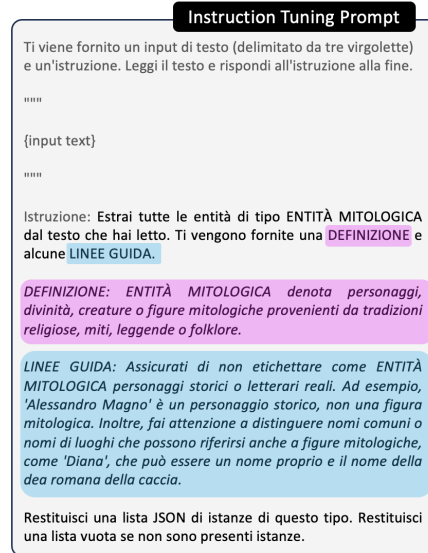

Figure 1: SLIMER-IT instruction tuning prompt. Dedicated entity *definition* and *guidelines* steer the model labelling.

However, little has been done for zero-shot NER in non-English data. More in general, as pointed out in [1], NER is understudied in languages like Italian, especially outside the traditional news domain and *person*, *location*, *organization* classes.

To this end, we propose in this paper an evaluation framework for Zero-Shot NER, and we apply it to the Italian language. In addition, we fine-tune a version of SLIMER for Italian, which we call SLIMER-IT[1]. In the experiments, we explore different LLM backbones and

[1]https://github.com/andrewzamai/SLIMER_IT

we assess the impact of Definition and Guidelines (D&G). When comparing SLIMER-IT with state-of-the-art approaches, either using models pre-trained on English or adapted for Italian, results demonstrate SLIMER-IT superiority in labelling unseen entity tags.

2. Related Work

Several works tackle Zero-Shot NER on English, such as InstructUIE [10], UniNER [4], GoLLIE [3], GLiNER [11], GNER [12] and SLIMER [13]. Most of them are based on the instruction tuning of an LLM and mainly differ in the prompt and output format design. GLiNER distinguishes itself by being a smaller encoder-only model, combined with a span classifier head, that achieves competitive performance at a lower computational cost.

As highlighted in SLIMER [13], most approaches mainly focus on zero-shot NER in Out-Of-Distribution input domains (OOD), since they are typically fine-tuned on an extensive number of entity classes highly or completely overlapping between training and test sets. In view of this, we proposed a lighter instruction-tuning methodology for LLMs, training on data overlapping in lesser degree with the test sets, while steering the model annotation process with a definition and guidelines for the NE category to be annotated. From this, the name SLIMER: Show Less, Instruct More Entity Recognition.

Although the authors of GLiNER propose also a multilingual model and evaluate zero-shot generalizability across different languages, neither they nor any other work has addressed the task of Zero-Shot NER specifically for the Italian language.

NER for Italian. While NER has been extensively studied on English, less has been done in other languages, particularly outside the traditional general-purpose domains and entity labels set [14]. Indeed, in Italian, most NER datasets focus on news and, more recently, social media contents [15, 16, 17]. Currently, there has been no research into zero-shot NER, only a few exploratory studies into multi-domain NER. This challenge was introduced in the NERMuD task (NER Multi-Domain) at EVALITA 2023[2], in which one sub-task required to develop a single model capable of classifying the common entities - *person*, *organization*, *location* - from different types of text, including news, fiction and political speeches. ExtremITA team [18] addressed the challenge proposing the adoption of a single LLM capable of tackling all the different tasks at EVALITA 2023, among which NERMuD. All the tasks were converted into text-to-text problems and two LLMs (LLaMA and T5 based) were instruction-tuned on the union of all the available datasets for the challenge.

3. Zero-Shot NER Framework

In traditional Machine-Learning theory, a model f, trained for a task (e.g. NER) represented by a dataset $\mathcal{X}, \mathcal{Y}$, is typically evaluated on an held-out test set sampled from the same task and distribution of the training. In zero-shot learning instead, a model is expected to go beyond what experienced during training. There are different levels of generalization indicating up to what extent the model goes beyond what directly learnt.

In the case of zero-shot NER, a model should be able to extract entities from inputs belonging to the same domain it was trained on (**in-domain**) and across other domains not encountered before (**out-of-domain**). Moreover, it should also generalize well to novel entity classes (**unseen named entities**). In our zero-shot evaluation framework we aim to measure each level independently. Hence, we define an evaluation benchmark that includes a collection of NER datasets divided by degree of generalization. In the following we describe the required properties to fit in.

In-domain. This evaluation helps measure how well the model can generalize from its training data to similar, but not identical, data. The model is evaluated on the same input-domains and named entities as those in the training set. This data often consists in the test partitions associated with each training set used for fine-tuning the model.

Out-Of-Domain (OOD). OOD evaluation tests the model's ability to generalize to input texts from domains that it has not encountered during training. While the named entities have been seen during training, this type of evaluation is particularly challenging because different input domains often exhibit unique linguistic patterns and domain-specific terminology.

Unseen Named Entities. This evaluation tests the model's ability to identify and classify entities that has not encountered during its training phase. The tag set comprises fine-grained categories which are often specifically defined for the domain in which NER is deployed. Because of this, the input data may often be also Out-Of-Domain (OOD), making this evaluation include the previously mentioned OOD scenario as well.

4. SLIMER-IT

To adapt SLIMER for Italian, we translate the instruction-tuning prompt of [13], as shown in Figure 1. The prompt is designed to extract the occurrences of one entity type per call. While this has the drawback of requiring |NE|

---

inference calls on each input text, it allows the model to better focus on a single NE type at a time.

As in [13], we query gpt-3.5-turbo-1106 via OpenAI's Chat-GPT APIs to automatically generate definition and guidelines for each needed entity tag. The definition for a NE is meant to be a short sentence describing the tag. The guidelines instead provide annotation instructions to align the model's labelling with the desired annotation scheme. Guidelines can be used to prevent the model from labelling certain edge cases or to provide examples of such NE. Such an informative prompt is extremely valuable when dealing with unfamiliar entity tags, and can also be used to distinguish between polysemous categories.

Finally, the model is requested to generate the named entities in a parsable JSON format containing the list of NEs extracted for the given tag.

5. Experiments

Experiments aim to assess our approach in Italian. We study the impact of guidelines and the usage of different backbones. Then, we compare our approach against state-of-the-art alternatives.

5.1. Datasets

We construct the zero-shot NER framework (described in Section 3) for Italian upon NerMuD shared task and Multinerd dataset. In particular, we use NerMuD to build in-domain and OOD evaluation sets, while Multinerd-IT is used to assess the behaviour in the unseen named entites scenario.

NERMuD. NERMuD [1] is a shared task organized at evalita-2023, built based on the Kessler Italian Named-entities Dataset (KIND) [19]. It contains annotations for the three classic NER tags: *person*, *organization* and *location*. Examples are organized in three distinct domains: news, literature and political discourses. Unlike NERMuD, we restrict fine-tuning to a single domain. In such a way, we can evaluate both in-domain and out-of-domain capabilities of the model. In particular, we designate WikiNews (WN) sub-set for training and in-domain evaluation, being the most generic domain, while Fiction (FIC) and Alcide De Gasperi (ADG) splits are kept for out-of-domain evaluation only.

Multinerd-IT. To construct the unseen NEs evaluation set, we exploit Multinerd[3] [20], a multilingual NER dataset made of 15 tags: *person, organization, location, animal, biological entity, celestial body, disease, event, food, instrument, media, plant, mythological entity, time* and

vehicle. We keep the Italian examples only. Such a dataset constitutes a perfect choice to assess models' capabilities on unseen NEs. Indeed, data belongs to the same news domain of the NERMuD split chosen for fine-tuning, but it includes a broader label set. Since we want to measure performance on never-seen-before entities, we exclude entity types seen in training, i.e. *person*, *organization* and *location*. We also remove *biological entity*, being poorly underrepresented, with a support of just 4 instances.

5.2. Backbones

We implemented several version of SLIMER-IT based on different backbone models. We consider similarly sized LLMs, all in the 7B parameters range. In particular, we selected five backbones: Camoscio[4] [21], LLaMA-2-7b-chat [22], Mistral-7B-Instruct [23], LLaMA-3-8B-Instruct, LLaMAntino-3-ANITA-8B-Inst-DPO-ITA[5] [24].

LLaMA-2-7b-chat was originally used in SLIMER [13], and LLaMA-3-8B-Instruct is the newest, improved version of it. As LLaMA family, Mistral-7B-Instruct is a multilingual model mainly English-oriented, but it has demonstrated greater fluency on Italian. Camoscio and LLaMAntino-3-ANITA-8B-Inst-DPO-ITA, instead, are two LLMs specifically fine-tuned on Italian instructions.

5.3. Compared Models

We compare the SLIMER-IT approach, implemented with different backbones, against other state-of-the-art approaches for zero-shot NER. All the methods are trained and evaluated in the defined zero-shot NER framework for a fair comparison. We evaluate against:

Token classification. Although certainly not being suited for zero-shot NER, due to its architectural inability to cope with unseen tags, we decided to evaluate the most known approach to NER as baseline. As in NERMuD [1], we use the training framework *dhfbk/bert-ner*[6]. We fine-tune two different base models, *bert-base-cased*, pretrained on English, and *dbmdz/bert-base-italian-cased*[7], an Italian version.

GNER. It is the best performing approach on zero-shot NER in OOD English benchmark. In GNER [12], they propose a BIO-like generation, replicating in output the same input text, along with a token-by-token BIO label. Here, we consider LLaMAntino-3 as its backbone.

Table 1
Comparing SLIMER-IT based on different backbones, with and without Definition and Guidelines (D&G) in the prompt. LLMs with † symbol were instruction-tuned on Italian. In parentheses the $(\pm\Delta F1)$ of performance given by the usage of D&G.

| Backbone | Params | w/ D&G | In-Domain | OOD | | unseen NEs |
			WN	FIC	ADG	MN
Camoscio †	7B	False	81.80	82.44	79.01	32.28
		True	81.50 (-0.3)	85.08 (+2.64)	76.00 (-3.01)	38.68 (+6.4)
LLaMA-2-chat	7B	False	80.69	80.45	73.81	32.38
		True	83.24 (+2.55)	88.81 **(+8.36)**	79.26 (+5.45)	35.16 (+2.78)
Mistral-Instruct	7B	False	82.71	85.61	75.80	35.63
		True	85.55 **(+2.84)**	**92.78** (+7.17)	80.56 (+4.76)	40.64 (+5.01)
LLaMA-3-Instruct	8B	False	85.93	82.85	80.00	27.62
		True	85.38 (-0.55)	84.38 (+1.53)	78.29 (-1.71)	50.74 (+23.12)
LLaMAntino-3-ANITA †	8B	False	84.12	77.06	74.35	30.90
		True	**85.78** (+1.66)	82.52 (+5.46)	**81.65 (+7.30)**	**54.65 (+23.75)**

GLiNER. Differently from all other methods, GLiNER is based on a smaller encoder-only model, combined with a span classifier head, able to achieve competitive performance on the OOD English benchmark at a lower computational cost. We fine-tune it both using its original *deberta-v3-large* English backbone and the Italian *dbmdz/bert-base-italian-cased* model.

extremITLLaMA. Already described in Section 2, it represents an interesting approach to compare against. Based on Camoscio LLM, we compare it with SLIMER-IT approach implemented with the same backbone.

5.4. Experimental setup

We kept the same training configuration of SLIMER [13] on English, except that we trained on all available samples. Depending on the backbone, the instruction-tuning prompt (see Figure 1) was adjusted accordingly to the structure of its template (e.g. [INST] or <|start_header_id|> formats). For all the competitors, we replicated their training setup using their scripts and suggested hyper-parameters. For the evaluation, we use the micro-F1 as computed in the UniNER[8] implementation.

5.5. Results

Impact of Definition and Guidelines (D&G). We compare SLIMER-IT with a version devoid of definition and guidelines in the prompt. To demonstrate the robustness of the approach, we train several SLIMER-IT instances, based on different LLM backbones. In Table 1, we report the results, highlighting the absolute difference in performance between the model steered by
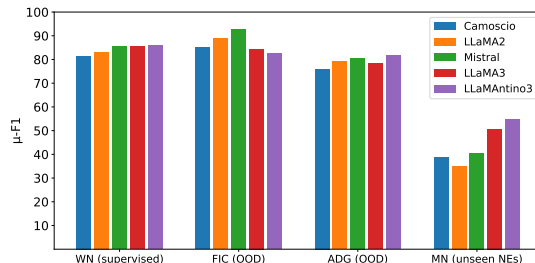

Figure 2: SLIMER-IT performance for different backbones.

Table 2
Comparison with existing off-the-shelf models for zero-shot NER on Italian. We omit in-domain evaluation to not disadvantage them against SLIMER-IT.

| Model | OOD | | unseen NEs |
	FIC	ADG	MN
Universal-NER-ITA	32.4	43.2	12.8 (all seen)
GLiNER-ITA-Large	36.6	42.0	15.5 (all seen)
GLiNER-ML	46.5	49.4	17.4 (all seen)
SLIMER-IT	**82.5**	**81.7**	**54.7**

D&Gs and the one not using them. Generally, definition and guidelines yield improvements in F1. In particular, the gap is contained when evaluating on in-domain data, whereas it becomes significant in OOD and even more substantial in unseen NEs. This is expected since D&G help the most in conditions unseen during training. Notably, LLaMA-3-based backbones benefit the most from definition and guidelines, with improvements beyond 23 absolute F1 points, surpassing all the other models by substantial margins in never-seen-before entity tags.

---

[8]https://github.com/universal-ner

Table 3

Comparing SLIMER-IT with state-of-the-art approaches trained in the same zero-shot setting, and adopting the same backbone when possible. *Note that extremITLLaMA was fine-tuned also on the FIC and ADG train sets for the NERMuD task, so these datasets are not actually OOD for this model.

Approach	Backbone	Language	Params	In-Domain	OOD		unseen NEs
				WN	FIC	ADG	MN
Token classification	BERT-base	EN	0.11B	83.9	75.6	75.0	-
Token classification	BERT-base	IT	0.11B	89.8	87.0	82.3	-
GLiNER	deberta-v3-large	EN	0.44B	87.8	77.2	80.3	0.2
GLiNER	BERT-base	IT	0.11B	89.3	87.5	**84.9**	0.6
extremITLLaMA	Camoscio	IT	7B	89.1	90.3*	83.4*	0.2
SLIMER-IT	Camoscio	IT	7B	81.5	85.1	76.0	38.7
GNER	LLaMAntino-3	IT	8B	**90.3**	**88.9**	82.5	1.2
SLIMER-IT	LLaMAntino-3	IT	8B	85.8	82.5	81.7	**54.7**

Some qualitative examples are shown in Appendix A.

Impact of Backbones. Regarding the choice of the SLIMER-IT backbone, we better illustrate results in Figure 2. We can observe no remarkable difference in in-domain evaluation, where most recent models outperform older ones, as one might expect. Also globally, Camoscio and LLaMA-2-chat obtain lower scores than the rest of the backbones, with the only exception of FIC dataset, where LLaMA-3 based architecture underperform. However, LLaMAntino-3-ANITA reaches the best performance on 3 out of 4 datasets, with a strong gap especially in unseen named entities scenario, the most challenging one. Interestingly enough, thanks to their better understanding capabilities, backbones specialized on Italian are particularly effective in the unseen NEs scenario. This is the case of LLaMAntino-3-ANITA and even Camoscio, which demonstrates higher F1 than LLaMA-2.

Off-the-shelf Italian NER models. Although there has been no prior work defining a Zero-Shot NER evaluation framework for Italian, there exist fine-tune specialized state-of-the-art zero-shot NER models for Italian language. In particular, we consider: GLiNER-ML [11], a multilingual instance of GLiNER, Universal-NER-ITA[9] and GLiNER-ITA-Large[10], both specialized on Italian. These models were trained on synthetic data covering a vast number of different entity classes (up to 97k). Thus, it is impossible to directly compare them in a pure zero-shot framework, since there are no entity tags actually never-seen-before during training. However, we still report their results against SLIMER-IT. Table 2 reports the results. Despite this advantage, SLIMER-IT outperforms all these models by large a margin.

State-of-the-art comparison. Thanks to the definition of our zero-shot evaluation framework, we can compare different state-of-the-art approaches fairly. Results are outlined in Table 3. When evaluating in the same domain where the model was trained, encoder-only architectures obtain strong results despite being much smaller models. This result is not surprising, given the acknowledged performance of these architectures for supervised NER. More unexpected is their ability to generalize well to OOD inputs. Also GNER proves to be quite competitive achieving the best results in in-domain evaluation, and in OOD on FIC dataset. However, all these approaches dramatically fail on never-seen-before tags, in contrast to SLIMER-IT that achieves almost 55 F1 score points. Compared with LLM-based approaches like GNER and extremITLLaMA, this proves once again that without definition and guidelines LLMs struggle in tagging novel kind of entities.

6. Conclusions

In this paper, we proposed an evaluation framework for Zero-Shot NER that we applied to Italian. Thanks to such a framework, we can better investigate different zero-shot properties depending on the scenario (in-domain, OOD, unseen NEs). On top of that, we compared several state-of-the-art approaches, with particular focus on SLIMER, which, thanks to the usage of definition and guidelines, is well suited to deal with novel entity types. Indeed, SLIMER-IT, our fine-tuned model based on LLaMAntino-3, surpasses other state-of-the-art techniques by large margins. In the future, we plan to further extend the zero-shot NER benchmark, and implement an input caching mechanism for scalability to large label sets.

---

[9]https://huggingface.co/DeepMount00/universal_ner_ita
[10]https://huggingface.co/DeepMount00/GLiNER_ITA_LARGE

Acknowledgments

References

[1] A. P. Aprosio, T. Paccosi, Nermud at evalita 2023: Overview of the named-entities recognition on multi-domain documents task (short paper), in: International Workshop on Evaluation of Natural Language and Speech Tools for Italian, 2023. URL: https://api.semanticscholar.org/CorpusID:261529782.

[2] J. Li, A. Sun, J. Han, C. Li, A survey on deep learning for named entity recognition, IEEE Transactions on Knowledge and Data Engineering 34 (2020) 50–70.

[3] O. Sainz, et al., Gollie: Annotation guidelines improve zero-shot information-extraction, 2024. arXiv:2310.03668.

[4] W. Zhou, S. Zhang, Y. Gu, M. Chen, H. Poon, Universalner: Targeted distillation from large language models for open named entity recognition, arXiv preprint arXiv:2308.03279 (2023).

[5] A. Radford, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[6] T. Brown, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[7] J. Wei, et al., Finetuned language models are zero-shot learners, in: International Conference on Learning Representations, 2022. URL: https://openreview.net/forum?id=gEZrGCozdqR.

[8] H. W. Chung, et al., Scaling instruction-finetuned language models, 2022. arXiv:2210.11416.

[9] Y. Wang, et al., Super-Natural Instructions: Generalization via declarative instructions on 1600+ NLP tasks, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 5085–5109. URL: https://aclanthology.org/2022.emnlp-main.340. doi:10.18653/v1/2022.emnlp-main.340.

[10] X. Wang, W. Zhou, C. Zu, H. Xia, T. Chen, Y. Zhang, R. Zheng, J. Ye, Q. Zhang, T. Gui, et al., Instructuie: multi-task instruction tuning for unified information extraction, arXiv preprint arXiv:2304.08085 (2023).

[11] U. Zaratiana, N. Tomeh, P. Holat, T. Charnois, Gliner: Generalist model for named entity recognition using bidirectional transformer, 2023. arXiv:2311.08526.

[12] Y. Ding, J. Li, P. Wang, Z. Tang, B. Yan, M. Zhang, Rethinking negative instances for generative named entity recognition, 2024. arXiv:2402.16602.

[13] A. Zamai, A. Zugarini, L. Rigutini, M. Ernandes, M. Maggini, Show less, instruct more: Enriching prompts with definitions and guidelines for zero-shot ner, 2024. URL: https://arxiv.org/abs/2407.01272. arXiv:2407.01272.

[14] 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.

[15] 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: N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, D. Tapias (Eds.), 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.

[16] 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.

[17] 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.

[18] C. D. Hromei, D. Croce, V. Basile, R. Basili, Extremita at EVALITA 2023: Multi-task sustainable scaling

to large language models at its extreme, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), Parma, Italy, September 7th-8th, 2023, volume 3473 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3473/paper13.pdf.

[19] T. Paccosi, A. Palmero Aprosio, KIND: an Italian multi-domain dataset for named entity recognition, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), 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.

[20] S. Tedeschi, R. Navigli, MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation), in: Findings of the Association for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, Seattle, United States, 2022, pp. 801–812. URL: https://aclanthology.org/2022.findings-naacl.60. doi:10.18653/v1/2022.findings-naacl.60.

[21] A. Santilli, E. Rodolà, Camoscio: an italian instruction-tuned llama, 2023. URL: https://arxiv.org/abs/2307.16456. arXiv:2307.16456.

[22] H. Touvron, et al., Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.

[23] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. URL: https://arxiv.org/abs/2310.06825. arXiv:2310.06825.

[24] M. Polignano, P. Basile, G. Semeraro, Advanced natural-based interaction for the italian language: Llamantino-3-anita, 2024. URL: https://arxiv.org/abs/2405.07101. arXiv:2405.07101.

A. SLIMER-IT on some NE tags

In Table 4 we compare SLIMER-IT (LLaMAntino-based) with a version of it devoid of Definition and Guidelines (D&G), in order to get a better insight into the usefulness of such components in zero-shot NER. We present results for both unseen named entities (from Multinerd) and previously seen tags *person*, *location* and *organization*, but in out-of-domain inputs (ADG and FIC datasets). The D&G components improve performance by up to 37 points for unseen named entities, serving as a source of additional knowledge to the model and providing annotation directives about what should be labeled. Particularly for these named entities, the D&G enhance precision by reducing the number of false positives the model would otherwise generate. The performance gain provided by D&G for known tags within out-of-domain inputs is smaller, with improvements of up to 17 points on some named entity tags. In this context, the definitions and guidelines serve more as a reasoning support than as a source of additional knowledge.

Table 4

Some examples of definition and guidelines. Absolute F1 gains between SLIMER-IT and its version without definition and guidelines are reported. In green we highlight examples on unseen named entities, in blue examples on known tags such person, organization and location, but in Out-Of-Domain input distributions.

NE (dataset)	Definition & Guidelines	w/o D&G F1	w/ D&G F1	Δ F1
Corpo celeste (MN)	Definizione: 'CORPO CELESTE' si riferisce a oggetti astronomici come pianeti, stelle, satelliti, costellazioni, galassie, comete e asteroidi. Linee guida: Evita di etichettare come 'corpo celeste' entità non direttamente collegate al campo dell'astronomia. Ad esempio, 'Vergine' potrebbe riferirsi anche a un segno astrologico, quindi il contesto è importante. Assicurati di non includere nomi di fenomeni non astronomici come 'alba' o 'tramonto'. Potresti incontrare ambiguità quando un termine è usato sia in campo astronomico che in contesti non astronomici, ad esempio 'aurora' che può riferirsi sia all'evento astronomico che al nome di persona.	27.07	64.00	+36.93
Pianta (MN)	Definizione: 'PIANTA' si riferisce a organismi vegetali come alberi, arbusti, erbe e altre forme di vegetazione., Linee Guida: Quando identifichi entità 'pianta', assicurati di etichettare solo nomi di specie vegetali specifiche, come 'Fagus sylvatica', 'Suaeda vera', 'Betula pendula', evitando generici come 'alberi' o 'arbusti' se non accompagnati da una specificazione della specie.	13.76	49.89	+36.13
Media (MN)	Definizione: 'MEDIA' si riferisce a entità come nomi di giornali, riviste, libri, album musicali, film, programmi televisivi, spettacoli teatrali e altre opere creative e di comunicazione., Linee Guida: Assicurati di etichettare solo nomi specifici di opere creative e di comunicazione, evitando generici come 'musica' o 'libro'. Presta attenzione alle ambiguità, ad esempio 'Apple' potrebbe riferirsi alla società tecnologica o ad un'opera d'arte. Escludi i nomi di artisti, autori o registi, che dovrebbero essere etichettati come 'persona', e nomi generici di strumenti musicali o generi letterari che non rappresentano opere specifiche.	47.78	65.86	+18.08
Luogo (FIC)	Definizione: 'LUOGO' denota nomi propri di luoghi geografici, comprendendo città, paesi, stati, regioni, continenti, punti di interesse naturale, e indirizzi specifici., Linee Guida: Assicurati di non confondere i nomi di luoghi con nomi di persone, organizzazioni o altre entità. Ad esempio, 'Washington', potrebbe riferirsi alla città di Washington D.C. o al presidente George Washington, quindi considera attentamente il contesto. Escludi nomi di periodi storici, eventi o concetti astratti che non rappresentano luoghi fisici. Ad esempio, 'nel Rinascimento' è un periodo storico, non un luogo geografico.	59.34	76.32	+16.98
Organizzazione (ADG)	Definizione: 'ORGANIZZAZIONE' denota nomi propri di aziende, istituzioni, gruppi o altre entità organizzative. Questo tipo di entità include sia entità private che pubbliche, come società, organizzazioni non profit, agenzie governative, università e altri gruppi strutturati. Linee Guida: Annota solo nomi propri, evita di annotare sostantivi comuni come 'azienda' o 'istituzione' a meno che non facciano parte del nome specifico dell'organizzazione. Assicurati di non annotare nomi di persone come organizzazioni, anche se contengono termini che potrebbero sembrare riferimenti a entità organizzative. Ad esempio, 'Johnson & Johnson' è un'azienda, mentre 'Johnson' da solo potrebbe essere il cognome di una persona.	55.56	71.85	+16.29
Persona (FIC)	Definizione: 'PERSONA' denota nomi propri di individui umani. Questo tipo di entità comprende nomi di persone reali, famose o meno, personaggi storici, e può includere anche personaggi di finzione. Linee Guida: Fai attenzione a non includere titoli o ruoli professionali senza nomi propri (es. 'il presidente' non è una 'PERSONA', ma 'il presidente Barack Obama' sì).	79.72	83.33	+3.61