Mesurer les risques de discrimination dans une tâche de détection d'entités nommées

Hugues de Mazancourt
CTO, Datapolitics
hugues@datapolitics.fr

Flavie Loshouarn
Lead developper, Datapolitics
flavie@datapolitics.fr

Alice Bruguier
NLP engineer, Datapolitics
alice@datapolitics.fr

Abstract

La reconnaissance d'entités nommées (NER) est une tâche classique en NLP, parfois sujette à des biais liés au corpus d'apprentissage. Nous présentons ici l'adaptation d'une méthode de détection de biais et son évaluation sur deux tâches de reconnaissance de noms de personnes à dans les chaînes d'analyse de textes politiques.

1 Introduction

Créée en 2021, Datapolitics est une start-up ayant pour but de produire une intelligence data liée à la vie démocratique française. Elle fournit à ses clients un outil de veille et d'analyse. Cette analyse est fondée sur des données structurées issus en partie de la reconnaissance des locuteurs dans la presse et les procès verbaux des mairies. Le locuteur est celui qui émet une parole dans l'espace publique. Pouvoir le reconnaître permet de suivre l'évolution de la position d'un homme ou d'une femme politique et d'être au plus près des opinions de la personne. Il permet aussi de pondérer les opinions détectées en fonction de la position de celui qui émet l'opinion.

2 Motivation

Datapolitics a été fondé dans un objectif d'aide à la démocratie. Son but est de rendre la vie politique française accessible à tous, permettant une meilleure représentativité et une meilleure capacité d'expression des causes et des personnes. L'invisibilisation de groupe d'appartenance est un problème social contre lequel Datapolitics veut lutter. Il nous a donc semblé majeur d'évaluer nos propres modèles de langue sous cet angle et de nous assurer de l'absence de biais discriminatoire dans notre traitement des données. Il s'agit avant tous de pouvoir certifier que Datapolitics n'introduit, ni n'aggrave des inégalités sociales.

Dans cette étude, nous évaluerons l'importance de l'origine du nom dans la qualité de sa détection. Est-ce que nous détectons mieux des noms de personnes françaises que de ressortissants d'autre nationalité ?

3 Contexte de l'étude

Notre détection de locuteur s'effectue sur deux contextes différents, conduisant à deux tâches de difficultés distinctes en termes de NLP. La première consiste à détecter les locuteurs dans un article de presse (*qui exprime l'opinion ?*) et la seconde à détecter les auteurs de prise de parole dans des procès verbaux de communautés territoriales (mairies, conseils d'arrondissement, etc.). Pour chacune de ces deux tâches (nommées **presse** et **mairie**, resp.), nous avons constitué des corpus manuellement annotés pour identifier le locuteur. L'apprentissage s'est ensuite effectué avec l'outil SpaCy (ref).

Il faut noter que ces corpus, exclusivement en français, sont relativement réguliers, et qu'ils mentionnent les noms de personne essentiellement sous la forme *Prénom Nom*, *Nom Prénom* ou *M/Mme Nom*, ce qui facilite la tâche d'évaluation.

4 Méthode d'évaluation

Nous avons constitué des corpus synthétiques à partir de données non-issues de l'apprentissage. Nous avons ensuite remplacé les noms identifiés manuellement comme *LOCUTEUR* dans notre corpus de test par d'autres noms de différentes origines afin de valider l'impact sur la reconnaissance automatique. Il s'agit d'une extension de la méthode présentée en [1] qui se limitait aux prénoms.

4.1 Sélection des noms

Nous avons choisi les pays d'origine en fonction de l'importance de leur immigration vers la France. Nos choix ont porté sur :

- En Afrique du nord, le Maroc, l'Algérie, la Tunisie, la Syrie.
- En Afrique subsaharienne, le Congo, le Mali, la Cote d'Ivoire, le Cameroun.
- En Asie occidentale, la Turquie, l'Arménie, le Liban.
- En Europe, le Royaume Uni, la Russie, l'Italie, le Portugal.
- En Amérique Latine, Haiti, le Venezuela, le Brésil.
- En Asie, le Vietnam, la Chine, le Sri Lanka.

Les noms servant à représenter ces pays ont été pris sur Wikipédia. Ils correspondent aux noms des élus des différentes chambres et aux noms de personnes publiques célèbres. Une liste semblable a été constituée selon la même méthode pour la France. Nous avons ainsi constitué une liste de **100** noms pour chacun des **22** pays, dont on trouvera un extrait ci-dessous. La liste exhaustive est publiée en open-source sur **GitHub**[1].

- **Maroc** : Aymen Benabderrahmane, Abdelmadjid Tebboune, Brahim Merad', Ahmed Attaf, Abderrachid Tabbi, Laaziz Fayed, Mohamed Arkab, Laid Rebiga, ...
- **Turquie** : Kâzım Karabekir, Fethi Okyar, 'Celâl Bayar, Ekrem Alican, Ragıp Gümüşpala, Süleyman Demirel, Bülent Ecevit, ...
- **Mali** : Modibo Keïta, Yoro Diakité, Mamadou Dembelé, Younoussi Touré, Abdoulaye Sékou Sow, Ibrahim Boubacar Keïta, Mandé Sidibé, ...
- **Vietnam** : Đng Th Minh Hnh, Nguyn Thùy Lâm, Trn Th Hng Giang, Võ Hoàng Yn, Thuy Diep, Thien LE, Chloe Dao, Chau Bui, ...
- **Chine** : Hong Cheong, Feng Xuemin, Fu Bingchang, He Chengyao, Lang Jingshan, Li Zhensheng, Miao Xiaochun, Stephen Chow, ...
- **Royaume-Uni** : Stephen Kinnock, Robin Millar, Kirsty Blackman, Stephen Flynn, Neil Gray, Leo Docherty, Wendy Morton, Graham Brady, ...

[1] https://github.com/datapolitics/ethics

4.2 Critères d'évaluation

Pour chacun des pays (y compris la France), nous avons remplacé aléatoirement les noms de locuteurs identifiés manuellement par un nom pris au hasard dans la liste. Nous avons ensuite validé si cette substitution impactait la reconnaissance, en mesurant l'écart de f-mesure par rapport à la substitution effectuée avec des noms de personnalités françaises. On mesure donc l'écart de performance de l'extraction par rapport à la performance de l'extraction sur des données non-connues et non préalablement testées, mais proches du corpus d'apprentissage.

5 Résultats

La table 1 reprend les résultats sur la tâche **mairie**, la table 2 ceux sur la tâche **presse**. Les plus mauvais résultats sont en gras. Les résultat montrent clairement qu'il n'existe pas différence notable entre la détection des noms de personnes de nationalité française et des noms de personnes d'autres nationalités. 9 pays sur 22 ont même des scores supérieurs à la France et sur l'ensemble de l'évaluation il n'y a que deux cas où le delta est inférieur à 1% ce qui reste insignifiant. On peut également remarquer que la difficulté de la tâche a plutôt tendance à gommer les différences entre pays, la tâche **presse** étant intrinsèquement plus complexe que la tâche **mairie**.

6 Conclusion

Nous avons mis en œuvre une méthode innovante pour mesurer les biais de détection de noms dans des corpus politiques et l'avons appliquée à des tâches majeures de la chaîne de traitement Datapolitics. Les résultats montrent que les analyseurs ne sont pas biaisés par le corpus d'entraînement et montrent des performances similaires quelles que soient les origines des noms mentionnés. Avec la publication en open-source de notre liste de référence, la méthode pourra être utilisée pour les évaluations de tâches dans le domaine de l'analyse de textes politiques en général.

References

[1] Shubhanshu Mishra, Sijun He, and Luca Belli. Assessing demographic bias in named entity recognition, 2020.

Pays	f-mesure	delta
France (témoin)	0.9974	
Maroc	0.9976	0.02%
Algérie	0.9986	0.12%
Tunisie	0.9987	0.12%
Syrie	0.9845	**-1.30**%
Congo	0.9974	-0.01%
Mali	0.9973	-0.01%
Cote d'Ivoire	0.9934	-0.40%
Cameroun	0.9974	-0.01%
Turquie	0.9940	-0.34%
Arménie	0.9983	0.08%
Liban	0.9830	**-1.45**%
Royaume Uni	0.9988	0.14%
Russie	0.9984	0.10%
Italie	0.9989	0.14%
Portugal	0.9970	-0.05%
Haiti	0.9993	0.19%
Venezuela	0.9987	0.13%
Brésil	0.9933	-0.42%
Vietnam	0.9902	-0.72%
Chine	0.9957	-0.17%
Sri Lanka	0.9972	-0.02%

Table 1: Résultats de l'évaluation sur la tâche "mairie"

Pays	f-mesure	delta
France (témoin)	0.8560	
Maroc	0.8569	0.10%
Algérie	0.8572	0.14%
Tunisie	0.8599	0.45%
Syrie	0.8555	-0.07%
Congo	0.8540	-0.24%
Mali	0.8543	-0.21%
Cote d'Ivoire	0.8581	0.24%
Cameroun	0.8546	-0.17%
Turquie	0.8537	-0.28%
Arménie	0.8552	-0.10%
Liban	0.8607	0.55%
Royaume Uni	0.8607	0.55%
Russie	0.8593	0.38%
Italie	0.8593	0.38%
Portugal	0.8534	**-0.31%**
Haiti	0.8540	-0.24%
Venezuela	0.8584	0.27%
Brésil	0.8540	-0.24%
Vietnam	0.8534	-0.31%
Chine	0.8546	-0.17%
Sri Lanka	0.8534	**-0.31%**

Table 2: Résultats de l'évaluation sur la tâche "presse"