Rapport du projet :

Reconnaisance des entités nommées

UE : TextMining

Lamiaa Snoussi *et Azhar* Taoual

Master 2 : Big Data 2019/2020

Table des matières

1 Introduction

La reconnaissance d'entité nommée est une sous-tâche d'extraction d'informations qui cherche à localiser et à classer l'entité nommée mentionnée dans un texte non structuré en catégories prédéfinies telles que les noms de personnes, les organisations, les emplacements, les codes médicaux, les expressions temporelles, les quantités, les valeurs monétaires, les pourcentages, etc. .

L'objectif de notre travail est de produire une annotation qui prend qu'une sous-catégorie des entités nommées (EN) contenant un terme du lexique fourni ainsi que les mots liés qui représentent des entités nommées étendue (ENE).

2 Bibliographies

En se basant sur l'article : Structured and Extended Named Entity Evaluation in Automatic Speech Transcriptions *(Olivier Galibert et al., 2010)*
Dans l'article, les auteurs présentent une représentation d'entités nommées structurées et des méthodes pour évaluer la reconnaissance de ces entités nommées structurées, ils fournissent une cartographie entre les éléments de référence et d'hypothèse qui permet d'énumérer les erreurs et de calculer la valeur du taux d'erreur de créneau.
Ces entités nommées étendues et algorithmes d'évaluation ont été utilisés dans l'évaluation des entités nommées des données vocales.
Ce travail est utile à la fois pour l'analyse des erreurs et pour se convaincre de la qualité de la mesure d'évaluation. Il permet également de fusionner toutes les sorties du système dans une évaluation et de rassembler les erreurs pour aider à corriger la référence plus efficacement si nécessaire.

En se basant sur l'article : What's missing in geographical parsing ? *(Milan Gritta et al., 2017)*
Cet article évalue et analyse les performances d'un certain nombre de géoparsers de premier plan sur un certain nombre de corpus et souligne les défis en détail,dans le géoparsing, les noms de lieux contenant les informations géographiques sont appelés toponymes, qui doivent d'abord être identifiés (appelés géolocalisation) et résolus en leurs coordonnées géographiques (appelés géocodage).
Ils fournissent une étude complète et une évaluation critique des géoparsers de pointe avec des ensembles de données hétérogènes.
Ils présentent également WikToR, un nouveau corpus Wikipedia à grande échelle, généré automatiquement et annoté géographiquement dans le but de réduire la pénurie de corpus open source dans la recherche sur le géoparking.
Ils ont évaluer le géoparsing comme un pipeline, puis évaluer chaque étape séparément, les performances de géolocalisation sont mesurées à l'aide du F-Score pour le corpus LGL, Précision pour WikToR.

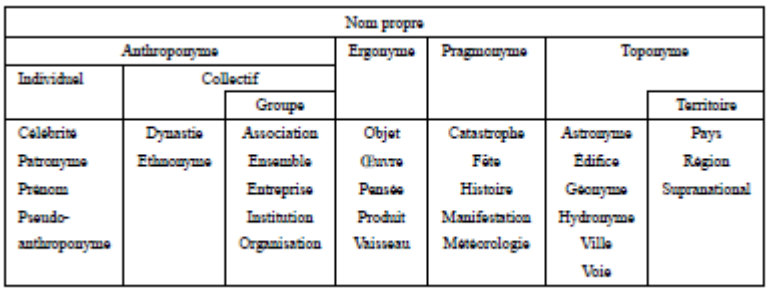

FIGURE 1 – La typologie primaire utilisée

La performance de géocodage est mesurée en utilisant l'ASC et l'erreur médiane pour les deux corpus.

L'article : A Linguistically Grounded Annotation Language for Spatial Information *(Pustejovsky et al., 2012)* nous apprend l'importance de l'utilisation des informations spatiales pour le langage naturel avec la spécification ISO-Space(un langage d'annotation pour l'encodage spatial et informations spatio-temporelles). Cette méthode est utilisée dans le traitement naturel du langage(TALN) pour la reconnaissance d'entité nommée et l'inférence basée sur du texte. Le but de ce projet est d'utiliser les annotations ISO-Space pour la description spatiale faite par une tierce personne par exemple : un cycliste pour aider à identifier une location générale.
Cette approche nous a inspiré à parcourir les fichiers XML donnés en annexe pour extraire des verbes de mouvements ou le nom de lieux afin d'enrichir nos dictionnaires.

L'article : Cascades de transducteurs autour de la reconnaissance des entitées nommées *(Maurel et al., 2011)* a pour but d'utiliser des cascades de transducteurs CasSys avec le logiciel Unitex. Ils ont crée deux cascades, une contenant des informations sur les locuteurs et un autre qui permet de crée le lien entre les entités nommées. Les graphes utilisés recherchent des informations sur des corpus etiquettés. Ces informations permettent de reconnaitre les éléments (Figure 1)[1]. Pour extraire ces informations, ils utilisent cinq catégories de graphes pour reconnaitre les entités, les outils, les masques et les étiqueteurs.
Cette approche nous a inspiré à utiliser plusieurs graphes pour créer une cascade
.

3 Synthèse des évaluations

3.1 Pré-traitement du corpus

Notre approche a été d'utiliser le corpus 2017 pour enrichir nos dictionnaires.

3.1.1 Reconnaissance des verbes de mouvements

En appliquant la bibliothèque de Python "MLconjug" sur le fichier contenant les verbes de mouvements,on a pu obtenir un ficher de sortie contenant tous les verbes de mouvements conjugué dans tous les temps,et avec ce résultat, on a créer un dictionnaire. Aussi,en parcourant les fichiers XML avec un balisage TEI sur Python, on extrait les verbes de mouvements présents dans le corpus 2017 fourni qui ont un type = 'V' afin d'enrichir le dictionnaire précédent.

3.1.2 Reconnaissance des entités de lieu

Avec des bibliothèques pour parcourir des fichiers XML avec un balisage TEI sur Python, on extrait les mots présents dans le corpus 2017 fourni qui sont balisés placeName. Cette approche nous permet d'enrichir les dictionnaires.

3.2 Création des graphes

On a crée huit graphes avec le logiciel Unitex :

Graphe verb : ce graphe reconnaît les verbes de mouvements présents dans le dictionnaire verb-mouv et affiche la variable verbe avec à droite le verbe qu'il reconnaît.

Graphe motCompose : ce graphe identifie les mots composés.

Graphe lieux : ce graphe identifie les lieux qui figurent dans le dictionnaire placename.

Graphe motguillemets : ce graphe identifie les mots entre guillemets.

Graphe Apostrophe : ce graphe identifie les mots contenant une apostrophe comme par exemple le mot **jusqu'au**.

Graphe distance : ce graphe identifie un chiffre suivi d'une unité de mesure de distance.

Graphe auto-ex-lexique-1 : ce graphe identifie tous les mots contenus dans le lexique fourni.

Graphe textmining : ce graphe permet de reconnaitre les entités nommées en se basant sur tous les graphes cités au-dessus.

3.3 Cascade

La création de la cascade se fait avec CasSys qui consiste à appliquer une liste de transducteurs au texte dans un ordre précis. Notre objectif est d'appliquer une cascade avec les graphes que nous avons créer et pouvoir détecter les verbes de mouvements ainsi que les entitées nommées et les entitées nommées étendue.

3.4 Lexiques utilisés

Le lexique utilisé est le lexique qui a été fournie ainsi que le dictionnaire Dela_fr.bin fourni par Unitex.

4 Analyse des résultats

On obtient un fichier texte avec une sortie annotée où chaque phrase contient obligatoirement un verbe de mouvement et un mot du lexique fourni (figure 2 et 3). Les résultats obtenus sont plus intéréssants si on applique le lexique DELAF sinon les transducteurs ne reconnaissent pas certaines prépositions. Aussi, il existe une certaine ambiguïté en utilisant le dictionnaire des verbes de mouvements, le graphe peut prendre un verbe conjugué dans la liste mais si par rapport au sens de la phrase ce mot est un nom et non un verbe par exemple la phrase suivante : Entrée du parc est reconnu comme un verbe.

FIGURE 2 – Le résultat obtenu sur le texte Boucle La Féclaz les fermes avec les entités nommées(EN).

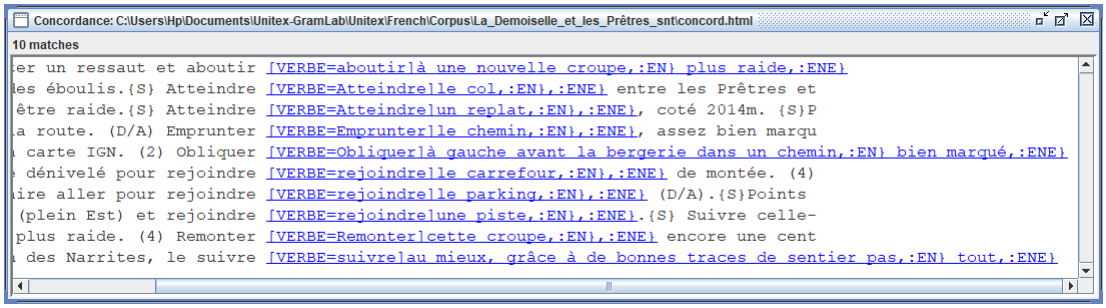

FIGURE 3 – Le résultat obtenu sur le texte La Demoiselle et les Prêtres avec les entités nommées étendues(ENE).

5 Conclusion

Dans le cadre de ce projet, il nous a été demandé de faire un travail de reconnaissance d'entités nommées contenues dans le corpus.
Mais d'après les résultats obtenus, on remarque que le graphe ne peut pas détecter toutes les entités nommées. Ce qui peut être amélioré dans notre travail, c'est le fait d'utiliser des graphes contenant des restrictions pour enlever les ambiguïtés

Références

[1] Denis Maurel, Nathalie Friburger,Jean-Yves Antoine,Iris Eshkol-Taravella,Damien Nouvel,(2011) *Cascades de transducteurs autour de la reconnaissance des entitées nommées*, TAL. Volume 52n1/2011.

[2] Olivier Galibert,Sophie Rosset,Cyril Grouin,Pierre Zweigenbaum,Ludovic Quintard,(2010)*Structured and Extended Named Entity Evaluation in Automatic Speech Transcriptions*

[3] Milan Gritta1, Mohammad Taher Pilehvar1,Nut Limsopatham1,Nigel Collier,(2017) *What's missing in geographical parsing ?*,Springerlink 2017.

[4] James Pustejovsky,Jessica Moszkowicz,Marc Verhagen*A Linguistically Grounded Annotation Language for Spatial Information*,2012.