TP2 : Analyse lexicale désambiguisée

BOUHOUCHE Sami

10468326

EL HELOU Myriam

11611686

UFR LLASIC

Master Sciences du langage

Parcours Industries de la langue

Année 2017/2018

Introduction :

Le but de ce TP est d'évaluer et de comparer la qualité de deux étiqueteurs morpho-syntaxique. Nous avions à disposition l'extrait d'un article du journal *Le Monde : Ce qu'il faut retenir des négociations européennes sur le glyphosate.* Nous avons utilisé deux analyseurs : l'analyseur de *XEROX* et *Treetagger.* Nous procéderons d'abord à une évaluation quantitative et ensuite nous effectuerons une analyse qualitative des résultats.

Comparaison des jeux d'étiquettes :

Nous avons remarqué que les analyseurs rajoutent parfois des catégories pour « fuir » l'ambiguïté. Par exemple, dans *Xerox,* nous remarquons que pour ne pas distinguer les différentes catégories grammaticales de « que » ou de « à », des catégories spécifiques ont été créé « +CONJQUE », « +PREP_A ». Ce choix peut être expliquer par le fait que les créateurs de cet analyseur souhaitaient optimiser le taux de précision, donnant ainsi l'illusion que leur outil est plus performant. Tandis que *Treetagger* indiquent des catégories « classiques » pour ces mots.

En ce qui concerne les verbes, *Xerox* nous fournit plus d'informations que *treetagger :*

Xerox		Treetagger	
a	+VAUX_P3SG	a	VER:pres

Si nous analysons ce cas, nous remarquons que *Xerox* indique que c'est un verbe auxiliaire alors qu'on ne retrouve pas cette notion avec *Treetagger,* quant aux traits, *Xerox* indique la personne et le nombre du sujet, tandis que *Treetagger* indique les temps verbaux.
Treetagger évite l'ambiguité inhérente à la personne du sujet pour augmenter sa précision (ex : je fais, tu fais). Tandis que *Xerox* évite celle inhérente aux temps verbaux (ex : qu'il fit à l'impératif du subjonctif et il fit au passé simple).

Evaluation quantitative :

A l'aide d'un tableau Excel, nous avons répertorié l'ensemble des résultats obtenus avec chaque étiqueteur pour faciliter la comparaison.
Nous avons obtenu un total de 503 mots. Les ponctuations et les symboles ne seront pas pris en compte dans le calcul. Nous avons repéré et comptabilisé les erreurs du jeu d'étiquettes que nous allons présenter sous forme de tableau.

Evaluation quantitative de *Xerox :*
Nous avons repéré 12 erreurs de catégories et 35 erreurs de traits et de catégories.
Nous avons repéré 16 unités polylexicales, seulement 4 ont été reconnues. Cela nous amène aux résultats suivant :

Types de catégories	Précision
% de catégories reconnues	97,6%
% de catégories + traits reconnus	93%
% d'unités polylexicales reconnues	25%

Evaluation quantitative de *Treetagger :*
Nous avons remarqué que cet étiqueteur n'indique pas le genre et le nombre des noms et des adjectifs. Nous avons donc procédé à deux calculs différents, un où on considérait que cela est une erreur, un autre où on considérait que ce n'est pas une erreur. Nous avons repéré 26 erreurs de catégories, 170 erreurs de traits et catégories si nous considérions que le non-étiquetage des noms et des adjectifs est une erreur et 29 si nous considérions que ce n'en était pas une. En ce qui concerne les unités polylexicales *Treetagger* n'en a détecté aucune.

Types de catégories	Précision
% de catégories reconnues	94,8%
% de catégories + traits reconnus (170 erreurs)	66,2%
% de catégories + traits reconnus (29 erreurs)	94,2%
% d'unités polylexicales reconnues	0%

Evaluation qualitative :

Après avoir observé en détails les erreurs relevées, nous remarquons que certaines erreurs sont récurrentes :

Les erreurs de *Xerox :*

- Les pronoms : l'étiqueteur de *Xerox* n'attribue aucun trait pour les pronoms.

Xerox		Treetagger	
Il	+PRON	Il	PRO:PER

Les sous-catégories des pronoms sont nombreuses et peuvent porter confusions. Exemple : « il » peut être un pronom personnel ou un pronom impersonnel. Il est possible que ce choix ait pour but d'éviter l'ambiguïté et d'optimiser les résultats.

- Les abréviations : l'étiqueteur de *Xerox* ne reconnait pas les abréviations.

Xerox		Treetagger	
UE	+NOUN_INV	UE	ABR

Cette erreur peut être liée à l'absence de la notion d'abréviation dans le lexique de *Xerox*.

- Les temps verbaux : l'étiqueteur de *Xerox* ne précise pas les temps verbaux.

Xerox		Treetagger	
ont	+VAUX_P3PL	Ont	VER:pres

Certaines formes de conjugaison se ressemblent, afin d'éviter les erreurs, il est possible que les développeurs aient décidé de ne pas l'indiquer.

Les erreurs de *Treetagger :*

- Les adjectifs : une erreur est la confusion adjectif (issus du participe passé) - verbe. Elle peut être liée à la taille du lexique de cet étiqueteur.

- Les catégories : une autre erreur récurrente est liée aux mots commençants par des majuscules. Ces mots ne sont pas reconnus correctement.
  Exemple : « Mais » qui est reconnu une fois en tant que conjonction de coordination, une autre fois comme un nom propre ou en tant que verbe. Nous pouvons en déduire que Treetagger est sensible à la casse et que la présence d'une majuscule en milieu de phrase l'a induit en erreur.
  Une amélioration de l'analyse textuelle pourrait éviter cette erreur.

- Le nombre : *Treetagger* ne précise pas le nombre des noms et des adjectifs.

- Les unités polylexicales : Treetagger n'en a reconnu aucune. Cette erreur peut être le résultat d'un lexique insuffisant et non spécialisé.

Une erreur commune aux deux étiqueteurs est la reconnaissance de la catégorie du sujet lorsque celui-ci est inversé dans la phrase :
« Sondée par la Commission européenne… »
Tous les deux l'ont reconnu en tant que NOM alors qu'il s'agit d'un PARTICIPE PASSE.
Quand la structure classique de la phrase n'est pas respectée, les étiqueteurs ont des difficultés à effectuer une bonne analyse catégorielle.

Conclusion :

Ces étiqueteurs donnent des résultats globalement satisfaisants. Les erreurs qui persistent sont liées à l'ambiguïté du langage naturel et à un lexique incomplet. Ces problèmes sont assez difficiles à résoudre.