A semi-automatic framework towards building Electricity Grid Infrastructure Management ontology: A case study and retrospective

Abdelhadi Belfadel, Maxence Gagnant, Matthieu Dussartre, Jérôme Picault, Laure Crochepierre, Sana Tmar

A semi-automatic framework towards building Electricity Grid Infrastructure Management ontology: A case study and retrospective

Abdelhadi Belfadel[1], Maxence Gagnant[1], Matthieu Dussartre[2], Jérôme Picault[2], Laure Crochepierre[2], and Sana Tmar[1]

[1] IRT SystemX,
2 Bd Thomas Gobert, 91120 Palaiseau, France
`firstname.lastname@irt-systemx.fr`
[2] RTE Réseau de transport d'électricité,
7C place du Dôme, 92073 Paris La Defense Cedex, France
`firstname.lastname@rte-france.com`

Abstract. Thanks to their extensive use in Internet-based applications, ontologies have gained significant popularity and recognition within the semantic web domain. They are widely regarded as valuable sources of semantics and interoperability in artificial intelligence systems. With the exponential growth of unstructured data on the web, there is a pressing need for automated acquisition of ontologies from unstructured text. This research area has seen the emergence of various methodologies that leverage techniques from machine learning, text mining, knowledge representation and reasoning, information retrieval, and high level natural language processing. These new techniques represent an opportunity to introduce automation into the process of ontology acquisition from unstructured text. To this end, this contribution offers a semi-automatic framework with a concrete usage of a tooled NLP-based approach to design an application ontology in a real-world industrial context. We discuss the state of the art analysis, the challenges met and the technological choices for the realization of this approach. Specifically, we explore its application in the real-world scenario of RTE's power grid event management.

Keywords: Ontology · NLP · Information Extraction · Power grid

1 Introduction

In the modern age of big data, vast amounts of diverse, unstructured information are generated daily across various professions. Adapting this data for real-time decision-making, while integrating expert knowledge and external sources, is a challenge. The semantic web, with ontologies, addresses this by providing meaningful information representation for humans and computers.

An ontology formally structures knowledge in a specific domain, including concepts, relations, attributes, and hierarchies. Building ontologies manually is time-intensive, requiring extraction of instances from unstructured text via ontology population. However, creating large ontologies manually is challenging [2], prompting a shift toward automated ontology population. This shift promotes exploring automatic ontology learning as an alternative to manual design.

This work introduces a semi-automated approach to build an application ontology for power grid event management. It addresses the lack of practical NLP-based ontology learning experiments and shares insights from the ongoing implementation and results. The following sections cover the conceptual framework, technical decisions, current progress, and encountered limitations in this real-world industrial context.

2 Industrial Case Study: Electricity Grid Infrastructure Management

RTE (Réseau de Transport d'Électricité) is the electrical transmission system operator (TSO) in France. It is an independent public company in charge of ensuring the smooth operation, safety and reliability of the French high-voltage electricity network. As a transmission system operator, RTE plays a crucial role in the coordination and management of electricity flows in France. In this context, electricity network operators handle critical documents for reporting grid operations and incident management. These reports enhance internal communication and inform decision-making, but they lack standardization in content and structure. Ontologies provide coordination among operators. Similarly, machine-generated documents like real-time monitoring data, archives, forecasts, network models and simulations play a crucial role in the efficient management and forecasting of power flows. Using ontology to model these documents bridges the gap between machines and operators, enhancing reliability and power system performance.

In order to meet RTE's ambitions, the proposed method involves four objectives: i) The generation of a Knowledge Graph (KG) from operating reports; ii) KG enrichment from information contained in the real-time (monitoring) or anticipatory databases; iii) The design of an application ontology allowing to represent RTE specific knowledge in a formal way; iv) The ultimate objective is to automatically generate operating reports, while allowing AI assistants to offer communication adapted to the operators' vocabulary.

The results of this research will benefit power grid operators by facilitating optimized decision making, refining the understanding of network phenomena and ensuring good continuity of service. The rest of this paper will focus on objectives (i) and (iii) to elaborate on these crucial steps.

3 State of The Art

In recent decades, there has been significant interest in ontology engineering [18], resulting in a multitude of studies exploring methodologies, guidelines, tools, resources, and ongoing research in various related areas such as ontology learning topic.

To shift from a handcrafting development process to a (semi)-automatic process, several techniques emerging from the fields of machine learning, natural language processing, data mining or information retrieval have been proposed. Authors in [2] summarizes the various steps required for ontology learning from an unstructured text. It starts by extracting terms and synonyms using linguistic techniques, then relations between these concepts are found based on statistical techniques such as co-occurrence analysis or clustering. Finally, axioms are extracted thanks to inductive logic programming techniques such as [4, 17].

Named Entity Recognition (NER) is a core NLP method that leverages machine learning and linguistic patterns to identify and categorize specific elements like people, organizations, places, and dates in text, offering insights into document structure. NER aids ontology development by extracting entities that can be mapped to ontology concepts [7]. In order to predict relations between entities in a given sentence, relation extraction techniques are of high value. Early rule-based methods [1] had difficulty generalizing as relation syntax differ within situations. Therefore deep learning models were introduced, in particular the max-pooling CNN model [20] which has long remained the most efficient structure at classifying relations. More recently, large language models [22, 11] proved to be better at capturing semantics in sentences. They achieve state-of-the art results on all benchmark datasets. However, training large supervised models typically requires huge amounts of data examples.

Using knowledge bases for extensive training sets, [16] introduced distant supervision for relation extraction. Graph neural networks perform well in such cases [3]. Yet, this supervision can be noisy, as it relies on the strong assumption that the relation to extract is the same as the one found in the knowledge base. Semi-supervision addresses data scarcity: self-training models [9] generate artificial examples, and self-ensembling models [13] jointly label and generate. Despite notable progress in relation classification and improved results on benchmarks, limited research targets extracting unknown relation types. Unsupervised models constitute a decent alternative as they do not require prior knowledge. They first represent sentences in specific semantic spaces then apply clustering to extract different relation types [10]. Further works integrate hierarchy information which appears to be useful to relation clustering [21].

4 Overview of the Framework and Implementation Steps

4.1 Framework for ontology learning

To achieve our objectives, we have established the following steps (Figure 1):

- *Pre-processing*: It involves syntactic analysis of unstructured text (example in Figure 2). It uses NLP techniques like part-of-speech tagging, French lemmatization, and sentence parsing to label words, normalize terms, and reduce data dimension.

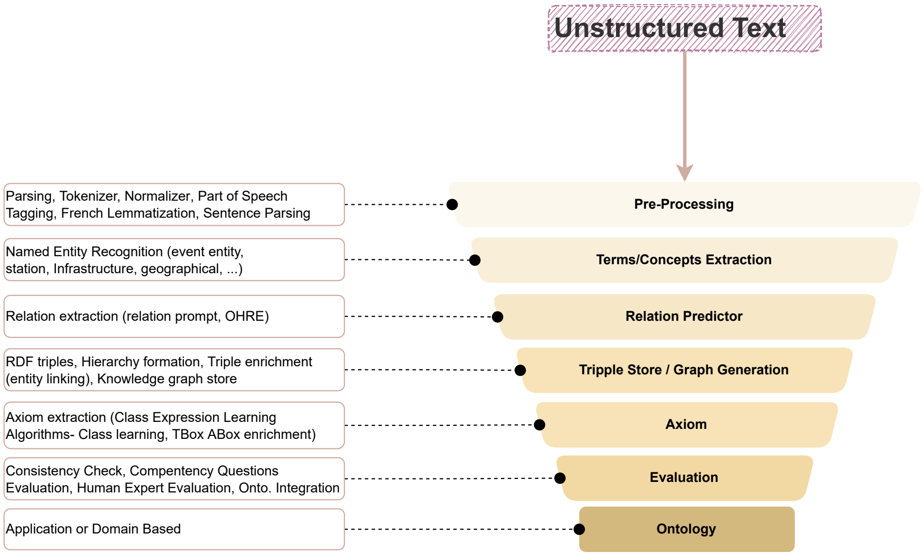

Unstructured Text
Parsing, Tokenizer, Normalizer, Part of Speech Tagging, French Lemmatization, Sentence Parsing
Pre-Processing
Named Entity Recognition (event entity, station, Infrastructure, geographical, ...)
Terms/Concepts Extraction
Relation extraction (relation prompt, OHRE)
Relation Predictor
RDF triples, Hierarchy formation, Triple enrichment (entity linking), Knowledge graph store
Tripple Store / Graph Generation
Axiom extraction (Class Expression Learning Algorithms- Class learning, TBox ABox enrichment)
Axiom
Consistency Check, Compentency Questions Evaluation, Human Expert Evaluation, Onto. Integration
Evaluation
Application or Domain Based
Ontology

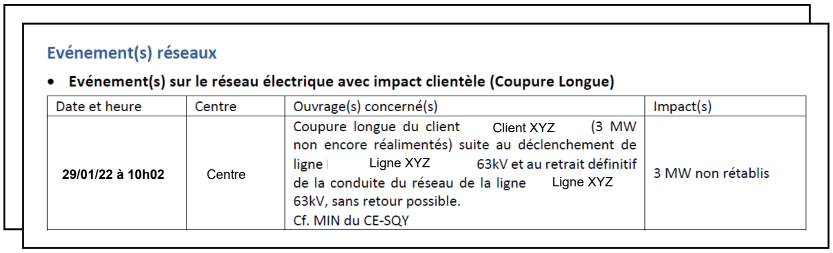
Fig. 1: Framework for ontology learning

- *Terms/Concepts Extraction*: This relies on the use of linguistic techniques for the extraction of important terms/concepts of a domain in the unstructured text, and the relations between them. This is achieved by using NER techniques to recognize important information in the text that refer to key subjects in our context, such as transmission lines, stations, infrastructures, as well as geographical locations, electrical grid events,...
- *Relation Predictor*: This important step helps identify the relation between the extracted terms/concepts, and enables during the next steps the creation of a knowledge graph. This step relies on language models which capture the semantics of sentences. More details about the chosen algorithms from the literature are described in the workflow of Figure 3.
- *Triple Store/Graph Generation*: This step uses previous results to form an initial data graph. Each term and relation becomes an atomic data entity in the Resource Description Framework (RDF) [6] data model, a W3C standard for web data interchange. Entity linking enhances resources with syntactic and semantic details from broad or domain-specific ontologies like DBpedia[3] or the Common Grid Model Exchange Standard (CGMES) ontology[4]. This establishes initial class expressions, concepts, and hierarchy. Ontology learning then constructs, learns, and enriches the graph by identifying class expressions and axioms, laying the foundation for the application ontology.
- *Axiom*: This step employs logic programming to uncover patterns among concepts in the knowledge graph, extracting axiom schemata and general ax-

[3] https://www.dbpedia.org/
[4] https://www.entsoe.eu/data/cim/cim-for-grid-models-exchange/

ioms. Techniques like supervised learning can identify expressed OWL classes using positive and negative examples [12]. Alternatively, a class learning approach employs class instances as positive examples, learning about the class and its relations in the graph.
- *Evaluation*: Evaluating ontology acquisition is vital to assess concept coverage, correctness, and suitability. It refines learning processes, aligning with user needs. Literature offers techniques like golden standard, application-based, data-driven, and human evaluation, outlined in [2].

The remainder of this paper focuses on the results of the framework's initial steps, including pre-processing, term/concept extraction, relation extraction, and RDF triple generation. It discusses the chosen algorithms and the lessons learned from analyzing existing methods. Figure 3 presents a workflow that depicts the implemented technical tasks.



Evénement(s) réseaux

- **Evénement(s) sur le réseau électrique avec impact clientèle (Coupure Longue)**

Date et heure	Centre	Ouvrage(s) concerné(s)	Impact(s)
29/01/22 à 10h02	Centre	Coupure longue du client Client XYZ (3 MW non encore réalimentés) suite au déclenchement de ligne Ligne XYZ 63kV et au retrait définitif de la conduite du réseau de la ligne Ligne XYZ 63kV, sans retour possible. Cf. MIN du CE-SQY	3 MW non rétablis

Fig. 2: Anonymized unstructured text example from a PDF file. Each PDF contains diverse event descriptions in various formats (tables, figures, paragraphs, etc.).

4.2 NLP Pipeline

The blue and yellow NLP tasks in Figure 3 are implemented using the Spacy framework [8], an open-source, flexible software library for advanced natural language processing, written in Python. Given the input documents specificity (technical french content in a specialized field), our NLP pipeline entails: 1) standard Spacy components, 2) third-party components, and 3) custom RTE components. Noteworthy design choices for components include:

- *Tokenization level*: in the RTE corpus, documents mention power grid equipments, including voltage levels like "le poste 400kV" or "le poste 400 kV". When using the standard Spacy tokenizer for French, "400kV" becomes separate tokens, which is not intended. Specific components in our NLP pipeline handle this situation.
- *Part-of-Speech (POS) tagging level*: the default Spacy French component provides quite poor performance for POS tagging, this is why we decided to replace the default component with a component based on LEFFF[19], developed by INRIA, which is superior by far.

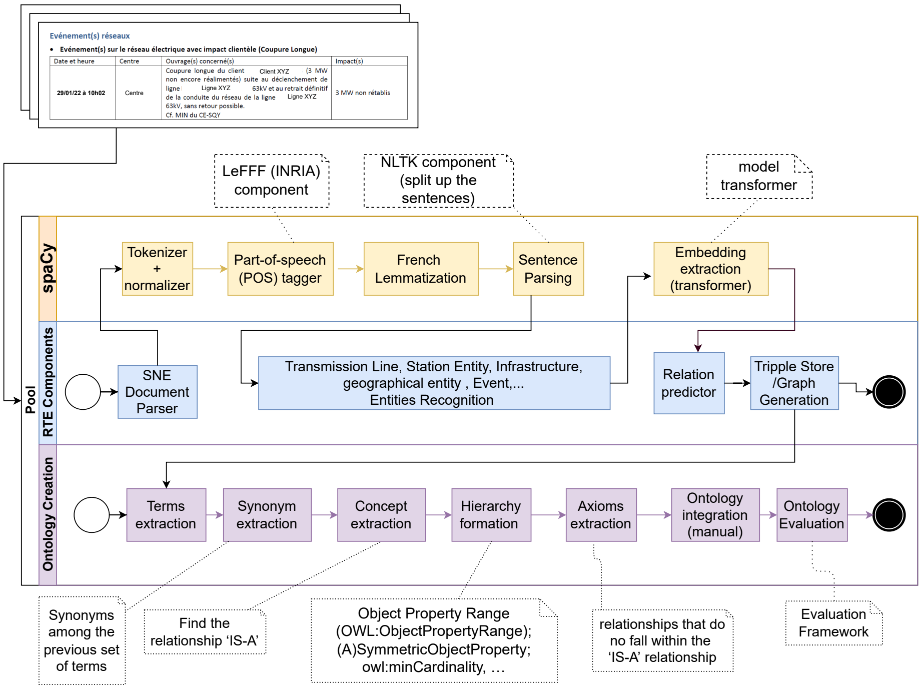
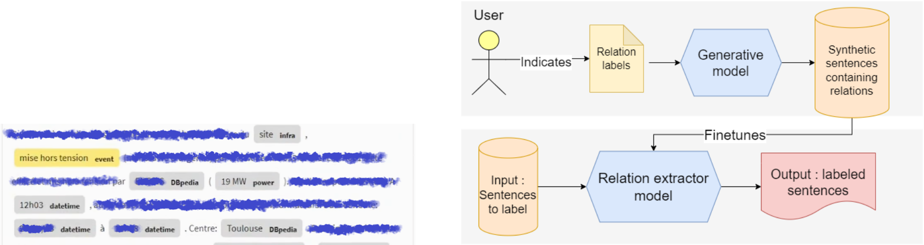

Evénement(s) réseaux
Evénement(s) sur le réseau électrique avec impact clientèle (Coupure Longue)
Date et heure
Centre
Ouvrage(s) concerné(s)
Impact(s)
29/01/22 à 10h02
Centre
Coupure longue du client Client XYZ (3 MW non encore réalimentés) suite au déclenchement de ligne Ligne XYZ 63kV et au retrait définitif de la conduite du réseau de la ligne Ligne XYZ 63kV, sans retour possible.
Cf. MIN du CE-SQY
3 MW non rétablis
LeFFF (INRIA) component
NLTK component (split up the sentences)
model transformer
Pool
spaCy
Tokenizer + normalizer
Part-of-speech (POS) tagger
French Lemmatization
Sentence Parsing
Embedding extraction (transformer)
RTE Components
SNE Document Parser
Transmission Line, Station Entity, Infrastructure, geographical entity , Event,... Entities Recognition
Relation predictor
Tripple Store /Graph Generation
Ontology Creation
Terms extraction
Synonym extraction
Concept extraction
Hierarchy formation
Axioms extraction
Ontology integration (manual)
Ontology Evaluation
Synonyms among the previous set of terms
Find the relationship 'IS-A'
Object Property Range (OWL:ObjectPropertyRange); (A)SymmetricObjectProperty; owl:minCardinality, …
relationships that do no fall within the 'IS-A' relationship
Evaluation Framework

Fig. 3: Proposed Workflow for a semi-automatic construction of an application Ontology

- *Sentencizer level*: similarly, the Spacy sentencizer component has a rather strange behaviour, which considers soft punctuation marks (such as ";", "-") as delimiters for sentences. This behaviour is not the one we expect, we used instead a component based on NLTK [14].
- *Entity level*: we targeted operations-related entities in the power grid, such as substations, lines, transformers, and associated grid events. Spacy's NER lacks the necessary specificity, therefore we developed custom components to detect a wider array of pertinent entities. Figure 4a depicts an output of this step.
- *Embeddings*: RTE documents contain a lot of technical vocabulary and even French terms meaning may diverge from common language usage. Thus, a specific transformation of tokens and entities is required to obtain meaningful numerical vectors (embeddings) describing accurately the documents. This is achieved by fine-tuning a language model, CamemBERT [15] (a French transformer model), with 5000+ technical documents.
- *Relation predictor*: a new component is designed to establish meaningful links between entities, forming the foundation for semi-automated ontology creation. In the unique context of power grid management, this involves extracting highly specific relation types. While traditional relation extraction models like [11] seemed promising at first, they didn't deliver the expected

real-case RTE results. This led to the exploration of open relation extraction models, in particular [10] and [21]. However, their unsupervised approach, while suitable for new relations and limited data, was discarded due to its tendency to produce relation clusters instead of labels. For automated knowledge graph and ontology creation, relation labeling is vital. Existing solutions lacked adaptability to specific contexts, relying on initial datasets. Therefore, the Relation Prompt solution [5] gained prominence. This dual relation extraction method consists of a powerful BART-based model operating the extraction, coupled with a generative model. Below, we outline the relation prompt process shown in Figure 4b.


[redacted] site infra ,
mise hors tension event [redacted]
[redacted] par [redacted] DBpedia (19 MW power) [redacted]
12h03 datetime , [redacted]
[redacted] datetime à [redacted] datetime . Centre: Toulouse DBpedia [redacted]
User
Indicates
Relation labels
Generative model
Synthetic sentences containing relations
Finetunes
Input : Sentences to label
Relation extractor model
Output : labeled sentences

(a) Named Entity Recognition result (anonymized extract)

(b) Workflow of the Relation Prompt solution

Fig. 4

Figure 4b illustrates the relation prediction step. The generator provides artificial sentences depicting a given relation. Then, the extractor is fine-tuned with these synthetic sentences in order to adapt classification to the desired relation types. This flexible approach only requires a very small amount of training sentences which fitted well to the context. Providing quality annotations and specific relations for our business data is indeed costly in terms of time and thus money. Table 1 shows the details of the RTE dataset and some targeted relations. Indeed, some very specific relations such as a relation between an `Event` and `Event Criticality` or `Event Type` are difficult to extract through this NLP process due to the absence of this knowledge in the unstructured text. In addition, only a very small amount of sentences was available with very specific relation types which constitutes a real challenge to training.

Results presented in Table 1 show that some relations such as `occurred at date`, `occurred at time`, or `of voltage` were effectively classified, even though there was very few training examples. Meanwhile, others are very bad classified. Several reasons can explain poor results : some relations involving proper nouns such as locations, infrastructures or clients are tougher to perceive as proper names inherently bear very little resemblance to each other. Also, some relations such as `geographical proximity` are rare in the set and thus not learnt enough.

Finally, the `ending time` relation is more ambiguous than others and appears in very different ways which explains why it is harder to classify.

Entities	Relation types	Training examples	Test examples	Precision
Event ; Datetime	occurred at date	18	6	100%
Event ; Datetime	occurred at time	17	4	100%
Event ; Geo. Region	occurred in geographical region	21	6	17%
Event ; Client	with participation of client	4	2	0%
Event ; Power	of electrical power	18	6	33%
Event ; Datetime	ending time	23	6	0%
Geo. Region ; Geo. Region	geographical proximity	4	2	0%
Infrastructure ; Voltage level	of voltage	39	3	100%
Event; Infrastructure;	occurred on infrastructure	16	2	0%
-	Total	160	37	43%

Table 1: Composition and results on RTE dataset

4.3 Knowledge graph generation

The transformation of NLP entities and relations into a knowledge graph involves two steps: 1) Entity Linking: this associates entity types with targeted RDF classes. A straightforward match currently based on existing ontology drafts, for instance, the "outage" extracted as an "event" type has to appear in RDF as an instance of an "Event" class. 2) Instance Uniqueness: in the knowledge graph, identical entities from different relation triples appear only once, grouped by textual similarity within a paragraph. However, dealing with different instances mentioned in different ways and in various contexts would require more extensive analysis.

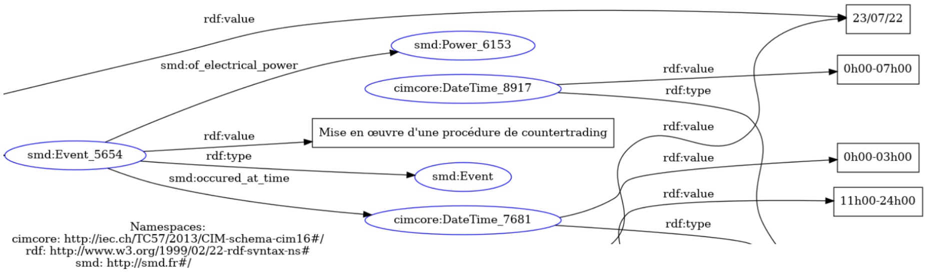


Fig. 5: Excerpt from a generated knowledge graph

5 Conclusion

This paper presents a framework that introduces automation into the process of ontology acquisition from unstructured text. We provide, as a retrospective in each step, the challenges met, the methodologies and technological choices for the realization of this approach on a real-world industrial context applied for power grid events management. In terms of implementation, our efforts were primarily directed towards the upper levels of the suggested framework. Our objective was to transform domain-specific unstructured text into a knowledge graph representation, with the aim of generating an initial draft of an application ontology as a subsequent step. To accomplish this, we employed a NLP-based approach that utilizes syntactic and linguistic techniques. Additionally, we employed a specialized language model that generates synthetic training data to support low-resource relation extraction methods.

As for the work to come in the next few months, our aim is twofold. Firstly, we intend to implement the OWL class expression learner algorithm to construct class expressions and relevant axioms using the extracted and observed dataset. This effort strives to create an initial and formal application ontology, capitalizing on the specified steps detailed in this paper, along with technical insights and subsequent assessment. Additionally, our ultimate objective is to develop a system that can automatically generate feedback documents and enable AI assistants to provide communication that is tailored to the operators' vocabulary. This will enhance the effectiveness and adaptability of RTE communication process.

6 Acknowledgement

This work has been supported by the French government under the "France 2030" program, as part of the SystemX Technological Research Institute.

References

1. Aone, C., Halverson, L., Hampton, T., Ramos-Santacruz, M.: SRA: Description of the IE2 system used for MUC-7. In: Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998 (1998), https://aclanthology.org/M98-1012
2. Asim, M.N., Wasim, M., Khan, M.U.G., Mahmood, W., Abbasi, H.M.: A survey of ontology learning techniques and applications. Database **2018** (2018)
3. Bastos, A., Nadgeri, A., Singh, K., Mulang', I.O., Shekarpour, S., Hoffart, J., Kaul, M.: Recon: Relation extraction using knowledge graph context in a graph neural network (2021)
4. Bühmann, L., Lehmann, J., Westphal, P.: Dl-learner—a framework for inductive learning on the semantic web. Journal of Web Semantics **39**, 15–24 (2016)
5. Chia, Y.K., Bing, L., Poria, S., Si, L.: Relationprompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction (2022)
6. Cyganiak, R., Wood, D., Lanthaler, M., Klyne, G., Carroll, J.J., McBride, B.: Rdf 1.1 concepts and abstract syntax. W3C recommendation **25**(02), 1–22 (2014)
7. Elgamal, M., Abou-Kreisha, M., Elezz, R., Hamada, S.: An ontology-based name entity recognition ner and nlp systems in arabic storytelling. Al-Azhar Bulletin of Science **31**, 31–38 (12 2020). https://doi.org/10.21608/absb.2020.44367.1088

8. Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017), https://spacy.io/
9. Hu, X., Zhang, C., Ma, F., Liu, C., Wen, L., Yu, P.S.: Semi-supervised relation extraction via incremental meta self-training (2021)
10. Hu, X., Zhang, C., Xu, Y., Wen, L., Yu, P.S.: Selfore: Self-supervised relational feature learning for open relation extraction. arXiv preprint arXiv:2004.02438 (2020)
11. Huguet Cabot, P.L., Navigli, R.: REBEL: Relation extraction by end-to-end language generation. In: Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 2370–2381. Association for Computational Linguistics, Punta Cana, Dominican Republic (Nov 2021), https://aclanthology.org/2021.findings-emnlp.204
12. Lehmann, J., Auer, S., Bühmann, L., Tramp, S.: Class expression learning for ontology engineering. Journal of Web Semantics **9**(1), 71–81 (2011)
13. Lin, H., Yan, J., Qu, M., Ren, X.: Learning dual retrieval module for semi-supervised relation extraction (2019)
14. Loper, E., Bird, S.: Nltk: The natural language toolkit (2002), https://arxiv.org/abs/cs/0205028
15. Martin, L., Müller, B., Suárez, P.J.O., Dupont, Y., Romary, L., de la Clergerie, É.V., Seddah, D., Sagot, B.: Camembert: a tasty french language model. CoRR **abs/1911.03894** (2019), http://arxiv.org/abs/1911.03894
16. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. pp. 1003–1011. Association for Computational Linguistics, Suntec, Singapore (Aug 2009), https://aclanthology.org/P09-1113
17. Muggleton, S., De Raedt, L., Poole, D., Bratko, I., Flach, P., Inoue, K., Srinivasan, A.: Ilp turns 20: biography and future challenges. Machine learning **86**, 3–23 (2012)
18. Poveda-Villalón, M., Fernández-Izquierdo, A., Fernández-López, M., García-Castro, R.: Lot: An industrial oriented ontology engineering framework. Engineering Applications of Artificial Intelligence **111**, 104755 (2022)
19. Sagot, B.: The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In: 7th international conference on Language Resources and Evaluation (LREC 2010). Valletta, Malta (May 2010), https://inria.hal.science/inria-00521242
20. Zeng, D., Liu, K., Lai, S., Zhou, G., Zhao, J.: Relation classification via convolutional deep neural network. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pp. 2335–2344. Dublin City University and Association for Computational Linguistics, Dublin, Ireland (Aug 2014), https://aclanthology.org/C14-1220
21. Zhang, K., Yao, Y., Xie, R., Han, X., Liu, Z., Lin, F., Lin, L., Sun, M.: Open hierarchical relation extraction. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 5682–5693 (2021). https://doi.org/10.18653/v1/2021.naacl-main.452
22. Zhou, W., Huang, K., Ma, T., Huang, J.: Document-level relation extraction with adaptive thresholding and localized context pooling (2020)