Automatic Nested Spatial Entity and Spatial Relation Extraction From Text for Knowledge Graph Creation: A Baseline Approach and a Benchmark Dataset

Helen Mair Rawsthorne, Nathalie Abadie, Eric Kergosien, Cécile Duchêne, Éric Saux

Automatic Nested Spatial Entity and Spatial Relation Extraction From Text for Knowledge Graph Creation: A Baseline Approach and a Benchmark Dataset

Helen Mair Rawsthorne
Nathalie Abadie
helen.rawsthorne@ign.fr
LASTIG, Univ Gustave Eiffel, IGN-ENSG
Saint-Mandé, France

Eric Kergosien
GERiiCO, Université de Lille
Villeneuve d'Ascq, France

Cécile Duchêne
LASTIG, Univ Gustave Eiffel, IGN-ENSG
Champs-sur-Marne, France

Éric Saux
IRENav, École navale
Brest, France

ABSTRACT

Automatically extracting geographic information from text is the key to harnessing the vast amount of spatial knowledge that only exists in this unstructured form. The fundamental elements of spatial knowledge include spatial entities, their types and the spatial relations between them. Structuring the spatial knowledge contained within text as a geospatial knowledge graph, and disambiguating the spatial entities, significantly facilitates its reuse. The automatic extraction of geographic information from text also allows the creation or enrichment of gazetteers. We propose a baseline approach for nested spatial entity and binary spatial relation extraction from text, a new annotated French-language benchmark dataset on the maritime domain that can be used to train algorithms for both extraction tasks, and benchmark results for the two tasks carried out individually and end-to-end. Our approach involves applying the Princeton University Relation Extraction system (PURE), made for *flat*, *generic* entity extraction and *generic* binary relation extraction, to the extraction of *nested*, *spatial* entities and *spatial* binary relations. By extracting *nested* spatial entities and the *spatial* relations between them, we have more information to aid entity disambiguation. In our experiments we compare the performance of a pretrained monolingual French BERT language model with that of a pretrained multilingual BERT language model, and study the effect of including cross-sentence context. Our results reveal very similar results for both models, although the multilingual model performs slightly better in entity extraction, and the monolingual model has slightly better relation extraction and end-to-end performances. We observe that increasing the amount of cross-sentence context improves the results for entity extraction whereas it has the opposite effect on relation extraction.

GeoHumanities '23, November 13, 2023, Hamburg, Germany

ACM ISBN 979-8-4007-0349-2/23/11…$15.00
https://doi.org/10.1145/3615887.3627754

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Natural language processing**; **Information extraction**.

KEYWORDS

geographic information, spatial knowledge, maritime data, nested spatial entity, binary spatial relation, deep learning, neural network, language model

ACM Reference Format:
Helen Mair Rawsthorne, Nathalie Abadie, Eric Kergosien, Cécile Duchêne, and Éric Saux. 2023. Automatic Nested Spatial Entity and Spatial Relation Extraction From Text for Knowledge Graph Creation: A Baseline Approach and a Benchmark Dataset. In *7th ACM SIGSPATIAL International Workshop on Geospatial Humanities (GeoHumanities '23), November 13, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3615887.3627754

1 INTRODUCTION

1.1 Motivation

Some spatial knowledge, current or historical, exists only in the form of text. Examples of such sources of unstructured spatial knowledge include travel guides, historical documents and social media posts. These sources can hold information about individual spatial entities that is absent from reference geographic resources[1] such as alternative names [2], and can even mention spatial entities that are missing entirely from reference geographic resources despite being or having been present in the local culture or being part of shared community knowledge [3, 23]. Such texts can also harbour spatial knowledge about the environment at a larger scale that does not exist elsewhere, for example how it is perceived, how it behaves and how it can be navigated [12, 16]. Text-based sources contain naturally heterogeneous spatial knowledge: they can be written by different authors, from different points of view, different names in potentially different languages can be used to refer to the same places [2, 14], they can cover large and diverse geographic areas, and crucially they can contain varied levels of detail and be vague or imprecise [9]. It is therefore difficult to integrate

[1] DBpedia is a frequently-used global reference geographic resource. BD TOPO® is a reference geographic resource for the French territory.

geographic information from text-based sources into geographic information system (GIS) models, which require highly-structured data. The open-world assumption of semantic Web technologies makes knowledge graphs a better solution for modelling and storing geographic information extracted from natural language text, and thus making it accessible and reusable [5, 13, 22, 33]. Harnessing the spatial knowledge contained within text by extracting it and structuring it as a geospatial knowledge graph opens up a vast range of possibilities. Structured geographic information can be queried or processed in order to provide access to it in other forms and it can be enhanced by linking it to other sources of information [13, 22]. It also makes it possible to verify its coherence and infer new facts thanks to reasoning [10, 27, 35]. The extraction of the geographic information contained within reference texts also allows the creation and enrichment of gazetteers. Essential for the structuring and processing of data in the humanities, gazetteers should provide alternative names across time, space and languages, as well quantitative and qualitative descriptions for places and their locations [8, 32].

During knowledge graph population, spatial entities become instances of ontological classes and spatial relations become assertions of object properties. To be able to correctly assign a spatial entity to its corresponding ontological class, it is necessary to know its *type*. We make the assumption that the geographic name of a spatial entity often contains a common noun that indicates its type, such as *port* in "Port of Liverpool". Although this holds for many of the Romance languages, it is not applicable to all languages. Sometimes, the geographic name of a spatial entity contains more than one type noun, such as in "Robben Island Lighthouse". This increases the complexity of identifying a spatial entity's true type from its geographic name.

Whilst *flat* spatial entity extraction would simply aim to capture "Port of Liverpool" or "Robben Island Lighthouse" as the name of a spatial entity without seeking any further definitions, *nested* spatial entity extraction allows defining multiple layers of labels for the same text. We use the labels introduced by [23], the definitions of which are as follows: geographic feature refers to common nouns that represent *types* of spatial entities, name refers to pure proper nouns and geographic name refers to the full name associated with a geographic feature. For our first example, *nested* spatial entity extraction would therefore aim to capture "Port of Liverpool" as the geographic name, "Port" as the geographic feature and "Liverpool" as the name. For our second example, nested spatial entity extraction would aim to capture initially "Robben Island" as one geographic name, with "Island" as the geographic feature and "Robben" as the name, and then also capture "Robben Island Lighthouse" as another geographic name, with "Lighthouse" as the geographic feature and "Robben Island" as the name. This layered approach facilitates the identification of the correct type in cases where the geographic name contains multiple instances of a geographic feature.

By extracting *nested* as opposed to *flat* spatial entities, the geographic feature type of the entity is already known and an instance of the right ontology class can be created automatically. In some cases, its name gives an indication of its geographical location. These two extra pieces of information, the entity *type* and its *name*, facilitate the disambiguation task of linking the instance to the correct entity in a reference geographic resource [32].

By extracting the spatial relations between entities, assertions of object properties can automatically be created between instances. This information can also be used to aid disambiguation of named and unnamed entities, and increase confidence in the results thanks to spatial reasoning [26]. In the case where a reference entity does not yet exist, a new entry can be created in the geographic resource, supported by the class and property information of the instance. The same reasoning applies to the creation and enrichment of gazetteers: by specifically identifying entity types during the extraction process, gazetteer entries can automatically be classed or assigned attributes and can more easily be disambiguated. The identification and extraction of the spatial relations in which spatial entities take part can increase the level of detail available in descriptions of gazetteer entries and their locations.

Texts that cover an international environment are likely to contain geographic names in languages other than the main language of the text. It is important that this does not hinder the extraction process: geographic names and entity types written in other languages should still be identified as such. The state of the art in information extraction from text relies on deep neural network language models [25]. Such models can be trained to deal with one or multiple languages and are referred to as pretrained *monolingual* or *multilingual* language models respectively. A multilingual ontology can then be used to aid the disambiguation of entities whose type is written in other languages [33].

1.2 Application Context

We developed the work presented in this paper as part of a project to structure the geographic information contained within the *Instructions nautiques*. The *Instructions nautiques* are a series of French-language books produced and published by the *Service hydrographique et océanographique de la Marine* (Shom), the French Naval Hydrographic and Oceanographic Service. Each volume contains essential information for navigating safely in the coastal waters of a specific geographic area, including instructions for entering ports and descriptions of the coastal maritime environment. Many national hydrographic services produce their own versions of the *Instructions nautiques*, which are commonly known as Sailing Directions, in other languages.

The aim of the project is to construct a geospatial knowledge graph of the content of the *Instructions nautiques*. This would offer new possibilities to improve the production chain, maintenance process and user experience of the *Instructions nautiques*.

A geospatial knowledge graph of the content of the *Instructions nautiques* could transform the manual processes for producing and updating the *Instructions nautiques* that are currently used by the Shom. Instead of manually analysing the entire series of *Instructions nautiques* to determine the impact that a new piece of information will have, and instead of manually searching for the lines to be updated, the knowledge base could be queried to automatically identify the relevant lines in the text.

To improve the efficiency and accuracy of these processes, and thereby increase the reliability of the *Instructions nautiques*, the

Shom could apply reasoning to the knowledge graph to automatically identify and correct errors that would otherwise put the users of the *Instructions nautiques* in danger. For example, the spatial relations between entities as described in the text could be verified by using the geographic positions of the entities and vice versa. To increase the exhaustiveness of the textual content of the *Instructions nautiques*, the Shom could use inference rules to infer new knowledge from the knowledge already present in the knowledge graph. For example, the description of a spatial entity whose geographic position is described only by geographic coordinates could be improved by adding a description of its position in relation to other nearby entities.

If the knowledge graph contained multilingual labels, a semi-automatic or automatic text generation system could be implemented to help produce high-quality automatic translations of the text of the *Instructions nautiques*, making them quickly and easily available in other languages and thereby increasing their potential user base. If the information contained within the other nautical publications produced by the Shom was also structured in geospatial knowledge graphs, the Shom could link related pieces of information from different sources. For example, instead of simply citing another publication, the *Instructions nautiques* text could be linked directly to the relevant text in the other publication.

To modernise user access to the content of the *Instructions nautiques*, the geospatial knowledge graph could be used as the basis of a digital platform, allowing users to interrogate the content of the *Instructions nautiques* via faceted search in different languages or even by selecting their area of interest on a nautical chart. The Shom could integrate knowledge and information from internal and trusted external sources to this platform to reduce the number of different resources needing to be consulted by users of the *Instructions nautiques* during itinerary planning. For example, live access to the tide predictions and weather forecast for the time and place indicated by the user could be provided.

For the remainder of this paper we will use examples from the corpus of *Instructions nautiques* and the coAsTaL mAritime Navigation InstructionS (ATLANTIS) Ontology[2] [29], an extract of which can be seen in figure 1, to illustrate our approach for spatial entity and relation extraction from text.

Nevertheless, the approach that we present is generalisable to other types of corpora and could be applied to train a base encoder for nested entity and binary relation extraction on any corpus, containing geographic information or not and written in any language. An annotation scheme specific to the new corpus would need to be developed and implemented to create a training dataset.

1.3 Contributions and Code Availability

The main contributions of this paper are as follows. First, we introduce a new annotated French-language dataset on the maritime domain for nested spatial entity and binary spatial relation extraction from text that we have published online[3]. Second, we present a baseline approach for nested spatial entity and binary spatial relation extraction from text that is an adaptation of the existing Princeton University Relation Extraction system (PURE) [44] for generic entity and relation extraction to the extraction of *nested*, *spatial* entities and *spatial* relations. We use the code provided in [44] combined with a modified annotation format that we demonstrate at the end of section 3.1. Our approach is suited to being applied to corpora in any language (provided that it features nested entity names) covering any domain, be it scientific or literary, historical or contemporary, fiction or non-fiction. Third, we provide benchmark results for our dataset for three tasks: nested spatial entity extraction, binary spatial relation extraction, and end-to-end spatial entity and relation extraction. The spatial entities and relations extracted from our corpus could directly enrich reference geographic resources or gazetteers, which could be used for applications in domains such as hydrography, maritime navigation or the humanities. Finally, we compare the performance of the *bert-base-french-europeana-cased*[4] pretrained monolingual French Bidirectional Encoder Representations from Transformers (BERT) language model with that of the *bert-base-multilingual-cased*[5] pretrained multilingual BERT model for the three tasks, and study the effect of the cross-sentence context window size on the performances of both models.

1.4 Outline

In section 2 we review the state of the art in entity and relation extraction from text. We also review literature that compares the performance of pretrained monolingual and multilingual language models in text-based deep learning tasks. In section 3 we present our annotated dataset and its preparation process, and then describe the approach that we implemented to perform nested spatial entity and binary spatial relation extraction from text. In section 4 we present and analyse the results obtained for these tasks using a pretrained monolingual French BERT model and using a pretrained multilingual BERT model before concluding and discussing future work in section 5.

2 RELATED WORK

In their infancy, entity and relation extraction from text were primarily performed using rule-based approaches that required manually developing rules built on grammar, syntax and punctuation to identify them. Classical machine learning approaches were then developed and achieved consistently higher performances in these tasks, and a trend away from rule-based approaches was documented [24]. Research in both tasks is now dominated by approaches that apply deep learning techniques, which is why we only consider such techniques for review, although slower progress is being made in relation extraction [20, 25, 42]. Given that the tasks of entity extraction and relation extraction are not always studied together, we review work that deals with either task or both. We consider approaches designed for generic entities and relations as well as those dedicated to spatial entities and relations, which are much less common.

2.1 Flat Spatial Entity Extraction

Current work on flat spatial entity extraction is dominated by the use of bidirectional long short term memory (BiLSTM) and

[2]https://github.com/umrlastig/atlantis-ontology
[3]https://github.com/umrlastig/atlantis-dataset
[4]https://huggingface.co/dbmdz/bert-base-french-europeana-cased
[5]https://huggingface.co/bert-base-multilingual-cased

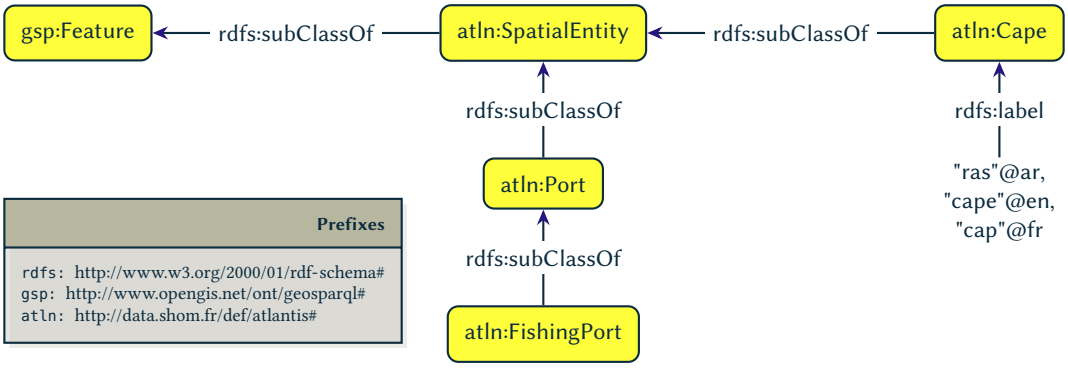

Figure 1: A simplified extract of the ATLANTIS Ontology [29].

Transformer [39] models. Unless explicitly stated otherwise, all approaches presented here aim to identify flat spatial entities in text as locations or place names without identifying their type. Novel BiLSTM and Transformer models developed specifically for the extraction of flat named spatial entities from English text are introduced in [3] and are shown to outperform existing solutions for entity extraction. A new Transformer model for flat spatial entity extraction from English text is also presented in [11], for named as well as unnamed spatial entities. Its performance is compared with many other systems and is shown to give better results in all cases. The sensitivity of spaCy convolutional neural network (CNN) models to optical character recognition (OCR) noise is compared with that of Stanza BiLSTM models during flat named spatial entity extraction from historical French text in [18]. Historical French texts undergo normalisation preprocessing in [17] to improve the results of flat named spatial entity extraction using various machine learning approaches as well as a spaCy CNN model. A Transformer model is coupled with a BiLSTM model in [36] to perform flat named spatial entity extraction from Chinese text. The approach is tested on different datasets, for some of which the entities are classified by type.

2.2 Nested Generic Entity Extraction

Nested entity extraction has been tackled using a variety of methods including layered approaches and joint labelling. Layered approaches consist of training different models for each possible class in the nested entity structure, effectively resulting in stacked flat entity extraction modules. The models can be trained independently, thereby avoiding error propagation, or information can be passed between them to improve context representations. This is the case in [15], where fine-grained entities are extracted first and each subsequent layer extracts more complex entities, using information encoded in previous layers, until no more entities are found. To avoid training multiple models, the joint labelling approach presented in [1] can be used. It requires training only one model as all the labels that correspond to a single token at different levels of nesting are concatenated to form one label. This technique has been improved in [38] by the addition of class-based weights in the loss function which penalise semantically-distant classes more severely.

2.3 Binary Spatial Relation Extraction

Spatial relation extraction has been little studied independently of spatial entity extraction. A CNN modified to deal specifically with spatial relations in Chinese is presented in [28]. The training and test datasets are composed of sentences that contain exactly two spatial entities and exactly one spatial relation each. A method for spatial relation extraction from Chinese text that combines a Transformer model with a bidirectional gated recurrent unit (BiGRU) and an attention mechanism is proposed in [43]. As with the previous study, it has the disadvantage of relying on training and test datasets composed of sentences that contain one spatial relation between one pair of spatial entities.

2.4 Combined Entity and Relation Extraction

Deep learning approaches that aim to tackle both entity and relation extraction can either separate the two tasks and dedicate an independent neural network to each, keep the two tasks separate with dedicated neural networks but allow information to be shared between them, or model the two tasks together and have a single neural network perform both tasks together. We refer to the former as a *pipelined* approach whilst the latter two are known as *joint modelling* approaches. It is shown in [41] that for spatial entity and relation extraction, a pipelined approach is the most effective of the two. A pipelined approach for generic flat entity and generic binary relation extraction from English text is presented in [44]. Known as the Princeton University Relation Extraction system (PURE), it trains two separate encoders, one for each task, from existing pretrained deep language models. It is shown that cross-sentence information should be taken into account during the training of both the entity model and the relation model, as well as during the prediction phases, to maximise results. This pipelined approach with cross-sentence context outperforms joint modelling systems on standard benchmarks using Bidirectional Encoder Representations from Transformers (BERT) models. Flat spatial entity and binary spatial relation extraction from French text are carried out using a pipelined approach in [4]. A BiLSTM neural network coupled with a BERT model is trained for spatial entity extraction whilst spatial relation extraction is based on a dependency parsing method using a Stanza BiLSTM model.

2.5 Monolingual vs. Multilingual Language Models

Many experiments have been done to determine whether the multilingual BERT language model first presented in [7] performs better than monolingual language models for monolingual texts. Such work includes [30] for Arabic, English, Finnish, Indonesian, Japanese, Korean, Russian, Turkish and Chinese, [21] for French, [6] for Dutch, [37] for Vietnamese and [40] for Marathi. All of these papers show better results using the monolingual language models for most if not all of the tasks evaluated.

2.6 Summary

We chose to implement PURE [44] and apply it to *nested*, *spatial* entities and the binary *spatial* relations between them. The authors of PURE do not attempt nested entity extraction nor do they specifically target spatial entities and relations. Our review shows that monolingual BERT language models perform better than the multilingual model on monolingual texts. We decided to investigate the effects of a dataset containing words in multiple languages other than the main language of the text, as is the case in our dataset, on the performances of monolingual and multilingual models. This is especially important given that the words in other languages are almost always part of a spatial entity name.

3 METHOD

3.1 Dataset Preparation

Our dataset is made up of extracts from each one of the 15 volumes of the *Instructions nautiques* that we had at our disposal. The volumes, which are written in French, cover coastal areas in Africa, Europe, North and South America, as well as in the Indian and Pacific Oceans. We extracted the text from the PDF documents using pdfminer.six[6].

We annotated our dataset by hand using the brat rapid annotation tool[7], which allows creating nested labelled annotations and creating directed labelled links between them [34]. The annotation scheme, described below, was designed and agreed-upon by all authors. One author carried out the annotation of the dataset and then extracts were cross-checked and validated by all authors. The source text was split on whitespace by brat, giving a dataset of 101,400 tokens.

Given that we wished to perform *nested* spatial entity extraction to simultaneously capture the full name of the spatial entity as well as its type and name, we implemented a nested labelling approach using the labels defined in section 1. Any token can be annotated with zero or one geographic feature or name label. A token cannot be annotated with both a geographic feature and a name label. A token cannot be annotated only with a name label. A token annotated with a name label must also be annotated one or more times with a geographic name label. Any token, already labelled or not, can be annotated zero or more times with a geographic name label.

An extremely large number of different types of spatial relations are used in our corpus so we decided to limit ourselves to extracting only those that would be the most useful during the disambiguation process. The cardinal directions are heavily relied upon in navigation because spatial relations that employ them are constructed using an absolute frame of reference, which means that no viewpoint is involved [19]. We chose to extract the spatial relations that employ the cardinal directions because of their frequent use and unambiguity. This amounts to 16 relation types in total: four that use the cardinal directions (N, E, S, W), four that use the intercardinal directions (NE, SE, SW, NW) and eight that use the secondary intercardinal directions (NNE, ENE, ESE, SSE, SSW, WSW, WNW, NNW). In our corpus, these spatial relations are always referred to by using these 16 one-, two- and three-letter abbreviations, for example "*le port est au NW de la ville*" ("the port is to the NW of the town") or "*la tour est à l'ESE du château*" ("the tower is to the ESE of the castle"). The 16 labels that correspond to these spatial relations are of the format "is XYZ of", where "XYZ" is one of the 16 cardinal direction abbreviation.

We identified three other types of spatial relations to capture more information about domain-specific spatial entities that are often unnamed in the corpus or are likely to be absent from reference geographic resources such as navigation marks (buoys, beacons, etc.), rocks or sandbanks. First, the "is off the coast of" label is used when it is indicated that one spatial entity is located off the coast of, or in the coastal waters of, another. It is therefore frequently used to locate isolated spatial entities. This type of spatial relation, which is also constructed using an absolute frame of reference, is always referred to using the same three words in our corpus: "*au large de*" ("is off the coast of"). Second, the "is marked by" label is used for any spatial entity that is marked or pointed out by another deliberately-placed entity, often a navigation mark, either when the former poses a danger to navigators or when it allows a safe passage: "*Son musoir est marqué par un feu.*" ("Its pierhead is marked by a light.)" [31]. This relation indicates a proximity between the two entities and is expressed in a number of different ways in our corpus: "*est marqué par*" ("is marked by") can alternatively be expressed as "*est signalé par*" ("is flagged by") or "*est indiqué par*" ("is indicated by"). Third, the "is an element of" label indicates a topological relation that includes entities that are situated *on* or *in* another such that a bird's eye view shows the spatial footprint of one as being within or partly within the other. This relation is expressed in a wide variety of different ways in our corpus, including implicitly, and rarely includes the word "*élément*" ("element"). For example, "*l'île porte un phare*" ("the island boasts a lighthouse"), "*le feu établi sur le quai*" ("the light located on the quay") and "*les haut-fonds de la baie*" ("the sandbanks of the bay") all indicate a "is an element of" relation.

All relation annotations must link two entity annotations, either geographic feature or geographic name labels. All relation annotations must have a direction. Instead of duplicating the relation labels to account for their inverses and create directed relation annotations that always go in the direction of the text: "A →is marked by→ B" and "C →marks→ D", we created one version for each label and allow directed relation annotations that go in either direction: "A →is marked by→ B" and "C ←is marked by← D".

After having annotated exactly one section from each of the 15 volumes of the *Instructions nautiques*, our dataset was considerably lacking in some relation labels, in particular those using

[6]https://github.com/pdfminer/pdfminer.six
[7]http://brat.nlplab.org

the secondary intercardinal directions. To increase the number of examples of these relations we semi-automatically extracted random sentences containing the keywords NNE, ENE, ESE, SSE, SSW, WSW, WNW and NNW from the remaining text of each volume, manually annotated them and added them to our dataset.

Figure 2 shows a sentence from the *Instructions nautiques* annotated according to our nested spatial entity and spatial relation annotation scheme. The specific labelling of the geographic feature "*ras*" ("cape") within the geographic name combined with multilingual label values in an ontology means that this spatial entity could automatically be instantiated in the correct class regardless of the language in which the geographic feature is written, which in this case is romanised Arabic. Figure 3 shows a set of Resource Description Framework (RDF) triples that could automatically be constructed from the information extracted from this sentence according to the extract of the coAsTaL mAritime NavigaTion InstructionS (ATLANTIS) Ontology shown in figure 1.

We split our annotated dataset into three parts: train, development and test, aiming to keep a 80:10:10 ratio of overall number of tokens and of numbers of entity labels. We also ensured that text covering each geographic area was present in all three parts. Our dataset of 101,400 tokens contains 16,777 entity labels (which can span one or more tokens) and 3,051 relation labels (which connect exactly two entity labels in a given direction). In total, 18,030 tokens are annotated with at least one entity label, which corresponds to almost one in five tokens. We will refer to these manual annotations in our dataset as *gold annotations*. The dataset composition is summarised in table 1 and table 2 shows for the label distribution.

We converted our dataset from the brat standoff format to the JSON Lines (JSONL)[8] format required for PURE using a Python script[9]. Figure 4 shows the same annotated sentence as in figure 2 converted to a JSON value in this format. In our case, each JSON value corresponds to one paragraph and contains a list of sentences (each of which is a list of tokens), a list of label and span combinations that correspond to the entity annotations (boundary token pair + label), and a list of label and span pair combinations that correspond to the relation annotations (ordered pair of boundary token pairs + label). This nested annotation format that allows any token to be annotated with zero or more labels makes it possible to perform nested entity extraction without using joint labelling.

Table 1: Number of tokens and labels per split in the dataset. A single entity label can span one or more tokens.

	Train	Dev.	Test	Total
Tokens	83,851	8,156	9,393	101,400
Unlabelled tokens	69,200	6,507	7,663	83,370
Entity-labelled tokens	14,651	1,649	1,730	18,030
Entity labels	13,582	1,476	1,719	16,777
Relation labels	2,507	222	322	3,051

Table 2: Detailed dataset composition. Entity and relation label distribution per dataset split.

Label	Train nb.	Dev nb.	Test nb.	Total nb.
all entity labels	13,582	1,476	1,719	16,777
geographic feature	6,602	692	801	8,095
name	3,486	391	462	4,339
geographic name	3,494	393	456	4,343
all relation labels	2,507	222	322	3,051
is an element of	1,300	109	190	1,599
is marked by	143	13	17	173
is off the coast of	21	1	1	23
is N of	84	9	8	101
is NNE of	46	1	3	50
is NE of	47	1	5	53
is ENE of	72	11	6	89
is E of	92	6	11	109
is ESE of	73	8	13	94
is SE of	42	4	1	47
is SSE of	51	10	3	64
is S of	84	8	12	104
is SSW of	86	6	7	99
is SW of	45	2	5	52
is WSW of	75	10	6	91
is W of	75	10	11	96
is WNW of	76	4	5	85
is NW of	32	2	8	42
is NNW of	63	7	10	80

3.2 Model Training and Testing

PURE [44] independently trains two base encoders from existing pretrained deep language models: one to identify and label entity spans, and one to identify related pairs of entity spans and classify the relation between them. We will refer to the former as the *entity model* and the latter as the *relation model*. It also allows the regulation of the size of the context window W, that is to say the amount of cross-sentence context that is made available for the model. The context made available during the processing of a given sentence spans from $(W-n)/2$ words to the left of the sentence to $(W-n)/2$ words to the right, where n is the number of words in the sentence. A cross-entropy loss is used for both models. For the base encoders we used *bert-base-french-europeana-cased*[10] as our pretrained monolingual French BERT model and *bert-base-multilingual-cased* as our pretrained multilingual BERT model. We used the default hyperparameters provided by [44], shown in table 4, and experimented over multiple context window sizes, within the ranges of default values, for both the entity model and the relation model. For the entity model we used context windows of 0, 50, 100, 150, 200 and 248 ($W = 250$ exceeded our available GPU memory usage) and for the relation model we used context windows of 0, 50 and 100. We

[8] A JSONL file contains one valid JSON value on each line.

[9] https://github.com/dwadden/dygiepp/blob/master/scripts/new-dataset/brat_to_input.py

[10] The popular pretrained monolingual French BERT model *CamemBERT* presented in [21] is not compatible with PURE. In order to keep an identical workflow for the training of the monolingual and multilingual models we chose to use *bert-base-french-europeana-cased*, which is compatible with PURE. This model is pretrained primarily on 20th-century texts. We judge that its pretraining is well suited to our corpus of the *Instructions nautiques*, which are written in formal language.

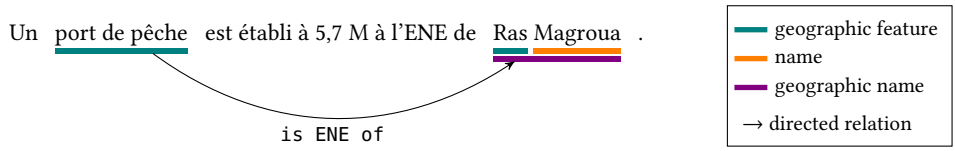

Figure 2: A sentence from the *Instructions nautiques* annotated according to our nested spatial entity and binary spatial relation annotation scheme. Sentence translated from the original French text into English: "A fishing port lies 5.7 M to the ENE of Ras Magroua." [31]

Table 3: Detailed dataset results. [Prec.|Rec.|F1] [ent] gives the mean [precision|recall|micro F1-score] over five runs for entity extraction for each entity label using the context window that gives the best overall results ($W = 248$ for monolingual, $W = 200$ for multilingual). [Prec.|Rec.|F1] [rel] gives the mean [precision|recall|micro F1-score] over five runs for relation extraction for each relation label from gold entity annotations using the context window that gives the best overall results ($W = 0$ for monolingual and for multilingual). [Prec.|Rec.|F1] [e2e] gives the mean [precision|recall|micro F1-score] over five runs for end-to-end entity and relation extraction for each relation label from the best predicted entity annotations using the context window that gives the best overall results ($W = 0$ for monolingual and for multilingual). For each task, the highest precision, recall and F1-score over both base encoders is in bold.

	Prec. [ent|rel]		Prec. [e2e]		Rec. [ent|rel]		Rec. [e2e]		F1 [ent|rel]		F1 [e2e]	
Label	Mono.	Multi.	Mono.	Multi.	Mono.	Multi.	Mono.	Multi.	Mono.	Multi.	Mono.	Multi.
all entity labels	94.6	**95.2**	-	-	**89.8**	89.6	-	-	92.2	**92.3**	-	-
geographic feature	94.1	94.4	-	-	95.8	95.1	-	-	95.0	94.8	-	-
name	97.7	97.4	-	-	78.0	78.4	-	-	86.7	86.9	-	-
geographic name	92.9	94.9	-	-	91.3	91.4	-	-	92.1	93.1	-	-
all relation labels	**70.8**	67.2	**70.2**	67.3	58.8	**59.9**	58.8	**59.8**	**64.2**	63.2	**63.9**	63.2
is an element of	70.5	67.9	70.2	68.2	60.2	60.3	60.2	60.2	64.9	63.8	64.8	63.9
is marked by	64.9	55.1	61.3	55.1	51.8	49.4	51.8	49.4	57.5	50.8	56.1	50.8
is off the coast of	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
is N of	48.6	39.8	47.7	41.8	52.5	50.0	52.5	50.0	49.9	44.2	49.7	45.2
is NNE of	76.7	37.3	76.7	39.3	73.3	53.3	73.3	53.3	74.1	43.3	74.1	44.8
is NE of	95.0	100.0	95.0	100.0	64.0	60.0	64.0	60.0	76.1	75.0	76.1	75.0
is ENE of	83.0	93.3	96.0	93.3	63.3	83.3	63.3	83.3	71.6	87.9	75.9	87.9
is E of	62.5	57.1	67.9	57.1	45.5	41.8	45.5	41.8	52.6	48.1	54.4	48.1
is ESE of	73.4	71.1	70.9	71.1	55.4	66.2	55.4	66.2	62.6	67.8	61.6	67.8
is SE of	90.0	60.0	90.0	60.0	100.0	100.0	100.0	100.0	93.3	73.3	93.3	73.3
is SSE of	73.7	81.3	71.7	81.3	73.3	66.7	73.3	66.7	68.8	69.3	67.1	69.3
is S of	84.7	82.1	84.7	82.1	71.7	81.7	71.7	81.7	77.3	81.1	77.3	81.1
is SSW of	87.4	74.1	87.4	74.1	68.6	85.7	68.6	85.7	76.0	79.3	76.0	79.3
is SW of	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
is WSW of	92.0	83.6	92.0	83.6	66.7	70.0	66.7	70.0	77.1	75.3	77.1	75.3
is W of	79.3	80.4	75.8	78.9	65.5	60.0	65.5	60.0	71.3	68.0	69.5	67.3
is WNW of	100.0	89.3	100.0	89.3	88.0	88.0	88.0	88.0	93.3	87.9	93.3	87.9
is NW of	67.3	87.4	80.0	87.4	35.0	50.0	35.0	50.0	45.7	63.0	47.0	63.0
is NNW of	61.7	64.1	63.5	64.1	44.0	40.0	44.0	40.0	51.1	49.0	51.8	49.0

made no other changes to the code released in [44]. We trained and evaluated the two base BERT encoders for nested spatial entity extraction thanks to our nested annotation format and then separately trained and evaluated the same two base encoders for relation extraction. During training, the models had access to the gold entity annotations. We performed two different evaluations on the relation models: one with the gold entity annotations and one with predicted entities. The relations predicted from gold entity annotations give solely an evaluation of the relation extraction process. The relations predicted from predicted entities give an evaluation of the end-to-end entity and relation extraction process. For each configuration, we trained and evaluated five individual models using different seed values and calculated the arithmetic mean and the standard deviation of the micro F1-scores obtained.

4 RESULTS AND ANALYSIS

The F1-scores for the three tasks with varying context window sizes are displayed in table 5 and table 3 shows the overall precision,

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix ent: <http://data.shom.fr/id/spatialentity/> .
@prefix atln: <http://data.shom.fr/def/atlantis#> .

ent:0001 rdf:type atln:FishingPort ; # entity number 1 is a fishing
      port
    atln:isENEof ent:0002 . # entity number 1 is ENE of entity number 2

ent:0002 rdf:type atln:Cape ; # entity number 2 is a cape
    rdfs:label "Ras Magroua" . # entity number 2 is called "Ras
         Magroua"
```

Figure 3: RDF triples constructed from the information annotated in figure 2 according to the ontological model presented in figure 1.

```
{ "doc_key": "d6_example_sentence",
  "dataset": "atlantis",
  "sentences": [["Un", "port", "de", "pêche", "est", "
      établi", "à", "5,7", "M", "à", "l'", "ENE", "de",
       "Ras", "Magroua", "."]],
  "ner": [[[1, 3, "geogFeat"], [13, 13, "geogFeat"], [1
      4, 14, "name"], [13, 14, "geogName"]]],
  "relations": [[[1, 3, 13, 14, "isENEof"]]] }
```

Figure 4: One line from a JSONL file formatted as required for PURE. It contains the text and the annotations illustrated in figure 2.

Table 4: Values of hyperparameters used for all experiments. The learning rate is the learning rate for BERT encoder parameters and the task learning rate is the learning rate for the classifier head after the encoder.

Hyperparameter	Entity Model	Relation Model
learning rate	1e-5	2e-5
task learning rate	5e-4	-
train batch size	16	32
training epochs	100	10

recall and F1-scores per label. Our experiments show that PURE [44] is capable of extracting nested spatial entities, and that it can do so via nested annotations. This dispenses with the need for joint labelling as in [1], which is more costly than producing nested annotations due to the additional pre-processing of the annotated dataset and post-processing of the predictions.

For entity extraction our experiments show that making cross-sentence context available during training and prediction improves micro F1-scores for both models, and that the multilingual BERT model slightly outperforms the monolingual French BERT model for all context window sizes, with its highest mean micro F1-score being 92.3 when $W = 200$ (table 5). We attribute this contrast in results compared to those in the literature reviewed in section 2 to a characterising feature of our dataset: although the main language of the text is French, it contains words from a large number of other languages. The words in question are primarily geographic features that are part of geographic names, meaning that they must be identified and correctly labelled by the entity model. The monolingual model loses its advantage over the multilingual model in these cases, as the multilingual model is able to understand the semantic meaning of a larger proportion of the words in the dataset.

For relation extraction and end-to-end extraction our experiments show that the monolingual French BERT model slightly outperforms the multilingual BERT model for all context window sizes, with its highest mean micro F1-scores being 64.2 and 63.9 respectively when $W = 0$ (table 5), which means that the monolingual model performs better at relation prediction whether provided with perfect or imperfect entity labels. The monolingual French BERT model achieves higher precision scores for relation extraction and end-to-end extraction than the multilingual BERT model, but the inverse is true of the recall scores (table 3). These results reflect the fact that relations are always expressed in French in the dataset, and sometimes require intricate semantic information to be understood. Taking a closer look at the results for the individual relation labels, we can see that the "is an element of" and the "is marked by" labels have overall lower results than many of the relation labels that involve the cardinal directions. This may be explained by the numerous ways in which these two relations are expressed in our corpus, in comparison with all the other relations that are always expressed using the same key words. The results for both relation models decrease slightly as the size of the context window increases (table 5). This may be attributed to the fact that all the information that categorises one relation is generally included in one sentence, meaning that cross-sentence context may not contribute useful information. Both models give results that are less stable than for entity extraction. This lack of stability may be attributed to the relatively small number of examples of certain relation types in our dataset.

5 CONCLUSION

We discussed and emphasised the importance of reliable nested spatial entity and spatial relation extraction to the construction of geospatial knowledge graphs or gazetteers from text and the disambiguation of spatial entities. We introduced a new annotated French-language dataset for these two extraction tasks, specific to the maritime domain. We provided benchmark results for our own dataset and thereby demonstrated that PURE [44], an existing approach for generic entity and binary relation extraction from text, can be used to extract *nested* entities. This was achieved by training a BERT encoder with nested annotations, without using joint labelling. We also showed that PURE is a suitable baseline approach for the extraction of domain-specific *spatial* entities and *spatial* relations. Our results reveal that the multilingual BERT model slightly outperforms the monolingual French BERT model for entity extraction, with a mean micro F1-score of 92.3, whilst for relation extraction and end-to-end entity and relation extraction the monolingual French BERT model performs slightly better, with mean micro F1-scores of 64.2 and 63.9 respectively. Our results show that making cross-sentence context information available during training and prediction favours entity extraction but hinders relation extraction.

Table 5: Mean micro F1-score with standard deviation over five runs for varying context window sizes for: entity extraction [ent], relation extraction [rel] from gold entity annotations, and end-to-end entity and relation extraction [e2e] from best predicted entity annotations ($W = 248$ for monolingual, $W = 200$ for multilingual). For each task, the highest F1-score over all context window sizes for each base encoder is in bold, and the overall highest F1-score over all context window sizes and both base encoders is underlined.

Task	Base Encoder	$W = 0$	$W = 50$	$W = 100$	$W = 150$	$W = 200$	$W = 248$
[ent]	Monolingual French BERT	91.1 ± 0.3	92.1 ± 0.2	91.9 ± 0.2	91.9 ± 0.2	92.0 ± 0.2	**92.2** ± 0.2
	Multilingual BERT	91.9 ± 0.2	92.3 ± 0.3	92.3 ± 0.2	92.2 ± 0.2	**92.3** ± 0.2	92.3 ± 0.2
[rel]	Monolingual French BERT	**64.2** ± 2.2	64.2 ± 1.4	63.7 ± 0.7	-	-	-
	Multilingual BERT	**63.2** ± 1.0	63.0 ± 1.7	62.9 ± 0.7	-	-	-
[e2e]	Monolingual French BERT	**63.9** ± 2.2	63.8 ± 1.4	63.6 ± 0.7	-	-	-
	Multilingual BERT	**63.2** ± 1.2	63.1 ± 1.7	62.9 ± 0.8	-	-	-

We hope to improve upon the end-to-end extraction results by combining the training of a multilingual BERT model for the entity extraction task with that of a monolingual French BERT model for the relation extraction task. Future work also includes extending this baseline approach to the extraction of values from the text to add properties to entities and relations, the extraction of n-ary relations, and the addition of class-based weights in the loss function to penalise impossible label combinations. Finally, we would like to apply our work to corpora from other fields, such as the humanities, and written in other languages.

ACKNOWLEDGMENTS

This work was co-financed by the Shom and the IGN and is being carried out at the LASTIG, a research unit at Université Gustave Eiffel. This work was granted access to the HPC resources of IDRIS under the allocation 2022-AD011013699 made by GENCI.

REFERENCES

[1] Ankit Agrawal, Sarsij Tripathi, Manu Vardhan, Vikas Sihag, Gaurav Choudhary, and Nicola Dragoni. 2022. BERT-Based Transfer-Learning Approach for Nested Named-Entity Recognition Using Joint Labeling. *Applied Sciences* 12, 3 (2022), 976. https://doi.org/10.3390/app12030976

[2] Jeffrey Beall. 2010. Geographical research and the problem of variant place names in digitized books and other full-text resources. *Library Collections, Acquisitions, & Technical Services* 34, 2-3 (2010), 74–82. https://doi.org/10.1080/14649055.2010.10766263

[3] Cillian Berragan, Alex Singleton, Alessia Calafiore, and Jeremy Morley. 2023. Transformer based named entity recognition for place name extraction from unstructured text. *International Journal of Geographical Information Science* 37, 4 (2023), 747–766. https://doi.org/10.1080/13658816.2022.2133125

[4] Lucie Cadorel, Alicia Blanchi, and Andrea G. B. Tettamanzi. 2021. Geospatial Knowledge in Housing Advertisements: Capturing and Extracting Spatial Information from Text. In *Proceedings of the 11th on Knowledge Capture Conference*. Association for Computing Machinery, USA, 41–48. https://doi.org/10.1145/3460210.3493547

[5] Hao Chen, Maria Vasardani, Stephan Winter, and Martin Tomko. 2018. A Graph Database Model for Knowledge Extracted from Place Descriptions. *International Journal of Geo-Information* 7, 6 (2018), 221. https://doi.org/10.3390/ijgi7060221

[6] Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 3255–3265. https://doi.org/10.18653/v1/2020.findings-emnlp.292

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1. Association for Computational Linguistics, Minneapolis, MN, USA, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[8] Vincent Ducatteeuw. 2021. Developing an Urban Gazetteer: A Semantic Web Database for Humanities Data. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities (GeoHumanities'21)*. Association for Computing Machinery, New York, NY, USA, 36–39. https://doi.org/10.1145/3486187.3490204

[9] Ignatius Ezeani, Paul Rayson, and Ian Gregory. 2023. Extracting Imprecise Geographical and Temporal References from Journey Narratives. In *Proceedings of Text2Story — Sixth Workshop on Narrative Extraction From Texts*, Vol. 3370. CEUR Workshop Proceedings, Dublin, Ireland. https://ceur-ws.org/Vol-3370/paper11.pdf

[10] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, José Emilio Labra Gayo, Sabrina Kirrane, Sebastian Neumaier, Axel Polleres, Roberto Navigli, Axel-Cyrille Ngonga Ngomo, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge Graphs. *Comput. Surveys* 54, 4 (2021), 1–37. https://doi.org/10.1145/3447772

[11] Xuke Hu, Zhiyong Zhou, Yeran Sun, Jens Kersten, Friederike Klan, Hongchao Fan, and Matti Wiegmann. 2022. GazPNE2: A General Place Name Extractor for Microblogs Fusing Gazetteers and Pretrained Transformer Models. *IEEE Internet of Things Journal* 9, 17 (2022), 16259–16271. https://doi.org/10.1109/JIOT.2022.3150967 Conference Name: IEEE Internet of Things Journal.

[12] Yingjie Hu, Chengbin Deng, and Zhou Zhou. 2019. A Semantic and Sentiment Analysis on Online Neighborhood Reviews for Understanding the Perceptions of People toward Their Living Environments. *Annals of the American Association of Geographers* 109, 4 (2019), 1052–1073. https://doi.org/10.1080/24694452.2018.1535886

[13] Krzysztof Janowicz, Pascal Hitzler, Wenwen Li, Dean Rehberger, Mark Schildhauer, Rui Zhu, Cogan Shimizu, Colby K. Fisher, Ling Cai, Gengchen Mai, Joseph Zalewski, Lu Zhou, Shirly Stephen, Seila Gonzalez, Bryce Mecum, Anna Lopez-Carr, Andrew Schroeder, David Smith, Dawn Wright, Sizhe Wang, Yuanyuan Tian, Zilong Liu, Meilin Shi, Anthony D'Onofrio, Zhining Gu, and Kitty Currier. 2022. Know, Know Where, KnowWhereGraph: A densely connected, cross-domain knowledge graph and geo-enrichment service stack for applications in environmental intelligence. *AI Magazine* 43, 1 (2022), 30–39. https://doi.org/10.1002/aaai.12043

[14] Diego Jiménez–Badillo, Patricia Murrieta–Flores, Bruno Martins, Ian Gregory, Mariana Favila-Vázquez, and Raquel Liceras-Garrido. 2020. Developing Geographically Oriented NLP Approaches to Sixteenth–Century Historical Documents: Digging into Early Colonial Mexico. *Digital Humanities Quarterly* 14, 4 (2020).

[15] Meizhi Ju, Makoto Miwa, and Sophia Ananiadou. 2018. A Neural Layered Model for Nested Named Entity Recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, LA, USA, 1446–1459. https://doi.org/10.18653/v1/N18-1131

[16] Junchul Kim, Maria Vasardani, and Stephan Winter. 2015. Harvesting large corpora for generating place graphs, Vol. 12. Santa Fe, NM, USA.

[17] Eleni Kogkitsidou and Philippe Gambette. 2020. Normalisation of 16th and 17th century texts in French and geographical named entity recognition. In *Proceedings of the 4th ACM SIGSPATIAL Workshop on Geospatial Humanities (GeoHumanities'20)*. Association for Computing Machinery, New York, NY, USA, 28–34. https://doi.org/10.1145/3423337.3429437

[18] Caroline Koudoro-Parfait, Gaël Lejeune, and Glenn Roe. 2021. Spatial Named Entity Recognition in Literary Texts: What is the Influence of OCR Noise?. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities (GeoHumanities'21)*. Association for Computing Machinery, New

York, NY, USA, 13–21. https://doi.org/10.1145/3486187.3490206
[19] Stephen C. Levinson. 1996. Language and Space. *Annual Review of Anthropology* 25 (1996), 353–382.
[20] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering* 34, 1 (2020), 50–70. https://doi.org/10.1109/TKDE.2020.2981314
[21] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7203–7219. https://doi.org/10.48550/arXiv.1911.03894
[22] Fernando Melo and Bruno Martins. 2017. Automated Geocoding of Textual Documents: A Survey of Current Approaches. *Transactions in GIS* 21, 1 (2017), 3–38. https://doi.org/10.1111/tgis.12212
[23] Ludovic Moncla. 2015. *Automatic reconstruction of itineraries from descriptive texts*. PhD. Université de Pau et des Pays de l'Adour, Pau.
[24] David Nadeau and Satoshi Sekine. 2007. A Survey of Named Entity Recognition and Classification. *Lingvisticæ Investigationes* 30, 1 (2007), 3–26. https://doi.org/10.1075/li.30.1.03nad
[25] Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2021. Named Entity Recognition and Relation Extraction: State-of-the-Art. *Comput. Surveys* 54, 1 (2021), 20:1–20:39. https://doi.org/10.1145/3445965
[26] Pierre-Henri Paris, Nathalie Abadie, and Carmen Brando. 2017. Linking Spatial Named Entities to the Web of Data for Geographical Analysis of Historical Texts. *Journal of Map & Geography Libraries* 13, 1 (2017), 82–110. https://doi.org/10.1080/15420353.2017.1307306
[27] Heiko Paulheim. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web* 8, 3 (2017), 489–508. https://doi.org/10.3233/SW-160218
[28] Qinjun Qiu, Zhong Xie, Kai Ma, Zhanlong Chen, and Liufeng Tao. 2022. Spatially oriented convolutional neural network for spatial relation extraction from natural language texts. *Transactions in GIS* 26, 2 (2022), 839–866. https://doi.org/10.1111/tgis.12887
[29] Helen Mair Rawsthorne, Nathalie Abadie, Eric Kergosien, Cécile Duchêne, and Eric Saux. 2022. ATLANTIS : Une ontologie pour représenter les Instructions nautiques. In *Journées Francophones d'Ingénierie des Connaissances (IC) Plate-Forme Intelligence Artificielle (PFIA 2022)*. Saint-Étienne, France, 154–163. https://hal.archives-ouvertes.fr/hal-03695242
[30] Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 3118–3135. https://doi.org/10.18653/v1/2021.acl-long.243
[31] Shom. 2021. *Instructions nautiques. D6 : Mer Méditerranée, côtes d'Afrique et du Levant [Version à jour au 13 octobre 2021]*. Brest, France.
[32] Humphrey Southall, Ruth Mostern, and Merrick Lex Berman. 2011. On historical gazetteers. *International Journal of Humanities and Arts Computing* 5, 2 (2011), 127–145. https://doi.org/10.3366/ijhac.2011.0028
[33] Claus Stadler, Jens Lehmann, Konrad Höffner, and Sören Auer. 2012. LinkedGeoData: A core for a web of spatial open data. *Semantic Web* 3, 4 (2012), 333–354.
[34] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Avignon, France, 102–107.
[35] Fabian M. Suchanek. 2014. Information Extraction for Ontology Learning. In *Perspectives on Ontology Learning*, Johanna Völker and Jens Lehmann (Eds.). Studies on the Semantic Web, Vol. 18. IOS Press, 135–151.
[36] Liufeng Tao, Zhong Xie, Dexin Xu, Kai Ma, Qinjun Qiu, Shengyong Pan, and Bo Huang. 2022. Geographic Named Entity Recognition by Employing Natural Language Processing and an Improved BERT Model. *ISPRS International Journal of Geo-Information* 11, 12 (2022), 598. https://doi.org/10.3390/ijgi11120598
[37] Huy Quoc To, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen, and Anh Gia-Tuan Nguyen. 2021. Monolingual vs multilingual BERTology for Vietnamese extractive multi-document summarization. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*. Association for Computational Lingustics, Shanghai, China, 692–699. https://aclanthology.org/2021.paclic-1.73
[38] Solenn Tual, Nathalie Abadie, Joseph Chazalon, Bertrand Duménieu, and Edwin Carlinet. 2023. A Benchmark of Nested Named Entity Recognition Approaches in Historical Structured Documents. In *Proceedings of the 17th International Conference on Document Analysis and Recognition*. Springer, San José, CA, USA. https://hal.science/hal-03994759
[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc.
[40] Abhishek Velankar, Hrushikesh Patil, and Raviraj Joshi. 2022. Mono vs Multilingual BERT for Hate Speech Detection and Text Classification: A Case Study in Marathi. In *Artificial Neural Networks in Pattern Recognition (Lecture Notes in Computer Science, Vol. 13739)*, Neamat El Gayar, Edmondo Trentin, Mirco Ravanelli, and Hazem Abbas (Eds.). Springer, Dubai, UAE, 121–128. https://doi.org/10.1007/978-3-031-20650-4_10
[41] Kehan Wu, Xueying Zhang, Yulong Dang, and Peng Ye. 2022. Deep learning models for spatial relation extraction in text. *Geo-spatial Information Science* 26, 1 (2022), 58–70. https://doi.org/10.1080/10095020.2022.2076619
[42] Vikas Yadav and Steven Bethard. 2018. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2145–2158. https://aclanthology.org/C18-1182
[43] Jiannan Yang, Hong Jia, and Hanbing Liu. 2022. Spatial Relationship Extraction of Geographic Entities Based on BERT Model. In *Journal of Physics: Conference Series*, Vol. 2363. IOP Publishing, Beijing, China. https://doi.org/10.1088/1742-6596/2363/1/012031 Article number: 012031.
[44] Zexuan Zhong and Danqi Chen. 2021. A Frustratingly Easy Approach for Entity and Relation Extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 50–61. https://doi.org/10.18653/v1/2021.naacl-main.5