Développement d'un étiqueteur de parties du discours et d'un étiqueteur de groupes pour le français

Rémi Roques
remi.roques.657@student.lu.se

Mathieu Martinet
mathieu.martinet.674@student.lu.se

Abstract

Ce document présente un projet réalisé dans le cadre du cours de *Language Processing and Computation Linguistic* dirigé par Pierre Nugues à Lunds Tekniska Högskola, Suède. Le but du projet est de développer un étiqueteur des parties du discours et un étiqueteur de groupes nominaux et verbaux pour le français. Le programme, écrit en Java et utilisant un outil d'apprentissage, a été entrainé sur un corpus annoté et a été testé sur un corpus non annoté. Le projet représente aussi une étude de faisabilité dans le but d'améliorer le projet Direkt Profil.

1 Introduction

Le projet Direkt Profil a pour objectif de réaliser une analyse grammaticale automatique de textes écrits en Français. Il est le résultat d'une collaboration entre l'Institut d'Etudes Romanes de l'Université de Lund et du Département d'Informatique de l'Institut de Technologie de Lund.
Nous avons travaillé avec Jonas Granfeld et Pierre Nugues pour proposer une amélioration du système actuel en développant un étiqueteur des parties du discours et d'un étiqueteur de groupes pour le français.
Un programme a été développé en Java. Les résultats sont présentés dans ce document.

2 Le Corpus

2.1 Corpus annoté

Un corpus annoté, compilé par Anne Abeillé de l'Université de Paris 7, a été utilisé pour entrainer les *classifiers.*
Il se compose d'articles issus du journal Le Monde et contient environ 600000 mots.
Les fichiers sont sous format xml. Les phrases sont repérées par la balise SENT, les mots par la balise W. A l'intérieur des balises W, l'attribut CAT indique la partie du discours et l'attribut COMPOUND indique si le mot est un mot composé ou non. D'autres attributs tels que LEMMA ou EI peuvent aussi renseigner sur la racine du mot ou le genre et le nombre. Ces catégories de classification n'ont pas été utilisées dans le cadre de ce projet.

2.2 Corpus non annoté

Un corpus non annoté a été utilisé pour tester le programme dans les conditions de Direkt Profil. Il a été compilé par Malin Ågren et se compose de textes écrits par des étudiants suédois apprenant le français. Les textes sont de différents niveaux grammaticaux.
Les textes sont regroupés dans un seul fichier xml. L'auteur, le niveau, le nom du lycée sont indiqués en tant qu'attributs xml.

3 Weka

Weka [2] est un outil d'apprentissage développé à l'Université de Waikato en Nouvelle-Zélande. Il permet de produire des arbres de décisions en fonction de données présentées en entrée.
Weka prend en entrée un fichier Attribute-Relation File Format (ARFF). La figure 3 présente la forme d'un tel fichier.

```
@relation <relation-name>

@attribute1 <attribute-name1> <datatype1>
{values}
@attribute2 <attribute-name2> <datatype2>
{values}
..

@data
valueAtt1, valueAtt2, ..
..
```

Figure 1 Fichier ARFF

Le but du fichier ARFF est de présenter les données à traiter par Weka. Le dernier attribut est l'attribut prépondérant. Celui que l'on veut obtenir en fonction des précédents attributs.
Différents algorithmes sont disponibles. Dans ce projet, nous avons utilisé les arbres de décision J48 avec une validation croisée de facteur 10.

4 L'étiqueteur des parties du discours

Le but de notre étiqueteur est d'attribuer à chaque mot d'un texte inconnu sa partie du discours correspondante.
Le corpus annoté d'Anne Abeillé utilise les notations suivantes pour les parties du discours:

- A (adjectif)
- Adv (adverbe)
- CC (conjonction de coordination)
- Cl (pronom clitique faible)
- CS (conjonction de subordination)
- D (déterminant)
- ET (mot étranger)
- I (interjection)
- NC (nom commun)
- NP (nom propre)
- P (préposition)
- PREF (préfix)
- PRO (pronom fort)
- V (verbe)
- PONCT (ponctuation)

Deux approches ont été considérées : l'utilisation de la partie du discours la plus fréquente pour chaque mot, ou l'utilisation d'un arbre de décision.

4.1 Partie du discours la plus fréquente

Cette méthode simple consiste uniquement à parcourir le corpus annoté, et à comptabiliser pour chaque mot la fréquence de chaque partie du discours associée à ce mot. Lors de l'annotation d'un texte non annoté, nous donnons à chaque mot sa partie du discours la plus fréquente.
Sur un corpus riche comme celui d'Anne Abeillé, les résultats sont plutôt convaincants : 90% de d'étiquetage positif.

4.2 Entrainement d'un *classifier*

L'inconvénient de l'approche précédente vient de l'existence de mots ambigus. Par exemple, le mot « sort » peut être considéré comme un verbe (sortir au présent à la troisième personne du singulier) ou un nom commun (« La sorcière jette un sort » par exemple). Suivant le corpus d'apprentissage, il apparaîtra que l'une de ces deux parties du discours est plus fréquente. D'où un mauvis étiquetage si l'autre partie du discours est la bonne.
L'utilisation d'un outil comme Weka permet de prendre en compte plusieurs paramètres pour définir une étiquette. Un contexte peut être choisi, c'est-à-dire, prendre en compte d'environnement d'un mot.
Nous avons décidé d'utiliser le vecteur suivant :

Mot, POS-1, POS

Il utilise le mot, la partie du discours du mot précédent, et la partie du discours la plus fréquente du mot en question.
Le choix de ce vecteur vient du fait que la prise en compte d'un contexte plus grand demandé plus de ressources informatiques que nous ne possédions pas.
Un exemple simple d'application de ce vecteur est le suivant :

```
Le      D
grand   A
garçon  N
danse   V
```

Figure 2 Vecteur de paramètres de l'étiqueteur de parties du discours

En plus de ce vecteur, nous avons implémenté une limitation à notre étiqueteur: la prise en compte d'une fréquence de coupure F. Pour limiter la taille dictionnaire, les mots présents moins de F fois dans le corpus, sont annotés comme "autre mot" (OTHER_WORD). Cette limitation vient de notre

faible ressource matérielle (mémoire vive de l'outil de calcul) qui ne permet pas d'utiliser des attributs avec trop de valeurs possibles.
Les résultats obtenus, avec une fréquence de coupure de 120, et l'analyse de 20 textes[1]:

TP Rate	FP Rate	Precision	Recall	F-measure	Class
0.851	0.149	0.656	0.851	0.741	**N**
0.408	0.002	0.938	0.408	0.569	**ADV**
0.892	0.065	0.722	0.892	0.798	**D**
0.692	0.045	0.654	0.692	0.672	**V**
0.604	0.001	0.966	0.604	0.743	**CL**
0.96	0.001	0.994	0.96	0.977	**PONC**

Tableau 1 Résultats étiqueteur de parties du discours

Les résultats se situent aux alentours de 70-72%, ce qui est honorable mais nettement inférieur à la seule utilisation de la partie du discours la plus fréquente.
Ces résultats peuvent être très nettement améliorés en utilisant un contexte plus grand et en n'utilisant pas de fréquence de coupure.

5 L'étiqueteur de groupes verbaux et nominaux

Le but de l'étiqueteur de groupes est de mettre en évidence les groupes nominaux et verbaux d'un texte.
Le corpus d'Anne Abeillé distingue ces groupes.
Nous avons alors utilisé Weka comme outil d'apprentissage afin d'entrainer un *classifier* permettant une annotation optimale des groupes verbaux et nominaux d'un texte où seuls les parties du discours sont précisées.
Le schéma d'annotation utilisé est de type IOB2 : Begin, Inside, Between. Utilisant ce schéma, nous avons les cinq différents tags :

- Groupes nominaux : B-NP, I-NP
- Groupes verbaux : B-VN, I-VN
- Autres : O

```
Six       PRO   B-NP
d'entre   P     O
eux       PRO   B-NP
ont       V     B-VN
pu        V     I-VN
se        CL    B-VN
réfugier  V     I-VN
```

Figure 3 Exemple annotation IOB2

La principale difficulté est alors de trouver un bon vecteur de paramètre afin de représenter les données à Weka. La solution optimale en fonction des résultats obtenues et de notre configuration matérielle est représentée figure 3.

```
Six       PRO   B-NP
d'entre   P     O
eux       PRO   B-NP
ont       V     B-VN
pu        V     I-VN
se        CL    B-VN
réfugier  V     I-VN
```

Figure 4 Vecteur de paramètres

Afin d'étiqueter le mot « pu », nous donnons a Weka les partie du discours de ce mot, du mot précédent et du mot suivant. Nous donnons aussi l'étiquette de groupe donnée au mot précédent.

Le vecteur est de la forme :

POS-1, POS, POS+1, CHUNK-1, CHUNK

Cette solution a le mérite de posséder un petit contexte (prise en charge des parties du discours des mots voisins) et d'être dynamique (prise en compte l'étiquette de groupe donnée précédemment).

[1]Les indices les plus significatifs sont les suivants :
precision = nbr d'instances proposées correctes / nbr d'instances proposées
recall = nbr d'instances proposées correctes / nbr d'instances possibles
F-measure = moyenne harmonique de ***recall*** et de ***precision***

5.1 Résultats de l'étiqueteur de groupes

Les résultats sont les suivants :

TP Rate	FP Rate	Precision	Recall	F-measure	Class
0.957	0.014	0.951	0.957	0.954	**B-NP**
0.929	0.019	0.948	0.929	0.938	**I-NP**
0.902	0.003	0.964	0.902	0.932	**B-VN**
0.94	0.008	0.888	0.94	0.913	**I-VN**
0.949	0.033	0.932	0.949	0.94	**O**

Tableau 2 Résultat étiqueteur de groupes

Le pourcentage d'instance correctement classifiées est de 94,2%.

6 Application à un corpus non annoté

L'étiqueteur de parties du discours et de groupes a été appliqué sur les textes du corpus non annoté collecté par Malin Ågren.

Les résultats sont présentés ici selon le niveau de l'étudiant auteur du texte.

6.1 Niveau débutant

```
C           CL              B-VN
est         V               I-VN
deux        D               B-NP
filles      N               I-NP
,           PONCT           O
une         D               B-NP
fille       N               I-NP
avec        P               O
les         D               B-NP
cheveux     N               I-NP
blondes     NOT_FOUND       O
,           PONCT           O
et          C               O
une         D               B-NP
fille       N               I-NP
avec        P               O
les         D               B-NP
cheveux     N               I-NP
brunes      NOT_FOUND       O
et          C               O
longes      NOT_FOUND       O
.           PONCT           O
```

Figure 5 Exemple niveau débutant

Nous remarquons que certains mots ne sont pas reconnus. Cela signifie que ces mots n'étaient pas présents dans le corpus d'entrainement (corpus d'Anne Abeillé). Nous leur donnant la valeur NOT_FOUND. Malgré le contexte de l'étiqueteur de groupes, les étiquettes de groupe pour ces mots inconnus au système sont mal devinées.

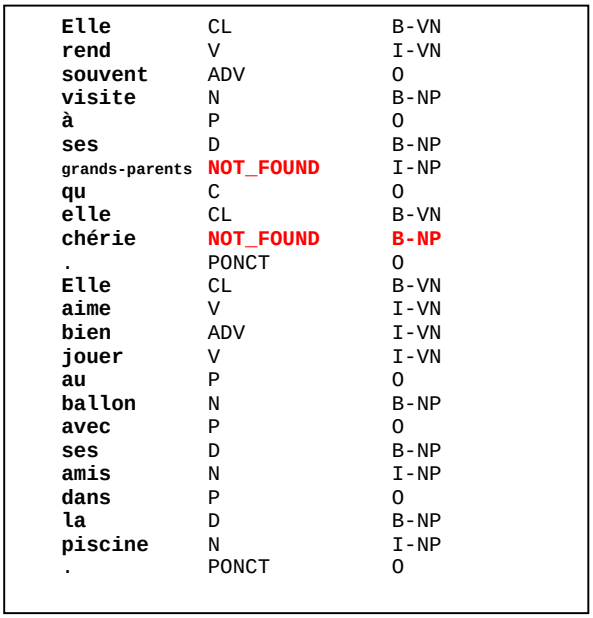
6.2 Niveau Intermédiaire

```
Je          CL              B-VN
fais        V               I-VN
tout        ADV             O
que         C               O
suis        V               B-VN
intresent   NOT_FOUND       B-NP
et          C               O
rigole      NOT_FOUND       O
,           PONCT           O
par         P               O
exemple     ADV             O
faire       V               B-VN
le          D               B-NP
cuisine     N               I-NP
,           PONCT           O
manger      V               B-VN
,           PONCT           O
faire       V               B-VN
du          P               O
jogging     NOT_FOUND       B-NP
,           PONCT           O
aller       V               B-VN
de          P               O
fêtes       N               B-NP
etc.etc     NOT_FOUND       I-NP
.           PONCT           O
```

Figure 6 Exemple niveau intermédiaire

Une remarque très intéressant est que, alors que l'outil ne reconnaît pas le mot « jogging », l'utilisation n d'un contexte dans l'étiqueteur de permet de proposer justement l'étiquette B-NP. C'est dans cet esprit que nous voulons que l'outil se comporte vis-à-vis des mots inconnus.

6.3 Niveau avancé



Elle	CL	B-VN
rend	V	I-VN
souvent	ADV	O
visite	N	B-NP
à	P	O
ses	D	B-NP
grands-parents	**NOT_FOUND**	I-NP
qu	C	O
elle	CL	B-VN
chérie	**NOT_FOUND**	**B-NP**
.	PONCT	O
Elle	CL	B-VN
aime	V	I-VN
bien	ADV	I-VN
jouer	V	I-VN
au	P	O
ballon	N	B-NP
avec	P	O
ses	D	B-NP
amis	N	I-NP
dans	P	O
la	D	B-NP
piscine	N	I-NP
.	PONCT	O

Figure 7 Exemple niveau avancé

Le même phénomène que celui évoqué dans l'exemple du niveau intermédiaire apparaît ici. Le mot « grand-parents » n'est pas reconnu. Cependant le contexte donné à l'étiqueteur de groupes permet au système de deviner son étiquette de groupe.

7 Conclusion

Les résultats obtenus sont très intéressants et encourageants. L'utilisation d'une machine d'apprentissage comme Weka avec un grand contexte permettent de deviner de façon optimale quelle étiquette est la plus probable malgré les fautes d'orthographe et de grammaire. C'est dans ce sens que nous désirons que le système se comporte.

Cependant il faut faire remarquer qu'un outil comme Weka demande des ressources matérielles (mémoire vive) non négligeable. Afin d'obtenir de meilleurs résultats, notamment dans l'étiqueteur des parties du discours, il est indispensable de disposer d'une autre ressources qu'un simple ordinateur personnel.

8 Remarques sur l'implémentation

L'outil a été développé en Java. Il utilise la librairie JDOM [3] pour lire les fichiers xml. Il utiliser aussi l'API fournie par Weka pour interfacer ce dernier avec notre application.

Références

[1] Pierre Nugues. 2006. *Introduction to Language processing with Perl and Prolog*. Springer.

[2] www.cs.waikato.ac.nz/ml/weka

[3] www.jdom.org