Défi TextMine'24 : Reconnaissance d'entités géographiques dans un corpus des Instructions nautiques

Helen Mair Rawsthorne, Nathalie Abadie, Adrien Guille, Pascal Cuxac, Cédric Lopez

Défi TextMine'24 : Reconnaissance d'entités géographiques dans un corpus des Instructions nautiques

Helen Mair Rawsthorne*, Nathalie Abadie*,
Adrien Guille**, Pascal Cuxac***, Cédric Lopez ****

* LASTIG, Univ Gustave Eiffel, IGN-ENSG, 73 avenue de Paris, F-94165 Saint-Mandé, France
**Université de Lyon, Lyon 2, UR 3083
adrien.guille@univ-lyon2.fr
***INIST-CNRS, UAR76, Nancy, France
pascal.cuxac@inist.fr
****Emvista, Jacou, France
cedric.lopez@emvista.com,

Résumé. Pour cette deuxième édition du défi TextMine, l'Institut national de l'information géographique et forestière (IGN) et le Service hydrographique et océanographique de la Marine (Shom) proposent de relever le défi de la reconnaissance d'entités géographiques nommées et non nommées à deux niveaux dans les *Instructions nautiques*, une série d'ouvrages publiée par le Shom. Neuf équipes académiques et industrielles ont participé à ce défi, les résultats de 5 de ces équipes, présentés lors de l'atelier, sont résumés dans cet article.

1 Introduction

Le 21 octobre 2022, l'association Extraction et Gestion des Connaissances (EGC) a lancé le groupe de travail TextMine [1]. Dans le cadre de ce groupe de travail, un objectif est de confronter l'état de l'art scientifique aux problèmes de fouille de texte (*text mining* en anglais) rencontrés par des industriels. Sous la forme de défis, le groupe de travail propose des jeux de données inédits et les partage avec la communauté scientifique. Le premier défi du groupe de travail TextMine a été lancé le 21 octobre 2022 en étroite collaboration avec la société Emvista, éditrice de logiciels fondés sur des technologies du Traitement Automatique du Langage Naturel, qui a fourni une partie des données. Cinq équipes ont partagé leurs expériences lors de l'atelier TextMine'23 qui s'est tenu à Lyon [2].

Pour cette deuxième édition, l'Institut national de l'information géographique et forestière (IGN) et le Service hydrographique et océanographique de la Marine (Shom) proposent de relever le défi de la reconnaissance d'entités géographiques nommées et non nommées à deux niveaux dans les *Instructions nautiques*, une série d'ouvrages publiée par le Shom. Chaque volume décrit l'environnement maritime côtier d'une zone géographique et donne aux navigateurs les informations nécessaires pour naviguer près des côtes et accéder aux ports en sécu-

1. https://textmine.sciencesconf.org/
2. https://textmine.sciencesconf.org/resource/page/id/4

rité. Les zones géographiques couvertes par les *Instructions nautiques* sont réparties autour du monde entier.

Le défi a été lancé le 24 avril 2023 et clôturé le 16 novembre 2023. Neuf équipes académiques et industrielles ont participé, avec un total de 202 soumissions de résultats sur la plateforme Kaggle hébergeant le défi [3]. Ce défi a donné lieu à 7 articles publiés : un article hors compétition de la part de l'instigateur du défi (Rawsthorne et al., 2024), cinq articles de la part des participants (CIAD, CRIT, OctopusMind/IRISA, Inist-CNRS, CIRAD/INRAE), et un article de synthèse de la part des organisateurs du défi (le présent article). Chaque article des participants est associé au code source qui permet de reproduire les modèles développés.

2 Présentation du jeu de données

Ce défi est inspiré par un projet qui vise à structurer en graphe de connaissances géospatial les informations géographiques contenues dans les *Instructions nautiques* (Rawsthorne, 2024). Une partie du corpus utilisé dans ce projet, publié par Rawsthorne et al. (2023), constitue le corpus de ce défi. Ce dernier est constitué d'extraits de 15 volumes des *Instructions nautiques*, annotés selon 3 étiquettes. Il compte au total 66030 jetons et 18537 étiquettes, dont une partie est proposée pour l'apprentissage et l'autre réservée pour le test. Les 3 étiquettes utilisées suivent la proposition de Moncla (2015) pour annoter des entités géographiques imbriquées :

- geogFeat : Pour les noms communs qui identifient une caractéristique géographique.
  - Par exemple : "À 8 M à l'ENE du phare de Nadji, le port de pêche de Sidi Abderrahmane (36° 29,7' N — 1° 05,7' E) est construit au bord du village de Soug el Bgar (pointe Rouge)." (Extrait de Shom (2021))
- name : Pour les noms propres purs.
  - Par exemple : "À 8 M à l'ENE du phare de Nadji, le port de pêche de Sidi Abderrahmane (36° 29,7' N — 1° 05,7' E) est construit au bord du village de Soug el Bgar (pointe Rouge)."
- geogName : Pour le nom associé à une caractéristique géographique.
  - Par exemple : "À 8 M à l'ENE du phare de Nadji, le port de pêche de Sidi Abderrahmane (36° 29,7' N — 1° 05,7' E) est construit au bord du village de Soug el Bgar (pointe Rouge)."

Chaque jeton peut recevoir entre 0 et 2 étiquettes selon les combinaisons indiquées dans la première colonne du tableau 1. Ce tableau donne le nombre d'annotations de chaque étiquette ou combinaison d'étiquettes présentes dans le jeu d'entraînement.

3 Expériences

Dans cette section nous synthétisons les expériences réalisées par chaque équipe ayant participé au défi.

- L'équipe CIAD (Armary et al., 2024) a expérimenté deux approches. D'une part, plusieurs modèles ont été affinés : BERT-Base-NER (un modèle BERT entraîne pour la tâche de NER (*Named Entity Recognition*) en anglais), CamemBERT-NER (un modèle BERT français entraîné pour la tâche de NER en français), TinyBERT (un réajustement

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

Étiquette(s)	Nombre d'annotations dans le jeu d'entraînement
geogFeat	4167
geogFeat geogName	1469
geogName	4490
name	2118
name geogName	2123
jeton sans étiquette	32668

TAB. 1 – Distribution des cinq étiquettes ou combinaisons possibles d'étiquettes dans le jeu d'entraînement, ainsi que le nombre de jetons sans étiquette dans ce dernier.

du modèle BERT), et un modèle de NER utilisant GPT2 avec un apprentissage basé sur des invites. D'autre part, une approche basée sur un GCN (*Graph Convolutional Network*) (Kipf et Welling, 2017) considère que les jetons dans du jeu de données sont des nœuds du graphe et que ces nœuds sont connectés en fonction de leur co-occurrence dans le contexte. Dans cette approche, les étiquettes sont également vues comme des nœuds et un jeton est connecté à son étiquette grâce à la connaissance du jeu d'entraînement. L'approche basée sur le GCN obtient un score d'exactitude faible (77,9%) en comparaison à l'approche Bert-Base-NER qui atteint 92,8%.

— L'équipe CRIT (Gutehrlé, 2024) a utilisé le modèle fr_core_news_lg distribué par spaCy [4] pour obtenir les parties du discours, les catégories grammaticales et les fonctions syntaxiques des jetons. Trois modèles CRF (*Conditional Random Fields*) (Lafferty et al., 2001) linéaires ont été entraînés dans l'objectif d'identifier quelles caractéristiques (formes, catégories grammaticales, fonctions syntaxiques, etc.) permettraient d'obtenir les meilleurs résultats. Il a été conclu que les performances des modèles diminuent dans le cas où les catégories grammaticales et les fonctions syntaxiques sont considérées comme caractéristiques dans l'entraînement. Le meilleur F-score obtenu est 96,4%.

— L'équipe OctopusMind/IRISA (Ahmia et al., 2024) a également expérimenté des modèles à base de CRF et de Transformers. Concernant les CRF : onze caractéristiques dont la forme du terme, son lemme, la présence de lettres dans le terme, la présence de ponctuation et les catégories grammaticales ont été utilisées. L'équipe a également utilisé des caractéristiques à longue portée qui modélisent le contexte d'occurrence des termes (les variables voisines d'un terme). L'optimisation des poids du CRF a été réalisée avec la méthode L-BFGS (*Limited-memory Broyden-Fletcher-Goldfarb-Shanno*) (Zhu et al., 1997). Par ailleurs, le modèle CamemBERT a été utilisé en ajoutant une couche linéaire dense en sortie. Un affinage a été réalisé sur les données TextMine. Contrairement à Gutehrlé (2024), les résultats montrent que le modèle obtient les meilleurs résultats (F-score de 98%) lorsque toutes les caractéristiques sont prises en compte. L'utilisation de caractéristiques à longue portée améliore le score à condition que des contraintes concernant la portée soient respectées.

— L'équipe Inist-CNRS (Anki et al., 2024) utilise un réseau de neurones récurrents (RNN) en utilisant le framework Flair (Akbik et al., 2019). Un embedding contextuel de dimen-

4. https://spacy.io/

sion 300 a été utilisé par défaut dans les expériences, contrairement aux expériences précédentes qui utilisent des embeddings de transformers. Le RNN mis en place est constitué de trois couches : une couche de reprojection de l'embedding et deux couches de type LSTM. Les paramètres par défaut de Flair ont été utilisés. Le meilleur F-score obtenu sur le jeu de test public est de 97,9%.

— L'équipe CIRAD/INRAE (Decoupes et al., 2024) a expérimenté deux approches. La première consiste en un réentrainement d'un modèle en français de la librairie spaCy (fr_core_news_lg), identique à celui utilisé par CRIT. La seconde approche consiste à réajuster des modèles pré-entrainés de type Transformers : deux modèles en langue française (CamemBERT-base et CamemBERT-large) et deux modèles multi-langues (XLM-RoBERTa). CamemBERT-large obtient les meilleurs résultats (sans optimisation des hyperparamètres), soit 97,6%.

4 Résultats

Le classement final des participants a été déterminé selon deux jeux de test : un jeu de test "public" couvrant la zone géographique présente dans le jeu d'entraînement (la côte Ouest de l'Afrique, la mer Méditerranée, la mer Rouge, Saint-Pierre-et-Miquelon, les Antilles françaises, la France hexagonale et les îles de l'océan Indien) et un jeu de test "privé" couvrant une zone géographique absente du jeu d'entraînement et du jeu de test public : les îles de l'océan Pacifique. Le F-score public a été calculé avec 41% du jeu de test. Les participants ont pu évaluer leurs modèles sur ce jeu de test public. Le jeu de test privé (59%) a été utilisé pour tester la version définitive du modèle de chaque participant. Le F-score moyen est la moyenne harmonique des F-score publics et privés.

Rang	Équipe	Soumissions	F-score public	F-score privé	F-score moy.
1	OctopusMind	16	**0.980**	**0.979**	**0,980**
2	Tetis	8	0.976	0.978	0,977
3	Inist	31	0.979	0.972	0,976
4	CRIT	18	0.964	0.943	0,953
5	Valentin Nyzam	29	0.975	0.974	0,974
6	CIAD	87	0.928	0.913	0,920
7	Nicolas Fouqué	5	0.954	0.951	0,952
8	Ben & Loïck	3	0.915	0.880	0,898
9	Skyens	5	0.842	0.834	0,838

TAB. 2 – Résultats obtenus sur les jeux de test public et privé.

Nous invitons les lecteurs à consulter les articles relatifs à chaque expérience (cf. section 3) afin d'obtenir plus de détails sur les résultats obtenus par chaque équipe.

Références

Ahmia, O., D. Cao, N. Béchet, et P.-F. Marteau (2024). OctopusMind @ Défi TextMine'24 Reconnaissance d'entités géographiques dans un corpus d'Instructions nautiques. In *Actes de*

l'atelier TextMine'24. Conférence Extraction et Gestion des Connaissances 2024 (EGC'24), Dijon, France.

Akbik, A., T. Bergmann, D. Blythe, K. Rasul, S. Schweter, et R. Vollgraf (2019). FLAIR : An easy-to-use framework for state-of-the-art NLP. In W. Ammar, A. Louis, et N. Mostafazadeh (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, Minneapolis, Minnesota, pp. 54–59. Association for Computational Linguistics.

Anki, L., L. Gaillard, et J. Revol (2024). Détection d'entités nommées géographiques par réseau de neurones récurrents. In *Actes de l'atelier TextMine'24*. Conférence Extraction et Gestion des Connaissances 2024 (EGC'24), Dijon, France.

Armary, P., C.-B. El-Vaigh, O. Labbani Narsis, et C. Nicolle (2024). CIAD System for Geographical Entity Detection at TextMine'24. In *Actes de l'atelier TextMine'24*. Conférence Extraction et Gestion des Connaissances 2024 (EGC'24), Dijon, France.

Decoupes, R., R. Interdonato, R. Kafando, M. Roche, S. Mehtab Alam, M. Teisseire, et S. Valentin (2024). TETIS @ Challenge TextMine 2024 : "Reconnaissance d'entités géographiques dans un corpus des Instructions nautiques". In *Actes de l'atelier TextMine'24*. Conférence Extraction et Gestion des Connaissances 2024 (EGC'24), Dijon, France.

Gutehrlé, N. (2024). Défi TextMine 2024 : "Reconnaissance d'entités géographiques dans un corpus des Instructions nautiques" - soumission équipe CRIT. In *Actes de l'atelier TextMine'24*. Conférence Extraction et Gestion des Connaissances 2024 (EGC'24), Dijon, France.

Kipf, T. N. et M. Welling (2017). Semi-supervised classification with graph convolu- tional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Lafferty, J. D., A. McCallum, et F. C. N. Pereira (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, San Francisco, CA, USA, pp. 282–289. Morgan Kaufmann Publishers Inc.

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

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, E. Kergosien, C. Duchêne, et É. 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, 31st International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL 2023)*, Hamburg, Germany, pp. 21–30. Association for Computing Machinery, New York, NY, United States.

Rawsthorne, H. M., N. Abadie, E. Kergosien, C. Duchêne, et É. Saux (2024). 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. In *Actes de l'atelier TextMine'24*. Conférence Extraction et Gestion des Connaissances 2024 (EGC'24), Dijon, France.

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

Zhu, C., R. H. Byrd, P. Lu, et J. Nocedal (1997). Algorithm 778 : L-bfgs-b : Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw. 23*(4), 550–560.

Summary

For this second edition of the TextMine challenge, the National Institute of Geographic and Forestry Information (IGN) and the Hydrographic and Oceanographic Service of the Navy (Shom) propose to take up the challenge of recognizing named and unnamed geographical entities at two levels in the *Nautical Instructions*, a series of works published by the Shom. Nine academic and industrial teams participated in this challenge, the results of 5 of these teams, presented during the workshop, are summarized in this article.