Extraction automatique d'entités spatiales imbriquées et de relations spatiales à partir de texte pour la création de graphes de connaissances : Une approche et deux jeux de données

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

Extraction automatique d'entités spatiales imbriquées et de relations spatiales à partir de texte pour la création de graphes de connaissances : Une approche et deux jeux de données

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

*LASTIG, Univ Gustave Eiffel, IGN-ENSG, 73 avenue de Paris, F-94165 Saint-Mandé, France
**GERiiCO, Université de Lille, F-59000 Villeneuve d'Ascq, France
***LASTIG, Univ Gustave Eiffel, IGN-ENSG, F-77420 Champs-sur-Marne, France
****IRENav, École navale, F-29240 Brest, France

Résumé. L'extraction automatique d'informations géographiques à partir de texte est essentielle pour exploiter l'ensemble des connaissances spatiales qui n'existent que sous cette forme non structurée. Les éléments clés sont les entités spatiales, leurs types et les relations spatiales entre elles. Structurées en graphe de connaissances géospatial, les connaissances spatiales ambiguës peuvent être désambiguïsées, ce qui facilite considérablement leur accessibilité et réutilisation. Nous présentons une approche pour l'extraction d'entités spatiales imbriquées et de relations spatiales binaires à partir de texte, un jeu de données annoté en français sur le domaine maritime qui peut être utilisé pour entraîner des algorithmes pour les deux tâches d'extraction, ainsi que des résultats de référence pour les deux tâches effectuées individuellement et de bout en bout. Notre approche applique le *Princeton University Relation Extraction system* (PURE), conçu pour l'extraction d'entités génériques plates et de relations binaires génériques, à l'extraction d'entités spatiales imbriquées et de relations binaires spatiales.

1 Introduction

Certaines connaissances spatiales, actuelles ou historiques, n'existent que sous forme de texte. Les guides de voyage, les documents historiques et les publications sur les réseaux sociaux sont quelques exemples de sources de connaissances spatiales non structurées. Les sources textuelles contiennent des connaissances spatiales naturellement hétérogènes : elles peuvent être écrites par différents auteurs, en utilisant un vocabulaire différent, à partir d'un point de vue différent. Elles peuvent par ailleurs couvrir des zones géographiques larges et diverses et contenir des niveaux de détail variés (Ezeani et al., 2023; Jiménez–Badillo et al., 2020; Hu et al., 2019; Kim et al., 2015; Beall, 2010). Pour toutes ces raisons il est difficile d'intégrer dans les modèles de systèmes d'information géographique (SIG) l'information géographique provenant de sources textuelles. L'hypothèse du monde ouvert des technologies du

Web sémantique induit que les graphes de connaissances sont une meilleure solution pour modéliser et stocker les connaissances géographiques extraites de textes hétérogènes, incomplets et imparfaits en langage naturel (Janowicz et al., 2022; Chen et al., 2018; Melo et Martins, 2017; Stadler et al., 2012). Structurées en graphe de connaissances géospatial, les connaissances spatiales ambiguës peuvent être désambiguïsées et liées formellement à des ressources géographiques de référence (telles que DBpedia [1] ou BD TOPO® [2]), ce qui les enrichit de références spatiales directes lorsque c'est possible et facilite considérablement leur accessibilité et réutilisation (Janowicz et al., 2022; Melo et Martins, 2017).

Pendant la population d'un graphe de connaissances, les entités spatiales deviennent des instances de classes ontologiques et les relations spatiales deviennent des propriétés d'objets. Afin de pouvoir attribuer une entité spatiale à sa classe ontologique correspondante, il faut connaître son *type*. Nous faisons l'hypothèse que le nom géographique d'une entité géographique contient souvent un nom commun qui indique son type, tel que *port* dans « Port de Marseille ». Bien que cette hypothèse soit valable pour de nombreuses langues romanes, elle ne s'applique pas à toutes les langues.

L'extraction d'entités spatiales *plates* vise à identifier simplement le nom d'une entité spatiale tel que « Port de Marseille » sans chercher à le définir plus finement, tandis que l'extraction d'entités spatiales *imbriquées* permet de définir plusieurs niveaux d'étiquettes pour la même section de texte. Nous utilisons les étiquettes introduites par Moncla (2015), dont les définitions sont les suivantes :

— geographic feature pour les noms communs qui identifient un *type* d'entité spatiale ;
— name pour les noms propres pures ;
— geographic name pour les noms complets associés à une entité spatiale.

Si nous reprenons notre exemple, l'extraction d'entités spatiales imbriquées viserait donc à identifier « Port de Marseille » comme geographic name, « Port » comme geographic feature et « Marseille » comme name. En extrayant des entités spatiales *imbriquées* au lieu de *plates*, le type de geographic feature de l'entité est déjà connu et une instance de la classe ontologique correspondante peut être créée automatiquement. Dans certains cas, comme dans notre exemple, le name donne une indication de la position géographique de l'entité. Ces deux informations supplémentaires, le type de geographic feature de l'entité et son name, facilitent la tâche de désambiguïsation qui lie l'instance à son entée correspondante dans une ressource géographique de référence (Southall et al., 2011).

En extrayant les relations spatiales entre entités, des assertions de propriétés d'objets peuvent être automatiquement créées entre instances. Ces informations peuvent également être utilisées pour faciliter la désambiguïsation des entités nommées et non nommées, et augmenter la fiabilité des résultats grâce au raisonnement spatial (Paris et al., 2017). Dans le cas où une entité de référence n'existe pas encore, une nouvelle entrée peut être créée dans la ressource géographique, appuyée par les informations relatives aux classes et aux propriétés de l'instance. Le même raisonnement s'applique à la création et à l'enrichissement de gazetiers : en identifiant spécifiquement les types d'entités pendant le processus d'extraction, les entrées de gazetiers peuvent être automatiquement classées ou dotées d'attributs et ainsi être plus facilement désambiguïsées. L'identification et l'extraction des relations spatiales dans lesquelles les entités

1. DBpedia (`https://www.dbpedia.org/`) est une ressource géographique de référence mondiale.
2. La BD TOPO® (`https://geoservices.ign.fr/bdtopo`) est une ressource géographique de référence pour le territoire français.

spatiales prennent part contribue à augmenter le niveau de détail dans les descriptions et les positions géographiques des entrées de gazetiers.

Les textes qui couvrent un environnement international sont susceptibles de contenir des noms géographiques en langues autre que la langue principale du texte. Il est important que cela n'empêche pas le processus d'extraction : les noms géographiques et les types d'entités écrits en autres langues doivent tout de même être identifiés. L'état de l'art en extraction d'information à partir de texte repose sur des modèles de langage à base de réseaux de neurones profonds (Nasar et al., 2021). L'état de l'art en matière d'extraction d'informations à partir de textes repose sur des modèles linguistiques de réseaux neuronaux profonds. Ces modèles peuvent être entraînés dans une ou plusieurs langues et sont appelés respectivement modèles de langage pré-entraînés *monolingues* ou *multilingues*. Une ontologie multilingue peut alors être utilisée pour faciliter la désambiguïsation des entités dont le type est écrit dans d'autres langues (Stadler et al., 2012).

Dans cet article nous présentons une approche pour l'extraction automatique d'entités imbriquées et de relations binaires de texte, qui peut être appliquée aux entités et aux relations *spatiales* si nécessaire (Rawsthorne et al., 2023). Notre approche est une adaptation du *Princeton University Relation Extraction system* (PURE) existant (Zhong et Chen, 2021), qui a été conçu pour l'extraction d'entités génériques plates et de relations binaires génériques. Elle s'insère dans la méthodologie *ATlantis Ontology and kNowledge base development from Texts and Experts* (ATONTE) pour la construction semi-automatique de graphes de connaissances, géospatiaux ou non, à partir de sources textuelles hétérogènes, des connaissances d'experts et des données de référence (Rawsthorne, 2024). Notre approche d'extraction peut être appliquée à des corpus dans n'importe quelle langue (à condition qu'ils contiennent des noms d'entités imbriqués) couvrant n'importe quel domaine, qu'il soit scientifique ou littéraire, historique ou contemporain, de fiction ou non.

Cet article est une version raccourcie et traduite d'un article en anglais par Rawsthorne et al. (2023). Ce dernier contient notamment une présentation complète de l'état de l'art de l'extraction d'entités et de relations à partir de texte en utilisant des approches en apprentissage profond.

2 Contexte d'application

Les travaux présentés dans cet article ont été réalisés dans le cadre d'un projet qui vise à structurer en graphe de connaissances géospatial les informations géographiques contenues dans les *Instructions nautiques* (Rawsthorne, 2024). Les *Instructions nautiques* sont une série d'ouvrages PDF rédigés en français qui sont produites et publiées par le Service hydrographique et océanographique de la Marine (Shom). Elles décrivent l'environnement maritime côtier et donnent des instructions de navigation côtière.

Les *Instructions nautiques* font partie d'une gamme de produits diffusés par le Shom qui servent à la planification d'itinéraires de navigation maritime. D'autres ouvrages du Shom, plus spécialisés, viennent compléter les connaissances sur l'environnement côtier et la navigation, parmi lesquels on trouve *Feux et signaux de brume*, *Radiosignaux*, *Courants de marée* ainsi que l'*Annuaire des marées*. Ils apportent des renseignements qui sont nécessaires à la préparation

d'un itinéraire adapté et sûr. Le type de navire, l'expérience du navigateur, la temporalité [3], les conditions météorologiques et les conditions océanographiques sont également à prendre en considération lors de la planification. Les *Instructions nautiques* contiennent principalement trois types de renseignements (Shom, 2020) :

1. elles donnent des informations complémentaires à celles qui sont affichées sur les cartes marines comme les caractéristiques physiques (couleur, forme, taille, etc.) d'un amer [4] ;
2. elles recensent les informations absentes des cartes marines telles que le climat typique de la zone décrite ;
3. elles donnent des instructions ou des informations à propos de la navigation telles que les routes conseillées, les conditions d'accès aux ports ou encore les réglementations en place.

Les *Instructions nautiques* sont divisées en plusieurs volumes, un par zone de couverture. Une zone de couverture peut être définie soit comme une section de trait de côte entre deux positions sur la côte, soit comme l'ensemble du trait de côte d'une île ou d'un ensemble d'îles. Chaque volume commence avec un chapitre de renseignements généraux. Le plan général du reste de l'ouvrage suit linéairement le trait de côte, chaque chapitre étant dédié à une section du trait de côte. En lisant un chapitre, le lecteur a l'impression d'être emmené le long de la côte par le rédacteur ; chaque repère, danger et autre particularité de l'environnement est décrit, et chaque mouillage, accès de port et entrée de chenal est détaillé. Les consignes mentionnent également les spécificités de la météorologie, la courantologie et la réglementation locales. Des photographies montrant les amers et les ports notables sont intercalées dans le texte. Elles illustrent également le positionnement relatif des différentes entités géographiques et doivent conforter le lecteur dans la représentation qu'il se fait de son environnement.

3 Plan de l'article et disponibilité des données

Dans la section 4 nous présentons le jeu de données *coAsTaL mAritime NavigaTion InstructionS* (ATLANTIS) [5], un nouveau jeu de données de référence en langue française, annoté manuellement, qui porte sur le domaine maritime (Rawsthorne et al., 2023). Il peut être utilisé pour entraîner des algorithmes pour l'extraction d'entités spatiales imbriquées et de relations spatiales binaires à partir de texte. Nous présentons également dans la section 4 le jeu de données TextMine'24 [6], qui est un sous-ensemble du jeu de données ATLANTIS et contient des annotations d'entités spatiales imbriquées (Rawsthorne et al., 2024).

Dans la section 5 nous présentons une approche pour l'extraction d'entités spatiales imbriquées et de relations spatiales binaires à partir de texte. Notre approche est une adaptation de PURE (Zhong et Chen, 2021), qui gère uniquement l'extraction d'entités plates génériques et de relations génériques. Nous la rendons applicable à l'extraction d'entités *spatiales* et *imbriquées*, et de relations *spatiales* en utilisant le code mis à disposition par Zhong et Chen (2021) ainsi qu'un format d'annotation modifié.

3. Une temporalité est une condition locale qui est dépendante sur le temps, par exemple l'heure, le mois de l'année ou le saison.

4. Un amer est un « objet remarquable situé à un endroit fixe sur la terre et pouvant être utilisé pour déterminer un emplacement ou une direction. » Traduit du Hydrographic Dictionary Working Group (2019).

5. https://github.com/umrlastig/atlantis-dataset

6. https://www.kaggle.com/competitions/defi-textmine-2024

Nous présentons dans la section 6 des résultats de référence pour le jeu de données ATLANTIS, obtenus grâce à l'approche décrite dans la section 5, pour les tâches d'extraction d'entités spatiales imbriquées, d'extraction de relations spatiales binaires, et d'extraction combinée d'entités et de relations spatiales de bout en bout. Les entités et relations spatiales extraites de notre corpus pourraient directement enrichir des ressources géographiques de référence ou des gazetiers, qui pourraient être utilisés pour des applications dans des domaines tels que l'hydrographie, la navigation maritime ou les humanités.

Enfin, dans la section 7 nous présentons les conclusions de ce travail et nous donnons quelques perspectives.

4 Préparation des jeux de données

4.1 Le jeu de données ATLANTIS

Le le jeu de données ATLANTIS est composé d'extraits de chacun des 15 volumes des *Instructions nautiques* dont nous disposions. Nous avons extrait le texte des PDF en utilisant *pdfminer.six* [7]. Le jeu de données contient 101 400 jetons.

Nous avons annoté notre jeu de données à la main en utilisant le *brat rapid annotation tool* [8]. Il permet la création d'annotations d'étiquettes imbriquées ainsi que la création de liens entre elles, avec un sens et une étiquette propre à chaque lien (Stenetorp et al., 2012). Le schéma d'annotation, décrit ci-dessous, a été concu et validé par l'ensemble des auteur·e·s. Une auteure a réalisé l'annotation du jeu de données et ensuite des extraits ont été vérifiés et validés par tout·e·s les auteur·e·s.

Étant donné que nous souhaitions réaliser l'extraction d'entités spatiales *imbriquées* afin de capturer simultanément le nom complet de l'entité spatiale ainsi que son type et son nom, nous avons mis en œuvre une approche d'étiquetage imbriqué en utilisant les étiquettes définies dans la section 1. Tout jeton peut être annoté de zéro ou d'une étiquette geographic feature ou name. Un jeton ne peut pas être annoté à la fois par une étiquette geographic feature et par une étiquette name. Un jeton ne peut pas être annoté uniquement avec une étiquette name. Un jeton annoté avec une étiquette name doit également être annoté une ou plusieurs fois avec une étiquette geographic name. Tout jeton, déjà étiqueté ou non, peut être annoté zéro ou plusieurs fois avec une étiquette geographic name.

Un très grand nombre de types différents de relations spatiales sont utilisés dans notre corpus. Nous avons donc décidé de nous limiter à l'extraction de ceux qui seraient les plus utiles au cours du processus de désambiguïsation.

Les directions cardinales sont très utilisées dans la navigation parce que les relations spatiales qui les utilisent sont construites en utilisant un cadre de référence absolu, ce qui signifie qu'il n'est pas nécessaire de préciser un point de vue spécifique (Levinson, 1996). Nous avons choisi d'extraire les relations spatiales qui utilisent les directions cardinales en raison de leur utilisation fréquente et de leur absence d'ambiguïté. Cela représente 16 types de relations au total : quatre qui utilisent les directions cardinales (N, E, S, W), quatre qui utilisent les directions intercardinales (NE, SE, SW, NW) et huit qui utilisent les directions intercardinales

7. https://github.com/pdfminer/pdfminer.six
8. http://brat.nlplab.org

secondaires (NNE, ENE, ESE, SSE, SSW, WSW, WNW, NNW). Dans notre corpus, ces relations spatiales sont toujours désignées par ces 16 abréviations à une, deux ou trois lettres, par exemple « le port est au NW de la ville » ou « la tour est à l'ESE du château ». Les 16 étiquettes qui correspondent à ces relations spatiales sont au format « est au XYZ de », ou « XYZ » est une des 16 abréviations des directions cardinales.

Nous avons identifié trois autres types de relations spatiales permettant de capturer plus d'informations sur les entités spatiales spécifiques à notre domaine qui ne sont souvent pas nommées dans le corpus ou qui sont susceptibles d'être absentes des ressources géographiques de référence telles que les marques de navigation (bouées, balises, etc.), les rochers ou encore les bancs de sable. Tout d'abord, l'étiquette « *is off the coast of* » est utilisée lorsqu'il est indiqué qu'une entité spatiale est située au large ou dans les eaux côtières d'une autre. Elle est donc fréquemment utilisée pour localiser des entités spatiales isolées. Ce type de relation spatiale, qui est également construit à partir d'un cadre de référence absolu, est toujours désigné par la même expression dans notre corpus : « est au large de ». Deuxièmement, l'étiquette « *is marked by* » est utilisée pour toute entité spatiale qui est marquée ou signalée par une autre entité délibérément placée, souvent une marque de navigation, soit lorsque la première représente un danger pour la navigation, soit lorsqu'elle permet un passage en toute sécurité : « Son musoir est marqué par un feu. » (Shom, 2021). Cette relation indique une proximité entre les deux entités et est exprimée de différentes manières dans notre corpus : « est marqué par » peut être exprimé alternativement par « est signalé par » ou « est indiqué par ». Troisièmement, l'étiquette « *is an element of* » indique une relation topologique qui inclut des entités situées *dans* ou *sur* d'une autre de manière à ce qu'une vue à vol d'oiseau montre l'empreinte spatiale de l'une comme étant à l'intérieur ou en partie à l'intérieur de l'autre. Cette relation est exprimée de différentes manières dans notre corpus, y compris de manière implicite, et inclut rarement le mot « élément ». Par exemple, « l'île porte un phare », « le feu est établi sur le quai » et « les hauts-fonds de la baie » indiquent tous une relation de type « *is an element of* ».

Toutes les annotations de relations doivent relier deux annotations d'entités, soit des étiquettes geographic feature, soit des étiquettes geographic name. Toutes les annotations de relation doivent avoir une direction. Au lieu de dupliquer les étiquettes de relation pour tenir compte de leurs inverses et de créer des annotations de relations orientées qui vont toujours dans le sens du texte (« A $\rightarrow$ *is marked by* $\rightarrow$ B » et « C $\rightarrow$ *marks* $\rightarrow$ D »), nous avons créé une version pour chaque étiquette et nous autorisons les annotations de relations orientées qui vont dans l'une ou l'autre direction (« A $\rightarrow$ *is marked by* $\rightarrow$ B » et « C $\leftarrow$ *is marked by* $\leftarrow$ D »).

La figure 1 montre une phrase des *Instructions nautiques* annotée selon notre schéma d'annotation d'entités spatiales imbriquées et de relations spatiales binaires. L'association de l'étiquetage spécifique du geographic feature « *ras* » (« cap ») au sein du geographic name et des valeurs d'étiquettes multilingues dans une ontologie signifie que cette entité spatiale pourrait être automatiquement instanciée dans la bonne classe, quelle que soit la langue dans laquelle la geographic feature est écrite (dans ce cas l'arabe romanisé). La figure 2 montre des triplets *Resource Description Framework* (RDF) qui pourraient être automatiquement construits à partir de cette phrase selon l'ontologie ATLANTIS (Rawsthorne et al., 2022).

Nous avons divisé notre jeu de données annoté en trois parties : entraînement, développement et test, en veillant à respecter un rapport de $80 : 10 : 10$ entre le nombre total de jetons et le nombre d'étiquettes d'entités. Nous avons également cherché à ce que du texte couvrant chaque zone géographique soit présent dans les trois parties. Notre jeu de données de 101 400

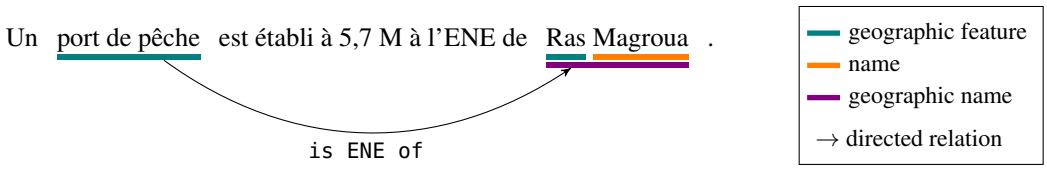

FIG. 1 – *Une phrase des Instructions nautiques annotée selon notre schéma d'annotation d'entités spatiales imbriquées et de relations spatiales binaires (Shom, 2021).*

```
@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 ; # entit  num ro 1 est un port de p che
    atln:isENEof ent:0002 . # entit  num ro 1 est ENE d'entit  num ro 2

ent:0002 rdf:type atln:Cape ; # entit  num ro 2 est un cap
    rdfs:label "Ras Magroua" . # entit  num ro 2 s'appelle Ras Magroua
```

FIG. 2 – *Triplets RDF construits à partir de l'information annotée dans la figure 1 selon l'ontologie ATLANTIS (Rawsthorne et al., 2022).*

jetons contient 16 777 étiquettes d'entités (qui peuvent s'étendre sur plusieurs jetons) et 3 051 étiquettes de relations (qui relient exactement deux étiquettes d'entités dans un sens donné). Au total, 18 030 jetons sont annotés avec au moins une étiquette d'entité, ce qui correspond à près d'un jeton sur cinq. La composition du jeu de données et affiché dans le tableau 1. La distribution des étiquettes peut être consulté sur le dépôt de notre jeu de données [9].

	Entraînement	Développement	Test	Total
Jetons	83 851	8 156	9 393	101 400
Jetons sans étiquette	69 200	6 507	7 663	83 370
Jetons avec étiquette d'entité	14 651	1 649	1 730	18 030
Étiquettes d'entité	13 582	1 476	1 719	16 777
Étiquettes de relation	2 507	222	322	3 051

TAB. 1 – *Nombre de jetons et d'étiquettes dans chaque partie du jeu de données ATLANTIS. Une étiquette d'entité peut s'étendre sur un jeton ou plus.*

Nous avons converti notre jeu de données du format *brat* au format *JSON Lines* (JSONL) [10] requis pour PURE à l'aide d'un script Python [11]. La figure 3 montre la même phrase annotée que dans la figure 1 convertie en valeur JSON dans ce format. Dans notre cas, chaque valeur

9. https://github.com/umrlastig/atlantis-dataset
10. Un fichier JSONL contient une valeur JSON valide sur chaque ligne.
11. https://github.com/dwadden/dygiepp/blob/master/scripts/new-dataset/brat_to_input.py

JSON correspond à un paragraphe des *Instructions nautiques* et contient une liste de phrases (dont chacune est une liste de jetons), une liste des annotations d'entités (les numéros des deux jetons frontaliers + l'étiquette), et une liste des annotations de relations (les numéros des deux fois deux jetons frontaliers + l'étiquette). Ce format d'annotation imbriquée, qui permet à tout jeton d'être annoté avec zéro ou plusieurs étiquettes, rend possible l'extraction d'entités imbriquées sans utiliser du *joint labelling* (Rawsthorne et al., 2023; Agrawal et al., 2022).

```
{ "doc_key": "d6_phrase_exemple",
  "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"], [14, 14, "name"], [13, 14
      , "geogName"]]],
  "relations": [[[1, 3, 13, 14, "isENEof"]]] }
```

FIG. 3 – *Une ligne d'un fichier JSONL formaté pour PURE. Elle contient le texte et les annotations illustrés dans la figure 1.*

4.2 Le jeu de données TextMine'24

Le jeu de données TextMine'24 est un sous-ensemble du jeu de données ATLANTIS. Il est composé de 66 030 jetons au total, et possède la même couverture géographique mondiale que le jeu de données ATLANTIS. Le jeu de données TextMine'24 contient uniquement des annotations d'entités spatiales, et ce dans la limite de deux niveaux d'imbrication. Il a été utilisé en tant que jeu de données de référence pour le Défi TextMine 2024 [12], un défi de reconnaissance d'entités spatiales organisé en collaboration avec le groupe de travail TextMine [13]. Une description complète du jeu de données TextMine'24 et le défi associé se trouve dans l'article par Rawsthorne et al. (2024).

5 Entraînement des modèles

PURE (Zhong et Chen, 2021) entraîne indépendamment deux encodeurs de base à partir de modèles de langage profonds pré-entraînés existants : l'un pour identifier et étiqueter les entités, et l'autre pour identifier les paires reliées d'entités et classer la relation entre elles. Nous appelons le premier modèle le *modèle entité* et le second le *modèle relation*. PURE permet également de réguler la taille de la fenêtre contextuelle W, c'est-à-dire la quantité de contexte inter-phrases qui est mise à la disposition du modèle. Le contexte disponible lors du traitement d'une phrase donnée s'étend de $(W-n)/2$ mots à gauche de la phrase à $(W-n)/2$ mots à droite, où n est le nombre de mots dans la phrase. Une perte d'entropie croisée est utilisée pour les deux modèles.

12. https://www.kaggle.com/competitions/defi-textmine-2024
13. https://textmine.sciencesconf.org/

Pour les encodeurs de base, nous avons utilisé *bert-base-french-europeana-cased* comme modèle *Bidirectional Encoder Representations from Transformers* (BERT) monolingue français et *bert-base-multilingual-cased* comme modèle BERT multilingue. Nous avons utilisé les hyperparamètres par défaut fournis par (Zhong et Chen, 2021) et nous avons fait des expériences avec plusieurs tailles de fenêtres contextuelles, dans les plages des valeurs par défaut, pour le modèle entité et le modèle relation. Pour le modèle entité, nous avons utilisé des fenêtres contextuelles de 0, 50, 100, 150, 200 et 248 ($W = 250$ a dépassé le mémoire GPU disponible) et pour le modèle relation, nous avons utilisé des fenêtres contextuelles de 0, 50 et 100. Nous n'avons apporté aucune autre modification au code publié par Zhong et Chen (2021). Nous avons entraîné et évalué les deux encodeurs de base BERT pour l'extraction d'entités spatiales imbriquées grâce à notre format d'annotation imbriquée, puis nous avons entraîné et évalué séparément les deux mêmes encodeurs de base pour l'extraction de relations spatiales. Pendant l'entraînement, les modèles ont eu accès aux annotations manuelles (*gold*) des entités. Nous avons effectué deux évaluations différentes sur les modèles de relations : l'une avec les annotations d'entités manuelles et l'autre avec les entités prédites. Les relations prédites à partir des annotations d'entités manuelles donnent uniquement une évaluation du processus d'extraction de relations. Les relations prédites à partir des entités prédites donnent une évaluation du processus d'extraction d'entités et de relations de bout en bout. Pour chaque configuration, nous avons entraîné et évalué cinq modèles individuels en utilisant différentes valeurs de graines aléatoires et nous avons calculé la moyenne arithmétique et l'écart-type des scores F1 micro obtenus.

6 Résultats et analyse

Les scores F1 pour les trois tâches avec différentes tailles de fenêtres contextuelles sont affichés dans le tableau 2. La précision globale, le rappel et les scores F1 par étiquette peut être consulté sur le dépôt de notre jeu de données [14].

Pour l'extraction d'entités, nos expériences montrent que la mise à disposition du contexte inter-phrases pendant l'entraînement et la prédiction améliore les scores F1 micro pour les deux modèles, et que le modèle multilingue BERT surpasse légèrement le modèle monolingue français BERT pour toutes les tailles de fenêtre contextuelle, avec son score F1 micro moyen le plus élevé étant de 92,3 lorsque $W = 200$ (tableau 2). Nous attribuons ce contraste dans les résultats par rapport à ceux de la littérature (Rawsthorne et al., 2023) à une caractéristique de notre jeu de données : bien que la langue principale du texte soit le français, il contient des mots provenant d'un grand nombre d'autres langues. Les mots en question sont principalement des geographic feature qui font partie de geographic name, ce qui signifie qu'ils doivent être identifiés et correctement étiquetés par le modèle entité. Dans ces cas, le modèle monolingue perd son avantage par rapport au modèle multilingue, car ce dernier est capable de comprendre le sens sémantique d'une plus grande proportion des mots de l'ensemble de données.

14. https://github.com/umrlastig/atlantis-dataset

W	Entités		Relations		Relations (de bout en bout)	
	Mono.	Multi.	Mono.	Multi.	Mono.	Multi.
0	91.1 ± 0.3	91.9 ± 0.2	**64.2** ± 2.2	**63.2** ± 1.0	**63.9** ± 2.2	**63.2** ± 1.2
50	92.1 ± 0.2	92.3 ± 0.3	64.2 ± 1.4	63.0 ± 1.7	63.8 ± 1.4	63.1 ± 1.7
100	91.9 ± 0.2	92.3 ± 0.2	63.7 ± 0.7	62.9 ± 0.7	63.6 ± 0.7	62.9 ± 0.8
150	91.9 ± 0.2	92.2 ± 0.2	-	-	-	-
200	92.0 ± 0.2	**92.3** ± 0.2	-	-	-	-
248	**92.2** ± 0.2	92.3 ± 0.2	-	-	-	-

TAB. 2 – *Score F1 micro moyen avec écart-type sur cinq exécutions pour différentes tailles de fenêtres contextuelles pour : l'extraction d'entités, l'extraction de relations à partir d'annotations manuelles (gold) d'entités, et l'extraction d'entités et de relations de bout en bout [e2e] à partir des meilleures annotations d'entités prédites ($W = 248$ pour le modèle monolingue et $W = 200$ pour le modèle multilingue). Pour chaque tâche, le score F1 le plus élevé parmi toutes les tailles de fenêtres contextuelles pour chaque encodeur de base est en* ***gras****, et le score F1 global le plus élevé parmi toutes les tailles de fenêtres contextuelles et les deux encodeurs de base est souligné.*

7 Conclusion et perspectives

Nous avons discuté et souligné l'importance d'une extraction fiable d'entités spatiales imbriquées et de relations spatiales pour la construction de graphes de connaissances géospatiales ou de gazetiers à partir de textes et pour la désambiguïsation d'entités spatiales. Nous avons présenté un nouveau jeu de données annoté en langue française pour ces deux tâches d'extraction, spécifique au domaine maritime. Nous avons fourni des résultats de référence pour notre propre jeu de données et avons ainsi démontré que PURE (Zhong et Chen, 2021), une approche existante pour l'extraction générique d'entités et de relations binaires à partir de textes, peut être utilisée pour extraire des entités *imbriquées*. Ceci a été réalisé en entraînant un encodeur BERT avec des annotations imbriquées, sans utiliser du *joint labelling*. Nous avons également montré que PURE est une approche de base adaptée à l'extraction d'entités *spatiales* et de relations *spatiales* spécifiques à un domaine. Nos résultats révèlent que le modèle multilingue BERT est légèrement plus performant que le modèle monolingue français BERT pour l'extraction d'entités, avec un score F1 micro moyen de 92,3, tandis que pour l'extraction de relations et l'extraction d'entités et de relations de bout en bout, le modèle monolingue français BERT est légèrement plus performant, avec des scores F1 micro moyens de 64,2 et 63,9 respectivement. Nos résultats montrent que la mise à disposition d'informations contextuelles inter-phrases lors de l'entraînement et de la prédiction favorise l'extraction d'entités mais entrave l'extraction de relations.

Une première perspective concerne la réduction du temps nécessaire pour annoter le jeu de données d'entraînement. Pour ce faire, on pourrait utiliser un outil d'annotation qui propose soit un étiquetage assisté par apprentissage machine, soit une annotation automatisée en fonction d'un ensemble de mots clés défini par l'utilisateur. Il faudrait évaluer et comparer les annotations produites à l'aide des différents outils afin de vérifier que la qualité ne diminue pas par rapport aux annotations que nous avons réalisées de manière entièrement manuelle. Dans le

but d'augmenter la fiabilité du jeu de données annoté, on pourrait adopter une approche d'annotation à plusieurs et calculer l'accord inter-annotateurs. On pourrait aussi produire automatiquement des données synthétiques sous la forme d'un texte annoté par le biais d'expressions régulières. Il serait également possible de combiner un apprentissage non supervisé avec un apprentissage supervisé en utilisant un plus petit jeu de données annoté manuellement lors de l'étape d'extraction. Si les résultats sont identiques ou meilleurs que ceux obtenus avec notre approche actuelle, l'apprentissage non supervisé pourrait être intégré à notre méthodologie et le volume de données annotées manuellement nécessaire pourrait être réduit. Une autre perspective concerne l'élargissement de l'approche d'extraction d'information à partir de texte afin de permettre la prise en compte de plusieurs types d'entités simultanément et la gestion de relations n-aires (ternaires ou plus). Ceci permettrait d'extraire des relations impliquant plus de deux entités telles que la relation spatiale *entre*, comme par exemple dans la phrase « la bouée est entre l'île Est et l'île Ouest ».

Références

Agrawal, A., S. Tripathi, M. Vardhan, V. Sihag, G. Choudhary, et N. Dragoni (2022). BERT-Based Transfer-Learning Approach for Nested Named-Entity Recognition Using Joint Labeling. *Applied Sciences 12*(3), 976.

Beall, J. (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), 74–82.

Chen, H., M. Vasardani, S. Winter, et M. Tomko (2018). A Graph Database Model for Knowledge Extracted from Place Descriptions. *International Journal of Geo-Information 7*(6), 221.

Ezeani, I., P. Rayson, et I. Gregory (2023). Extracting Imprecise Geographical and Temporal References from Journey Narratives. In *Proceedings of Text2Story — Sixth Workshop on Narrative Extraction From Texts*, Volume 3370, Dublin, Ireland. CEUR Workshop Proceedings.

Hu, Y., C. Deng, et Z. 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), 1052–1073.

Hydrographic Dictionary Working Group (2019). S-32 IHO Hydrographic Dictionary.

Janowicz, K., P. Hitzler, W. Li, D. Rehberger, M. Schildhauer, R. Zhu, C. Shimizu, C. K. Fisher, L. Cai, G. Mai, J. Zalewski, L. Zhou, S. Stephen, S. Gonzalez, B. Mecum, A. Lopez-Carr, A. Schroeder, D. Smith, D. Wright, S. Wang, Y. Tian, Z. Liu, M. Shi, A. D'Onofrio, Z. Gu, et K. 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), 30–39.

Jiménez–Badillo, D., P. Murrieta–Flores, B. Martins, I. Gregory, M. Favila-Vázquez, et R. Liceras-Garrido (2020). Developing Geographically Oriented NLP Approaches to Sixteenth–Century Historical Documents : Digging into Early Colonial Mexico. *Digital Humanities Quarterly 14*(4).

Kim, J., M. Vasardani, et S. Winter (2015). Harvesting large corpora for generating place graphs. Volume 12, Santa Fe, NM, USA.

Levinson, S. C. (1996). Language and Space. *Annual Review of Anthropology 25*, 353–382.

Melo, F. et B. Martins (2017). Automated Geocoding of Textual Documents : A Survey of Current Approaches. *Transactions in GIS 21*(1), 3–38.

Moncla, L. (2015). *Automatic reconstruction of itineraries from descriptive texts*. PhD, Université de Pau et des Pays de l'Adour, Pau, France.

Nasar, Z., S. W. Jaffry, et M. K. Malik (2021). Named Entity Recognition and Relation Extraction : State-of-the-Art. *ACM Computing Surveys 54*(1), 20 :1–20 :39.

Paris, P.-H., N. Abadie, et C. Brando (2017). Linking Spatial Named Entities to the Web of Data for Geographical Analysis of Historical Texts. *Journal of Map & Geography Libraries 13*(1), 82–110.

Rawsthorne, H. M. (2024). *Creation of Geospatial Knowledge Graphs From Heterogeneous Sources*. PhD, Université Gustave Eiffel, Champs-sur-Marne, France.

Rawsthorne, H. M., N. Abadie, A. Guille, P. Cuxac, et C. Lopez (2024). Défi TextMine'24 : Reconnaissance d'entités géographiques dans un corpus des Instructions nautiques. In *Actes de l'atelier TextMine'24*, Dijon, France, pp. 87–92.

Rawsthorne, H. M., N. Abadie, E. Kergosien, C. Duchêne, et 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, pp. 154–163.

Rawsthorne, H. M., N. Abadie, E. Kergosien, C. Duchêne, et 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*, Hamburg, Germany, pp. 21–30. Association for Computing Machinery.

Shom (2020). Guide de rédaction des Instructions Nautiques du Shom. Procédure spécifique, Shom.

Shom (2021). *Instructions nautiques. D6 : Mer Méditerranée, côtes d'Afrique et du Levant [Version à jour au 13 octobre 2021]*. Brest, France.

Southall, H., R. Mostern, et M. L. Berman (2011). On historical gazetteers. *International Journal of Humanities and Arts Computing 5*(2), 127–145.

Stadler, C., J. Lehmann, K. Höffner, et S. Auer (2012). LinkedGeoData : A core for a web of spatial open data. *Semantic Web 3*(4), 333–354.

Stenetorp, P., S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, et J. 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*, Avignon, France, pp. 102–107. Association for Computational Linguistics.

Zhong, Z. et D. 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*, Online, pp. 50–61. Association for Computational Linguistics.

Summary

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. We present 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.