Instruct Large Language Models for Public Administration Document Information Extraction

Salvatore Carta, Alessandro Giuliani*, Marco Manolo Manca, Leonardo Piano*, Alessia Pisu and Sandro Gabriele Tiddia

Department of Mathematics and Computer Science, University of Cagliari, via Ospedale 72, Cagliari, 09124, Italy

Abstract

With the rapid digitization of institutions, there is an ever-increasing problem of effectively organizing and accessing information. Public Administrations (PAs) manage large volumes of disparate data from a variety of sources. Thus, these organizations would greatly benefit from AI, particularly Natural Language Processing solutions that help organize, structure, and search for information effectively. In the context of Italian PA, which we address in this paper, there are two main challenges: the lack of ontologies and the limited tools available for Italian information extraction. In this paper, we attempt to advance Information Extraction for Italian PAs by instructing a Large Language Model on a set of automatically labeled triplets of public tenders.

Keywords

Large Language Models, Public Administration, Tenders, Italian Open Information Extraction

1. Introduction

The pervasive impact of Information and Communication Technologies (ICT) on our society over the past two decades is undeniable. This technological revolution has permeated every aspect of society. Such a revolution has also affected Public Administrations (PAs), radically transforming how these entities operate and interact with citizens. Digital technologies have enabled PAs to streamline processes, improve service access, and increase transparency. However, along with these opportunities, significant challenges also arise in terms of data management and internal organization. Public administrations handle vast amounts of sensitive and often disparate data from various sources. Lack of data standardization, information security, and citizen privacy are crucial issues to be addressed. In addition, data fragmentation among different systems and departments can inhibit effective information sharing and analysis. For the aforementioned reasons, PAs would benefit from technology solutions based on Machine Learning and, in particular, Natural Language Processing (NLP) to improve the organization of such fragmented information.

However, there are two major challenges. The first is the lack of appropriate resources to adequately organize PA documents. Indeed, it is crucial to organize, access, understand, and utilize information with proper structures, such as knowledge graphs or ontologies, which represent a powerful solution in many domains, e.g., in online news platforms [1], health and life sciences [2], or cultural heritage [3]. In this context, Open Information Extraction (OIE) [4] represents the unique solution to structure and organize PA information. OIE systems usually adopt a domain-agnostic method and can extract entities and relationship triples (the main components of knowledge graphs) from any sentence written in natural language.

The second challenge is that a predominant part of the research conducted on OIE is mainly oriented toward the English language. While advancements in OIE have been notable, they often must encompass the complexities inherent in non-English languages. This linguistic bias significantly hinders the widespread applicability and effectiveness of OIE systems in multilingual contexts.

In this paper, we aim to advance the research on Open Information Extraction applied to PA by testing and exploiting the potential offered by Large Language Models (LLMs). In particular, a proper LLM is instructed with an effective strategy, employing proper Italian PA data.

The rest of the paper is structured as follows: Section 2 gives an overview of the state-of-the-art; our methodology is detailed in Section 3, whereas the experiments are described in Section 4. Section 5 reports and discusses the results, and Section 6 ends the paper with the conclusions.

2. Related Works

The advent of Open Information Extraction (OIE) enabled the transcendation of domain-specific constraints inherent in conventional IE methodologies. OIE meth-

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy

*Corresponding authors.

salvatore@unica.it (S. Carta); alessandro.giuliani@unica.it (A. Giuliani); marcom.manca@unica.it (M. M. Manca); leonardo.piano@unica.it (L. Piano); alessia.pisu96@unica.it (A. Pisu); sandrog.tiddia@unica.it (S. G. Tiddia)

ods aim to identify linguistic extraction patterns, either hand-crafted or automatically learned from the data [5]. Therefore, they are subdivided into rule-based or neural methods. The former include *ClausIE* [6], an OIE framework based on dependency parsing to detect clauses in an input sentence and subsequently extract proposition. *REVERB* [7] extract the tuples by isolating relation phrases that satisfy syntactic and lexical constraints. Similarly, *TEXTRUNNER* [8] first identifies a pair of noun phrases that are not too far apart, and then it applies a classifier to determine whether or not to extract a relationship. Further works rely on a proper strategy for combining different OIE tools for triplet generation and filtering [9]. A pioneering proposal regarding the more recent Neural methods is the work of Stanovsky et al. [10], wherein OIE is treated as a sequence labeling problem, and an LSTM-transducer automatically extracts triplets. Zhan and Zhao [11] introduced a span model for n-ary Open Information Extraction. More recently, Kolluru et al. [12] introduced *IMOJIE* a neural Open Information Extraction system that follows an iterative approach where the triplet extraction is conditioned by the previously retrieved triplets, with the aim of reducing redundancy.

The methods above have been developed or tested specifically for English textual corpus. Regarding the Italian language, no significant research has been conducted on Italian Open IE until the last decade. To date, only a few works have addressed such a challenge. Damiano et al. proposed *ItalIE* [13], a clause-based OIE system inspired by *ClausIE* aimed at extracting n-ary coherent propositions from simple sentences. Sentences are analyzed to identify and categorize clauses based on seven predefined patterns specific to the Italian language. Guarasci et al. [14] presented an OIE method for Italian single-verb sentences based on Lexicon-Grammar tables. The system employs linguistic structures and patterns of verbal behavior to identify arguments, match patterns, and generate propositions, demonstrating effectiveness in generating syntactically and semantically valid propositions for the Italian language. Finally, [15] proposed OIE4PA, an Open IE framework that can identify facts from Public Administration documents. Leveraging the proposal of Siciliani et al. [15], in this work, we proposed an Instructed Large Language model for Italian Open Information Extraction specialized in Public Administration Documents.

3. Methodology

We propose a novel model for automated Information Extraction for Italian PAs by *instructing* an LLM on a set of automatically labeled triplets of public tenders. To this end, we devise a proper strategy to train an LLM with a suitable set of triplets and instructions. The entire process is depicted in Figure 1.

Our method involves two stages. In detail, the process first performs a step aimed at obtaining a correctly annotated set of triplets (*Triplet Auto-Labeling*), which is subsequently used to train the LLM (*Instruction Tuning*). Each step is described in the following.

3.1. Triplet Auto-Labeling

The first step of our methodology is training a Sequence Classifier Language Model to identify meaningful triplets within the PA context. To accomplish this, we leveraged the dataset OIE4PA, consisting of a collection of triplets extracted from Italian tenders of the Apulia region [15]. In particular, each triplet is extracted with the WikiOIE framework [16]. Specifically, the dataset is organized into two sets: a labeled set $\mathcal{L}$, which contains a subset of 2000 binary triplets labeled by humans as valid or not, and an unlabeled set $\mathcal{U}$ of 14,096 triplets, together with the original sentences. Then, at this stage, we exploited the $\mathcal{L}$ set to properly train a classifier to distinguish between valid and invalid triplets. To do this, we treated this task as a sentence classification problem, concatenating triplets into a single sentence and separating subject, predicate, and object by a semicolon. To this end, we identified three suitable Language Models (LMs) for this task, namely Italian-bert, LegalBert [17], and BureauBERTo [18]. The former is a Bert base model [19] fine-tuned on Italian corpus, the second is a fine-tuned version of Italian Bert on Italian civil law corpora, and the last is an UmBERTO model fine-tuned on PA, banking, and insurances corpus. Table 1 outlines the results obtained by these three Language Models on the triplet classification task. Finally, the trained most accurate classifier has been employed to label the triplets of the U set, forming a new $\mathcal{AL}$ (Auto-Labeled) set, which in turn will be exploited to instruct the Large Language Model for the OIE task.

3.2. Instruction Tuning

Instruction tuning is an innovative strategy that involves guiding a language model through human-like instructions to improve its performance on a specific task. Unlike traditional methods that rely solely on large-scale training data, instruction tuning provides targeted guidance, allowing the model to adapt and refine its behaviour toward desired outcomes. Incorporating human-like instructions enhances the model's understanding and improves its ability to generate contextually relevant responses. In summary, given a source text and task-specific instructions, the model is trained to create a sequence of tokens representing the desired output.

To instruct an LLM to perform Open Information Extraction, we transformed the $\mathcal{AL}$ triplets set into an *instruction dataset*- In particular, each auto-labeled triplet

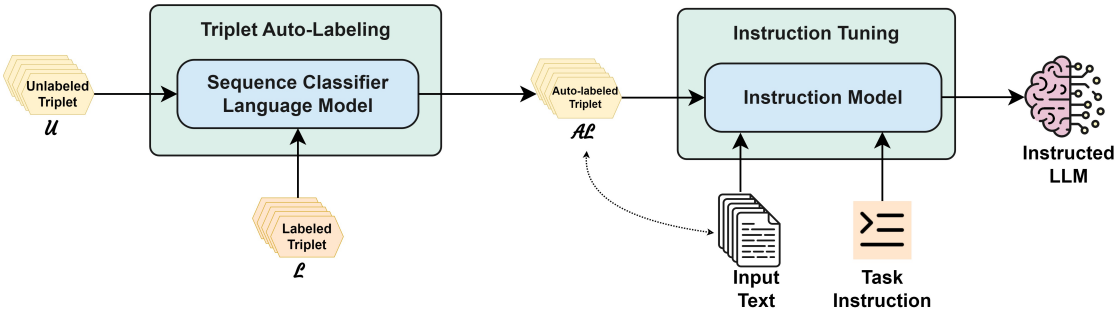

Figure 1: Instructed model training.

is used to train the Instruction model following the template: *Task Instruction*, *Input Text*, and *Response*.

3.2.1. Task instruction

Task instructions provide a detailed statement on accomplishing the desired task and properly structuring the output. In detail, we formulated the following instruction to query LLM:

```
<Trova quali triple semantiche esistono
nel testo. Formatta l'output come
[Soggetto;Predicato;Oggetto]>.
```

We formulate the instruction in Italian to make the model immediately understand that we are referring to the Italian language. The translation in English of the instruction is:*"Find which semantic triples exist in the text, Format the output as [Subject; Predicate; Object]"*.

3.2.2. Input text

The input text represents the sentences in which LLM has to perform the task defined by the instructions. In detail, each sentence is the original text excerpt from which a triplet belonging to the dataset OIE4PA has been extracted.

3.2.3. Response

The response represents the desired output. In our case, the input sentence was transformed into an open triplet. We also specify that to instruct the model to distinguish sentences where a triplet can be extracted from sentences where no useful triplets exist, we included the triplet as a response if it was labeled as valid by the classifier; otherwise, we leave an empty string.

4. Experimental settings

We adopted the `Flan-T5` family [20] as an instruction model. Such a choice is motivated by two reasons: first, prior research [21] has demonstrated the potential of such models in Information Extraction tasks, eventually outperforming larger models such as LLama2 or similar, resulting in a perfect trade-off between speed of inference and prediction quality. The other main benefit is that `Flan-T5` is a multi-language model, which is also suitable for tasks related to understanding Italian. We tested with two different T5-Flan sizes *flan-xxl (11b)* and *flan-xl (3b)* adopting for both the OIE4PA dataset, relying on a split of 80% and 20% for training and test, respectively. We fine-tuned the models for efficiency and hardware reasons by exploiting `QLora`[1] with a 4-bit quantization, allowing faster training and saving GPU memory. All experiments were conducted with an Nvidia RTX A6000 GPU machine with 48 GB of VRAM. We train both models for one epoch, and we adopt the following QLora settings and hyperparameters:

Lora-rank	16
Lora-alpha	32
Lora-dropout	0.05
Learning rate	0.003
Batch Size	8

4.1. Evaluation Metrics

To properly apply such metrics for the triplets evaluation, we considered as *true positive* (TP) a non-empty triplet that matches with the corresponding triplet in the ground truth (i.e., the triples belonging to the $\mathcal{AL}$ set), *true negative* (TN) a triple returned as an empty string by the model and labeled as invalid in the ground truth, *false positive* (FP) a triplet that was labeled as invalid but retrieved by the model, and *false negative* (FN) when

[1]https://github.com/artidoro/qlora

the model returned an empty string rather than a valid triplet.

In doing so, we can evaluate the performances in terms of classical confusion matrix metrics, i.e., *accuracy (a)*, *precision (p)*, *recall (r)*, and *F1 score (F1)*; whose formulae are:

$$a = \frac{TP + TN}{TP + TN + FP + FN}$$

$$p = \frac{TP}{TP + FP}$$

$$r = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * p * r}{p + r}$$

5. Results

Table 1 reports the comparisons of three different Italian Bert models for the triplet classification task. In detail, the selcted models are `LegalBERT-ITA`[2], `BertBase-ITA`[3], and `BureauBERTo`[4]. The best model turns out to be BureauBerto, probably due to the fact that it is the only model pre-trained on Public Administration corpora.

Table 1
Bert triplet classification results in terms of accuracy (a), precision (r), recall (r), and F1 score ($F1$).

Model	a	p	r	$F1$
LegalBERT-ITA	0.935	0.953	0.897	0.919
BertBase-ITA	0.927	0.935	0.894	0.911
BureauBERTo	**0.945**	**0.963**	**0.901**	**0.932**

Table 2 outlines the result of the two fine-tuned Flan-T5 models on extracting triplets in procurement texts. Both model sizes show excellent results for all metrics; in particular, recall is significantly high, demonstrating that the models are quite effective in finding a large number of true positives (e.g., valid triplets). It is also good to note that the values are higher for the model with a higher number of parameters. Therefore, the promising results support the thesis of leveraging Instruction Tuning to build strong Open Information Extraction models for Italian public administrations. To this end, we plan to create new datasets in the future to develop a new set of foundational models for information extraction in Italian, with a particular focus on PAs and other administrative entities.

[2]https://huggingface.co/dlicari/Italian-Legal-BERT
[3]https://huggingface.co/dbmdz/bert-base-italian-uncased
[4]https://huggingface.co/colinglab/BureauBERTo

Table 2
FLAN-OpenIE results on OIE4PA dataset in terms of accuracy (a), precision (r), recall (r), and F1 score ($F1$).

Model	a	p	r	$F1$
T5-xl	0.78	0.74	0.97	0.84
T5-xxl	**0.82**	**0.78**	**0.99**	**0.87**

6. Conclusions

Considering the significant gap between information extraction available for English and other resource-constrained languages such as Italian, we explored an Instruction Tuning approach to perform Open Information Extraction on Italian Public Tenders in this paper. A proper LLM is instructed with an effective two-stage strategy, in which a language model-based classifier is trained on a proper Italian PA dataset to obtain a set of correct triplets, which are used to instruct a suitable LLM. The promising experiments have validated the assumptions pointed out in the paper and incentivized future developments aimed at developing new datasets and models capable of theoretically understanding and structuring technical texts in Italian in the form of semantics triplets.

Acknowledgments

This work has been partially carried out thanks to the Ministerial Decree no. 351 of 9th April 2022, based on the NRRP – funded by the European Union - NextGenerationEU - Mission 4 "Education and Research", Component 1 "Enhancement of the offer of educational services: from nurseries to universities" - Investment 4.1, that provided a financial support for the Leonardo Piano's doctoral pathway.

Also, Alessia Pisu acknowledge MUR and EU-FSE for financial support of the PON Research and Innovation 2014-2020 (D.M. 1061/2021).

Furthermore, we acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5 - Call for tender No.3277 published on December 30, 2021 by the Italian Ministry of University and Research (MUR) funded by the European Union – NextGenerationEU. Project Code ECS0000038 – Project Title eINS Ecosystem of Innovation for Next Generation Sardinia – CUP F53C22000430001- Grant Assignment Decree No. 1056 adopted on June 23, 2022 by the Italian Ministry of University and Research (MUR).

References

[1] C. Rudnik, T. Ehrhart, O. Ferret, D. Teyssou, R. Troncy, X. Tannier, Searching news articles using an event knowledge graph leveraged by wikidata, in: S. Amer-Yahia, M. Mahdian, A. Goel, G. Houben, K. Lerman, J. J. McAuley, R. Baeza-Yates, L. Zia (Eds.), Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019, ACM, 2019, pp. 1232–1239.

[2] P. Ernst, C. Meng, A. Siu, G. Weikum, Knowlife: A knowledge graph for health and life sciences, in: 2014 IEEE 30th International Conference on Data Engineering, 2014, pp. 1254–1257. doi:10.1109/ICDE.2014.6816754.

[3] S. Carta, G. Fenu, A. Giuliani, M. M. Manca, M. Marras, L. Piano, A. S. Podda, L. Pompianu, S. G. Tiddia, Empowering digital transformation in tourism through intelligent methods for representation and exploitation of cultural heritage knowledge, volume 3536, 2023, p. 83 – 91. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85177612618&partnerID=40&md5=7e8334f126d9385a733fbfb0d1674f19.

[4] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, O. Etzioni, Open information extraction from the web, in: Proceedings of the 20th International Joint Conference on Artifical Intelligence, IJCAI'07, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2007, pp. 2670–2676.

[5] C. Niklaus, M. Cetto, A. Freitas, S. Handschuh, A survey on open information extraction, in: E. M. Bender, L. Derczynski, P. Isabelle (Eds.), Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 3866–3878. URL: https://aclanthology.org/C18-1326.

[6] L. Del Corro, R. Gemulla, Clausie: clause-based open information extraction, in: Proceedings of the 22nd international conference on World Wide Web, 2013, pp. 355–366.

[7] A. Fader, S. Soderland, O. Etzioni, Identifying relations for open information extraction, in: Conference on Empirical Methods in Natural Language Processing, 2011.

[8] A. Yates, M. Banko, M. Broadhead, M. J. Cafarella, O. Etzioni, S. Soderland, Textrunner: Open information extraction on the web, in: North American Chapter of the Association for Computational Linguistics, 2007. URL: https://api.semanticscholar.org/CorpusID:1455080.

[9] S. Carta, P. Fariello, A. Giuliani, L. Piano, A. S. Podda, S. G. Tiddia, Sailgenie: Sailing expertise to knowledge graph through open information extraction, in: G. A. Tsihrintzis, C. Toro, S. A. Ríos, R. J. Howlett, L. C. Jain (Eds.), Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 27th International Conference KES-2023, Athens, Greece, 6-8 September 2023, volume 225 of *Procedia Computer Science*, Elsevier, 2023, pp. 2224–2233. URL: https://doi.org/10.1016/j.procs.2023.10.213. doi:10.1016/J.PROCS.2023.10.213.

[10] G. Stanovsky, J. Michael, L. Zettlemoyer, I. Dagan, Supervised open information extraction, in: North American Chapter of the Association for Computational Linguistics, 2018.

[11] J. Zhan, H. Zhao, Span model for open information extraction on accurate corpus, in: AAAI Conference on Artificial Intelligence, 2019. URL: https://api.semanticscholar.org/CorpusID:208138002.

[12] K. Kolluru, S. Aggarwal, V. Rathore, Mausam, S. Chakrabarti, Imojie: Iterative memory-based joint open information extraction, ArXiv abs/2005.08178 (2020). URL: https://api.semanticscholar.org/CorpusID:218674382.

[13] E. Damiano, A. Minutolo, M. Esposito, Open information extraction for italian sentences, in: 2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA), 2018, pp. 668–673. doi:10.1109/WAINA.2018.00165.

[14] R. Guarasci, E. Damiano, A. Minutolo, M. Esposito, G. De Pietro, Lexicon-grammar based open information extraction from natural language sentences in italian, Expert Systems with Applications 143 (2020) 112954. URL: https://www.sciencedirect.com/science/article/pii/S0957417419306724. doi:https://doi.org/10.1016/j.eswa.2019.112954.

[15] L. Siciliani, E. Ghizzota, P. Basile, P. Lops, Oie4pa: open information extraction for the public administration, Journal of Intelligent Information Systems (2023) 1–22.

[16] L. Siciliani, P. Cassotti, P. Basile, M. de Gemmis, P. Lops, G. Semeraro, A. Moro, Extracting relations from italian wikipedia using self-training (2021).

[17] D. Licari, G. Comandè, ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law, in: D. Symeonidou, R. Yu, D. Ceolin, M. Poveda-Villalón, D. Audrito, L. D. Caro, F. Grasso, R. Nai, E. Sulis, F. J. Ekaputra, O. Kutz, N. Troquard (Eds.), Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management, volume 3256 of *CEUR Workshop Proceedings*, CEUR, Bozen-Bolzano, Italy, 2022. URL: https://ceur-ws.org/Vol-3256/#km4law3, iSSN: 1613-0073.

[18] S. Auriemma, M. Madeddu, M. Miliani, A. Bondielli, L. C. Passaro, A. Lenci, Bureauberto: adapting

umberto to the italian bureaucratic language, in: Ital-IA, 2023. URL: https://api.semanticscholar.org/CorpusID:262088765.

[19] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: North American Chapter of the Association for Computational Linguistics, 2019. URL: https://api.semanticscholar.org/CorpusID:52967399.

[20] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, arXiv preprint arXiv:2210.11416 (2022).

[21] S. Wadhwa, S. Amir, B. C. Wallace, Revisiting relation extraction in the era of large language models, Proceedings of the conference. Association for Computational Linguistics. Meeting 2023 (2023) 15566–15589. URL: https://api.semanticscholar.org/CorpusID:258564662.