1.5em

Table des matières

Université des Sciences et de la Technologie Houari Boumediene

Traitement Automatique du Langage Naturel

Rapport mini projet
Développement d'un outil pour l'extraction des entités nommées (Named Entity Extraction) pour la langue arabe

---

Quadrinôme
BENDALI Omar Manil
DJEBRI Mounira
SLIMANI Kenza
SMATI Yasmine

5 janvier 2020

1 Introduction

Le concept d'entité nommée est apparu au milieu des années 90 [1] comme étant une sous-tâche de l'extraction d'information. Elle consiste à identifier certains objets textuels tels que les noms de personne, d'organisation et de lieu.[4]

La reconnaissance d'entités nommées ou REN constitue encore de nos jours un champ de recherche très actif. Toutefois, même si la reconnaissance des Entités Nommées en langue arabe est un prétraitement potentiellement très utile pour de nombreuses applications du traitement du langage naturel, cette problématique reste peu abordée par les chercheurs en raison du manque de ressources informatise dans cette langue et éventuellement car cette tache se heurte à plusieurs défis inhérents aux caractéristiques de cette dernière.

En effet l'absence de la distinction majuscule/minuscule est un obstacle majeur pour la langue arabe. En fait, la majuscule est un indicateur très utile pour identifier les noms propres dans les langues utilisant l'alphabet latin. Il est aussi à noter que la langue arabe est très agglutinative ayant une morphologie dérivationnelle et flexionnelle assez complexe ne facilitant en rien la tache de la REN.

2 Analyse de l'approche utilisées pour l'implémentation du SAIE :

Différentes approches existent pour l'extraction d'entité nommées, l'une d'entre elle est présente au travers du projet SAIE. C'est dans l'optique de d'implémenter le système présente dans la thèse de Fatma Nasser Al Shamsi que nous abordons ce projet.

D'après la thèse qui nous a été fournie[1], l'architecture du SAIE a été séparée en deux parties :
La première consiste à faire une analyse lexicale du texte en entrée, appelée le module NLP (Natural Language Processing). Cette partie comprend 3 étapes :

- Tokenisation : durant cette étape un segmenteur, utilisant l'espace pour séparer les tokens, a été développé afin d'avoir une liste de mots qu'il faudra passer à un POS Tagger.
- Stemmatisation : dans cette étape, un stemmer (basé sur le stemmer de Buckwalter) a été utilisé pour séparer les mots de leurs affixes.
- Pos Tagger : un modèle de langage trigramme a été construit dans le but de construire un POS Tagger, et pour développer le modèle HMM les probabilités des trigrammes ont été utilisées.

Les séquences de mots représentent les observations et les états du HMM sont représentés par les POS Tags auxquels on rajoute Start et End.
Les probabilités ont été obtenues grâce au modèle de langage trigramme. Etant donne qu'un modèle trigramme a été utilisé, alors il peut y avoir des cas où on ne trouve pas certains trigrammes, donc pour éviter d'avoir des probabilités à 0, la technique du back-off a été utilisée, avec un modèle bigramme et un autre unigramme.
L'algorithme Viterbi basé sur les trigrams avec un backoff vers le modèle bi-gramme …a été utilisé pour trouver le POS Tag d'une séquence qui maximise les probabilités.

La deuxième partie, appelée le module d'extraction, qui consiste à faire l'extraction des entités nommées. Un modèle HMM a aussi été utilisé dans cette partie.

Les observations du HMM sont exprimées par les séquences de mots, les probabilités de transition sont obtenues à partir du modèle trigramme contextuel et les probabilités de transmission à partir du modèle trigramme

1. Le concept d'entité nommée est apparu à l'occasion de conférence d'évaluation MUC (Message Understanding Conference).

lexical (trigramme, bigramme et unigramme).
Les états du HMM sont les tags. Deux tags spéciaux ont été rajoutés ; start et end pour délimiter l'observation. L'algorithme de Viterbi a servi à trouver la séquence qui maximise les probabilités.

- Extracteur d'entités nommées :

Concernant la partie d'extraction d'entités nommées, un corpus de news a été stemmé puis pos-tagué et ensuite annoté manuellement selon le tagset des entités nommées. Une partie de l'annotation a automatisé en se basant sur un lexicon. Comme pour le module de POS-Tag, des modèles du langage trigramme avec back-off (Méthode de Katz) bi-gramme et unigramme furent créé et serviront de matrices de transitions et émissions pour l'algorithme de Viterbi.

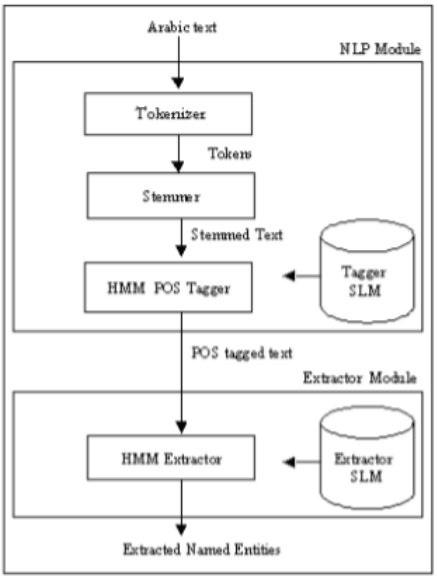

FIGURE 1 – Architecture SAIE

3 Corpus d'apprentissage

Afin d'effectuer l'apprentissage mentionné précédemment, nous devons traiter notre corpus et lui appliquer les transformations nécessaires[7].

3.1 Presentation du Corpus

Le corpus utilisé dans ce projet est : WikiFaneGold. C'est un corpus d'entités nommées arabe qui a été annoté manuellement. On y trouve les étiquettes suivantes :

- L'étiquette 'O' : pour désigner les mots autres que les entités nommées.
- Les étiquettes des entités nommées commencent toutes par B suivi de :
  - Politician, Athlete, Businessperson, Artist, Scientist, Police, Religious, Group, Lawyer : pour désigner une personne.
  - Government, NonGovernmental, Commertial, Educational, Media, Religious, Sports, MedicalScience, Etertainment : pour désigner une organisation.

- Address, Boundary, Kater-Body, Celestial, LandRegionNatural, RegionGeneral, RegionInternational, Continent, Nation, StateOfProvince, County-Or-District, PopulationCenter, GPE-Cluster, Special, BuildingGrounds, SubareaFacility, Path, Airport, Plant : pour désigner un lieu.
- Land, Air, Water, Subare-vehicle, Unspecified : pour designer vehicle.
- Blunt, Exploding, Sharp, Chemical, Biological, Shooting, Projectile, Nuclear, Unspecified : pour désigner une arme.
- Book, Movie, Sound, Hardware, Software, Food, Drug, Other : pour désigner un produit.

On peut aussi trouver ceci : BTAG suivi d'une ou plusieurs ITAG. Dans ce cas-là, l'étiquette qui commence par B représente le début d'une entité nommée ou un mot indice d'une entité nommées (CLUE). Contrairement à la logique utilisée dans le corpus utilisé pour l'elaboration de l'extracteur, le corpus WikiFane repose sur le principe d'extension (idafa).

Cependant, le corpus utilisé dans le développement du SAIE, n'a pas la même structure que celui que nous avons utilisé :

- L'étiquette OTHER a été attribuée aux mots qui ne sont pas des entités nommées.
- Les étiquettes pour les entités nommées sont : PERSON (personne), LOC (location), ORG (organisation), DATE (date).
- Les étiquettes pour les indices des entités nommées sont : PCLUE (Person clue), OCLUE (organisation clue) et LCLUE (location clue).
- NPREFIX pour désigner les préfixes
- PREP pour les prépositions.
- CONJ c'est pour les conjonctions.
- DEF pour les articles.

3.1.1 TagSet proposé :

Par conséquent, le tag set proposé dans ce projet est : LOC PERSON ORG PREP CONJ DEF NPREFIX OCLUE LCLUE PCLUE.

4 Transformation du Corpus

La conversion du corpus WikiFaneGold vers le NE-lexicon tag set s'est fait comme suit :

- L'étiquette LOC a été attribuée à tous les mots qui ont les tags qui désignent le lieu, mentionné précédemment (Address, Boundary, …etc.).
- L'étiquette PERSON a été attribuée à tous ceux qui représentent une personne comme dit plus haut (Artist, Athlete, …etc.).
- Le tag ORG pour regrouper les tags qui représentent les organisations.
- OCLUE, LCLUE, PCLUE sont les indices qui montrent que c'est ORG, LOC, PERSON respectivement.
- PREP pour dire que c'est une préposition.
- CONJ pour dire que c'est une conjonction.
- DEF pour dire que c'est un article.
- NPREFIX est une combinaison des préfixes déjà mentionnés, (ex : CONJ+PREP+DEF).*
- START et END pour les débuts et fin de phrases <s> et <e>

Remarque : Le WikiFane corpus est déjà segmenté, cependant, les NE qui sont précédé par un préfixe ne le sont pas. Cette contrainte rend le traitement de ce corpus difficile. La solution proposée à ce problème est de baser la segmentation sur l'existence du mot dans le lexicon.
Il subsiste tout de même des ambigüités si le mot n'existe pas dans le lexicon NELexicon : Exemple : le mot بالمر n'existe pas dans le lexicon et sera donc segmenté en بالمر

بال سانية <--- بالسانية

بالمر -----> بال مر

Bien que le corpus, WikiFane soit nommé 'Gold Standard', il existe un certain nombres d'erreurs comme certains mots étiqueté à O (OTHER) mais qui sont en réalité des entités nommées. La transformation de O vers le tag correspondant se fera en se basant sur le lexicon.

5 Enrichissement du NE-lexicon

Dans cette étape, nous avons enrichi le NELexicon étiqueté avec le tag set décrit précédemment, avec d'autres lexicons étiquetés de la même manière. Le premier, ArNameLexicon, contient essentiellement des noms de personne, ainsi que quelques noms d'organisations et de lieux. Ces derniers sont uniques pour chaque mot.

Le deuxième, NELexicon31old, qui contient plusieurs entités nommées ainsi que des mots étiquetés 'OTHER'. Aussi, après transformation du corpus WikiFane, les entités nommées inexistantes dans le ArNameLexicon sont ajoutés, nous créerons pour ce faire ArNameLexicon2019 issu de cette fusion.

De la même manière, le NELexicon sera fusionné avec les mots du corpus WikiFane.
Aussi, les mots du WikiFane corpus qui seront transformé en XCLUE ne seront pas ajouté puisque cette Le nombre des nouvelles entités nommées par type est donné par le tableau suivant :

PERSON	VEHICULE	WEAPON	PRODUCT	LOC	ORG
4634	368	126	1275	3966	1850

6 Modèle du Langage

Apres le transformation du corpus, nous procédons à la génération d'une distribution de fréquence des mots du corpus et affectons aux mots dont l'occurrence est inférieure à 3 la valeur 'UNK' représentant un mot inconnu. Cette étape s'avère importante pour la gestion des mots inexistants dans le lexicon lors de la phase d'étiquetage. En effet, dans une séquence à étiqueter, un mot qui n'existe pas dans le lexicon et qui n'est pas transformé en UNK engendra une étiquetage biaisé puisque toutes les probabilités qui le concernent seront à 0.
Nous procédons ensuite à la création des modèles de langage trigramme, bigramme et unigramme.

6.1 Backoff

Le calcul des probabilités de transition et d'émission se fera sur la base de la méthode de backoff smoothing appelée communément 'Stupid' backoff pour sa simplicité. Cette méthode a été introduite par Brants, Thorsten al[2] et repose sur la formule suivante :

$$S(w_i|w_{i-k+1}^{i-1}) = \begin{cases} \frac{f(w_{i-k+1}^{i})}{f(w_{i-k+1}^{i-1})} & \text{si} f(w_{i-k+1}^{i}) > 0 \\ \alpha S(w^{i}|) & \text{sinon.} \end{cases}$$

Ou $\alpha = 0.4$

6.2 l'algorithme Viterbi

Avant de procéder au tag d'une séquence, nous recherchons les mots inexistants dans le lexicon et leur affectons la valeur 'UNK'. Nous appliquerons ensuite l'algorithme de viterbi qui permettra l'extraction d'entités nommées en utilisant une matrice backpointer afin de retrouver la séquence de tag la plus probable.

Input: a sentence $x_1 \dots x_n$, parameters $q(s|u,v)$ and $e(x|s)$.
Initialization: Set $\pi(0,*,*) = 1$, and $\pi(0,u,v) = 0$ for all (u,v) such that $u \neq *$ or $v \neq *$.
Algorithm:

- For $k = 1 \dots n$,
  - For $u \in \mathcal{K}$, $v \in \mathcal{K}$,

$$\begin{aligned} \pi(k,u,v) &= \max_{w \in \mathcal{K}} (\pi(k-1,w,u) \times q(v|w,u) \times e(x_k|v)) \\ bp(k,u,v) &= \arg\max_{w \in \mathcal{K}} (\pi(k-1,w,u) \times q(v|w,u) \times e(x_k|v)) \end{aligned}$$

- Set $(y_{n-1}, y_n) = \arg\max_{(u,v)} (\pi(n,u,v) \times q(\text{STOP}|u,v))$
- For $k = (n-2) \dots 1$,

$$y_k = bp(k+2, y_{k+1}, y_{k+2})$$

- **Return** the tag sequence $y_1 \dots y_n$

FIGURE 2 – Viterbi

7 Tests et Evaluations

Nous testerons le modèle obtenu sur des phrases du corpus wikiFane :

Nombre de phrases	Précision	Recall	f-score
455	0.79	0.96	0.86
115	0.94	0.95	0.94

7.1 Interprétation :

On remarque qu'en augmentant le nombre de phrases de l'ensemble d'apprentissage les performances s'améliorent. Cependant, la segmentation des mots lors de la transformation du corpus wikiFane peut expliquer ces résultats. Exemples de phrases etiquetées :

| توجد/OTHER جمهورية/LCLUE فرعيزيا/LOC في/PREP
ال/DEF جزء/OTHER ال/DEF شرقي/OTHER من/PREP
اسيا/LOC الوسطى/LOC

و/CONJ من/PREP اشهر/OTHER ال/DEF اسواق/LOC في/PREP
مدينة/OTHER ال/DEF موصل/LOC هي/OTHER :/OTHER سوق
/LOC السرجخانة/LOC و/CONJ باب/LOC السراي/LOC و/CONJ
باب/LOC الطوب/LOC

8 Application

Afin d'exploiter le modèle résultant de ce travail, nous avons opté pour une application web qui sert d'outil d'annotation. 'Aleef' permet une annotation facile est rapide au travers d'une interface intuitive, c'est un outil extensible qui prend en charge différents type de donnes textuelles. Bien que nous proposons une application web, il est importer de noter que les données restent en interne, aucun de vos documents annotes ne reste dans nos bases de donnes.

9 Conclusion générale

A l'issue de ce mini-projet, nous pouvons dire que l'objectif principale à été atteint. En effet, nous avons pu mettre en pratique toutes les notions vues en TP et ainsi concevoir notre propre system d'annotation d'entités nommées. Ce travail nous a permis d'en apprendre d'avantage sur l'état de l'art et d'étudier un cas concret, SAIE, ainsi que de faire son implémentation.
L'enjeu majeur de ce projet a été le traitement de la langue arabe, langue encore tres peu traite en TAL due a sa complexité morphologique et au manque de ressources numérisées. Après des tests concluants (un f1-score de 0.86), et une intégration à l'application réussie, nous vous présentons "Aleef", notre outil d'annotation de documents.

Table des figures

Références

[1] Statistical Arabic Information Extraction System BY Fatma Nasser Al Shamsi.

[2] Brants, Thorsten Popat, Ashok Xu, Peng Och, Franz Dean, Jeffrey. (2007)
Large Language Models in Machine Translation.. 858-867

[3] A book recommendation system based on named entities.
Sariki, Tulasi Kumar, B.G.. (2018). Annals of Library and Information Studies.
65. 77-82.

[4] Les entités nommées, de lalinguistique au TAL : statut théorique et méthodesde désambiguïsation.
Maud Ehrmann. 2008. Ph.D. thesis, Univ. Paris 7.

[5] Columbia NLP courses.
http://www.cs.columbia.edu/ mcollins/courses/nlp2011/notes/hmms.pdf

[6] using named entities in post-click news recommendation
Arash Koushkestani, 2015

[7] A Hybrid Approach to Features Representation for Fine-grained Arabic Named Entity Recognition.
Alotaibi, Fahd Lee, Mark. (2014).

A Annexes

A Manuel d'utilisation

A.1 Installation des dépendances

Afin de lancer l'application, il est nécessaire d'avoir certaines librairies installées sur la machine, ces librairies sont indiquées dans le fichier requirements.txt, ce fichier ce trouve dans le dossier principale
Aleef, Toutefois avant de lancer pip install -r requirements.txt, il est necessaire de se connecter a l'environnement virtuel TAL en executant TAL (Nous avons prefere travailler sur dans un environnement virtuel afin d'eviter des conflits dues aux differentes dependances).

A.2 Lancer l'application

Une fois toutes les libraries installées, il suffit d'exécuter python manage.py runserver dans le repertoire , ce script se charge de lancer l'application web.

A Systeme de Recommendations d'articles de presses

Depuis la popularisation des journeaux en ligne, les habitudes des utilisateurs ont complètement changé, preferant acceder a leur contenu prefere en un click. Toutefois ayant maintenant facilement acces a toute l'information disponible. Les utilisateurs surexposes deviennent de plus en plus exigeants et se lassent rapidement, et si les anciens canneaux de communications n'avait aucun mal a combler leur lecteurs , leurs contemporains eux peinent à fidéliser leurs utilisateurs.
Le problème de garder les utilisateurs sur le site Web peut etre resolu en offrant a l'utilisateur ce qu'il veut avant qu'il n'aille le chercher ailleurs. Les systemes de recommendations ont proliferes ces dernieres annees, cependant tres peu se base la REN. Pour cette application nous avons voulu implementer l'approche suivie par Sariki, Tulasi Kumar, B.G dans 'A book recommendation system based on named entities'[3].

A.1 Collecte d'articles

Notre agregateurs de news regroupe un ensemble d'articles de presses collectes de differentes sources. Etant donne que les API disponibles gratuitement ne propose que le resume des articles, nous avons preferes ecire notre propore script aui se charge de scrapper ces dernier de facon automatique.

A.2 Approche de recommendation

La recommendation d'un article se fait comme tel :
A chaque fois qu'un nouvel article est collecte (scrappe), un script se charge de le resumer, puis ce resume est annote par notre systeme. Plus d'entite nommees, cet article a, en commun avec les articles deja considere comme apprecies par l'utilisateurs (le degree d'aprecibilite est represente par le nombre de fois ou l'article est ouvert) plus il est considere comme 'recommendable' est donc sera affiche en haut de la liste des articles de presses proposes par notre agregateur.

A Captures d'écran

A.1 Outil d'annotation

A.1.1 Parametres

La page dediee aux parametrages est divisee en 3 parties distinctes :
La partie corpus, l'utilisateur a la possibilite de choisir le corpus sur lequel l'apprentissage sera fait.

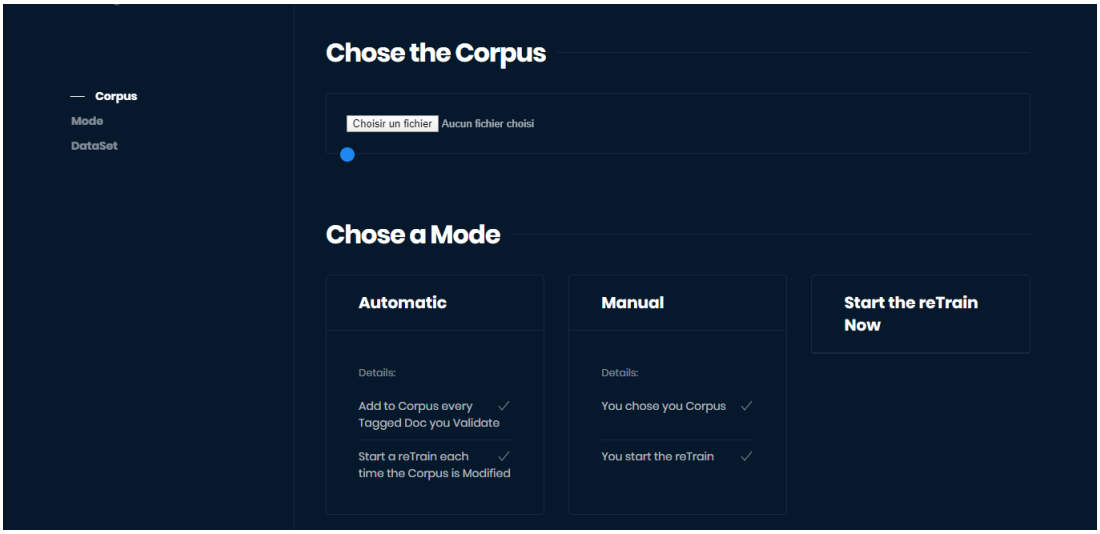

FIGURE 3 – Choix du Corpus d'apprentissage

Une partie nomee 'Mode', ou il y a le choix entre deux modes.

- Le mode automatique : l'apprentissage se relance automatiquement lorsqu'il y a changement du corpus.
- Le mode manuel : l'apprentissage doit etre lance manuellement en cliquant sur le bouton de droite.

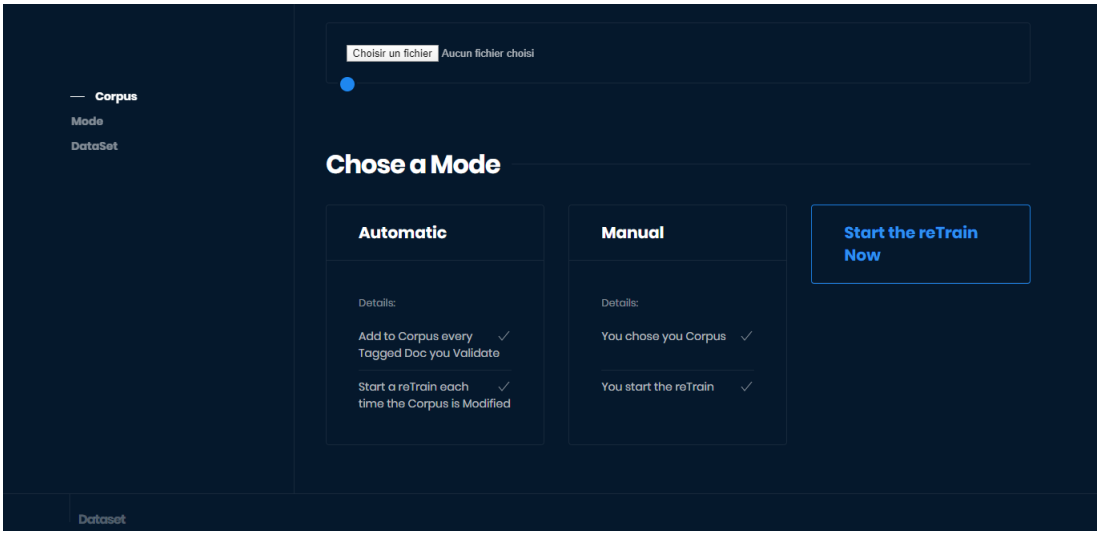

FIGURE 4 – Choix du Mode

Tout en bas se trouve les options liees aux datasets manipules au travers de l'outil d'annotation. Ces options sont :

Creer un document manuellement. Pour cela il suffit simplement d'ecrire le texte que l'on souhaite annoter et lui attribuer un titre. Il est egalement possible d'annoter le document lors de sa creation en assossiant aux termes des tags.

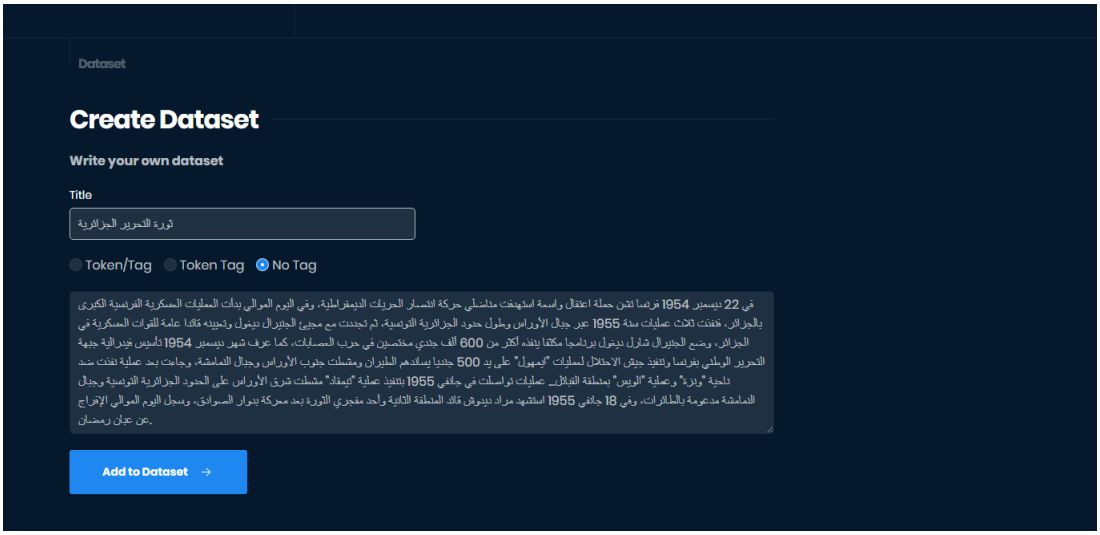

FIGURE 5 – Creer un document

Importer un document. Le document importe peut etre annote ou pas.

FIGURE 6 – Importer un document

Importer un document. Le dataset qui se verra annote automatiquement sera disponible de telechargement via le bouton ci-dessous.

FIGURE 7 – Exporter un document

Les fichiers generes suite al'annotation sont :

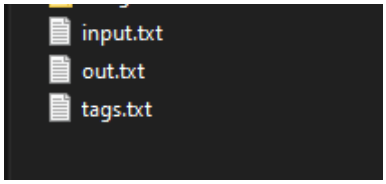

FIGURE 8 – Les fichiers generes

Le fichier annote est de la forme suivante :

و/CONJ في/PREP ال/DEF مجالات/OTHER ال
/DEF عالمية/OTHER اشتهر/OTHER من/PREP هم
/OTHER فرانك/PERSON زابا/PERSON و/CONJ
شاكيرا/PERSON و/CONJ مساري/PERSON و/CONJ
ميكا/PERSON و/CONJ كارل/PERSON وولف
/PERSON

FIGURE 9 – Fichier annote

A.1.2 Annotation

Dans la page 'Dataset', l'utilisateur a une vue d'enseble de tous les documents disponibles a l'annotation. En appuyant sur l'icone poubelle il peut supprimer un document de son dataset.

FIGURE 10 – Visualiser tous les datasets ajoutes/crees disponible a l'annotation

L'annotation se fait au niveau de cette fenetre. A droite l'utilisateur a une vue sur l'ensemble des document disponibles dans son dataset, pour se deplacer d'un document a un autre, il peut tout aussi bien selectionner le document en cliquant dessus ou encore se deplacer a l'aide des bouton de navigation tout en bas. Les boutons de navigation cites a l'exterieur aide a aller d'un document a l'autre tandis aue ceux de l'interieur font defiler les pages dans un meme document.
Toujours dans le but de faciliter le travail d'annotation a l'utilisateur, il peut organiser ces documents selon leur date de creation (ordre ascendant ou bien descendant).

Pour attribuer manuellement une etiquette a un mot il faut simplement selectionner le mot que l'on souhaite etiqueter puis appuyer sur les touche indiquees a cote du tag en question. Par exemple, pour l'etiaquette 'VEHICLE' il y a de note a cote 'C-v' ce qui correspond aux touches 'Ctrl' et 'v'. On appuie dans simultanement sur les deux touches indiquees (avec C : 'Ctrl' et S : 'Sharp'). Une autre facon d'annoter et de selectionner le mot puis cliquer simplement sur le label auquel on souhaiterais l'associer. [2]

2. Il est important de preciser que l'etiquette 'OTHER' n'est pas representee car tout ce aui ne sera pas annote sera par defaut associe a ce label

FIGURE 11 – Annotation d'un document

L'outil d'annotation 'Aleef' permet aussi bien l'annotation manuelle qu'automatique. En effet le petit bouton en haut des etiquettes une fois selectionne permet d'afficher une annotation resultante de notre modele.

FIGURE 12 – Annotation automatique d'un document

Outil de recherche :

FIGURE 13 – Outil de recherche rapide

A.2 Recommendation d'articles de presses

Comme application a notre systeme d'extraction d'entite nomees, nous proposons un agregateur d'articles de presses, ou la recommendation d'articles se base sur les entites nommees.

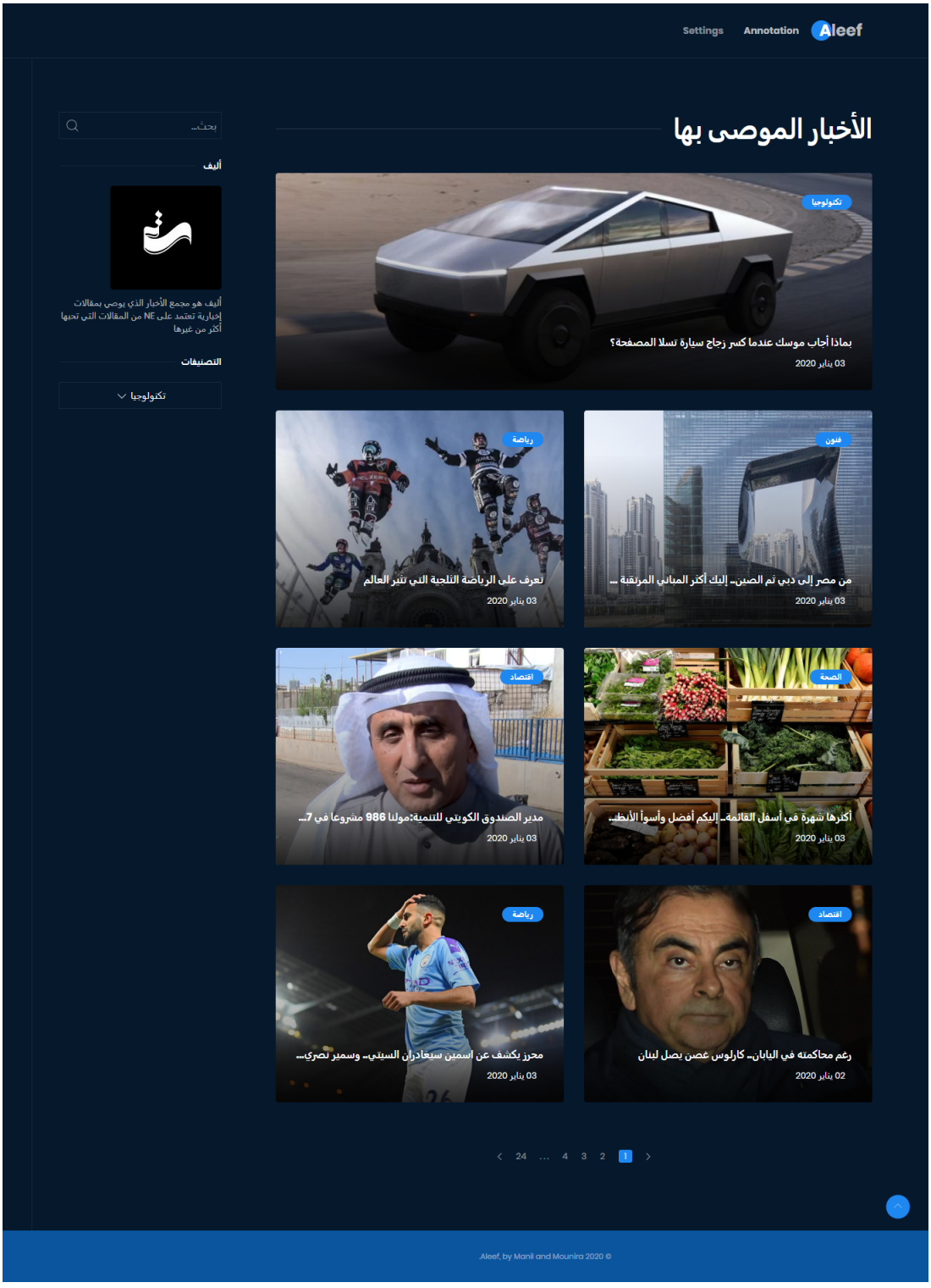

FIGURE 14 – Agregateur d'articles de presses

FIGURE 15 – Article de presses

A Répartition des tâches

Tâche	DJEBRI Mounira	BENDALI Omar Manil	SLIMANI Kenza	SMATI Yasmine
Redaction du rapport	30%	30%	5%	35%
Etude de l'etat de l'art	25%	25%	25%	25%
Transformation du corpus	5%	50%	40%	5%
Enrichissement du NE-lexicon	5%	50%	40%	25%
Implementation de Viterbi avec Backoff	5%	70%	20%	5%
Écriture des scripts d'annotation	20%	45%	25%	10%
Tests et analyse des résultats	15%	80%	5%	0%
Conception de l'application	50%	50%	0%	0%
Realisation de l'inteface	90%	10%	0%	0%
Écriture des fonctions de gestion de la BD	60%	40%	0%	0%