Rapport de projet
présenté par :

Alexane Jouglar
Master 1
« Langue et informatique »
Sémantique computationnelle
Enseignante :
Victoria Eyharabide
Soutenu le :
14/12/2020

Table des matières

Table des figures

Introduction

Le travail mené ce semestre dans le cadre de l'enseignement de Sémantique computationnelle constitue une introduction à l'annotation automatique de texte. Réalisé sur GATE (General Architecture for Text Engineering), il pose les bases du fonctionnement de ce logiciel et d'une notion fondamentale qui le caractérise : l'annotation. En effet, le résultat du travail effectué résulte en de multiples annotations, chacune réalisée avec des méthodes différentes. Les annotations ont été réalisées par ordre de difficulté croissante : annotations manuelles d'abord, puis annotations à partir de dictionnaires, et enfin, annotations à l'aide de grammaires JAPE (Java Annotation Patterns Engine). Deux corpus ont été choisis pour accueillir ces annotations : un corpus en français et un corpus en anglais. Le corpus français contient 25 chapitres extraits de l'œuvre de Victor Hugo nommée *Les Travailleurs de la mer* [Hugo, 1866]. Quant au corpus anglais, il contient 8 chapitres de l'œuvre de Lewis Carroll intitulée *Alice's adventures in Wonderland* traduite en français comme *Les Aventures d'Alice au pays des merveilles* [Carroll, 1865]. Ces œuvres fictionnelles au vocabulaire très riche permettent de nombreuses possibilités d'annotations. A ce titre, nous reviendrons dans la seconde partie de notre développement sur l'avantage des annotations faites avec dictionnaires. Mais avant toute chose, nous aborderons les annotations manuelles et leurs limites. Ce n'est qu'à la fin et en troisième partie que nous évoquerons les grammaires JAPE et l'étendue de leurs possibilités. A la fin de ce rapport, nous connaîtrons les tenants et aboutissants du travail d'annotation.

Chapitre 1

Les annotations manuelles

Les annotations manuelles ont été réalisées à partir de schémas écrits en format xml (Extensible Markup Language). Quatre schémas d'annotations ont été produits pour les deux corpus. Ils annotent respectivement : la structure du texte (titre et dialogues), les personnages (principaux et secondaires), les lieux (intérieurs ou extérieurs), les pensées (à voix haute ou bien sans guillemets). Un schéma d'annotation supplémentaire a été créé pour le corpus français : sexe (figurant féminin ou masculin).

1.1 Le choix des annotations

Les choix qui ont été faits pour les annotations manuelles relèvent du pragmatisme. En effet, les éléments à annoter sont suffisemment communs aux deux textes pour pouvoir les réutiliser dans l'un et l'autre (lorsqu'il s'agit des quatre premiers schémas d'annotation), mais suffisemment intéressants pour que l'on prenne conscience de la différence de contenu des textes. On remarque ainsi dans *Alice's adventures in Wonderland* que les annotations de pensées sont extrêmement présentes, à la différence du texte de Victor Hugo où le narrateur choisit de ne pas rapporter les pensées des personnages et où le personnage principal est plutôt taciturne et mystérieux. Avec ce genre d'annotation, on peut donc mettre en valeur le style d'un texte.

1.2 Début du travail d'annotation

Aux débuts de ce genre de travail, il est aisé de prendre conscience de la facilité de ce type d'annotation. Avec un simple schéma comme celui présenté en figure 1, nous pouvons ensuite, à la main, aller surligner dans le texte tous les éléments pertinents et leur assigner un type et éventuellement un

sous-type. Pour certaines des annotations, des expressions régulières ont été utilisées. Ces dernières permettent de surligner dans le texte des passages similaires ou des apparitions identiques sans qu'il n'y ait besoin de le faire point par point. Ainsi ont été annotées avec des expressions régulières : les personnages, les figurants (féminins ou masculins), certains lieux : "Le Bû de la rue" par exemple [Hugo, 1866].

Par ailleurs, ce type de schéma n'étant pas soumis aux contraintes d'une chaîne de traitement conditionnelle, il est possible d'utiliser un schéma pour n'importe quel texte indépendemment de sa langue (si tant est que les textes contiennent des éléments similaires à annoter).

1.3 Limites du travail d'annotation

Comme indiqué dans l'introduction, le corpus se compose de 8 documents en anglais, et de 25 documents en français. Soit 33 documents au total. Le premier document est, certes, facile à annoter, mais la difficulté survient sur la longueur. Que ce soit avec des expressions régulières ou non, annoter des documents avec cette méthode se révèle vite contraignant. D'ailleurs, le corpus anglais a été annoté entièrement, mais on remarquera que le corpus français n'a été annoté que partiellement. En effet, la contrainte du temps est d'autant plus forte lorsque l'on perd des annotations. C'est ce qui a pu se produire à de multiples reprises (à cause du traitement "Document Reset", compris dans l'application Annie) lors de la réalisation de ce travail et qui m'a conduite à délaisser une partie du corpus français lors des annotations manuelles afin d'optimiser ce travail et de le rendre aussi utile qu'agréable.

Chapitre 2

Utilisation des dictionnaires

L'utilisation des dictionnaires permet d'automatiser un peu plus le travail d'annotation. Les dictionnaires utilisés pour nos annotations sont légion et se divisent en deux groupes : les dictionnaires non-flexibles, et les dictionnaires flexibles. Les premiers ont été appliqués aux deux corpus. Les seconds au corpus anglais seulement.

2.1 Un pas dans l'automatisation : les dictionnaires

Les dictionnaires permettent, grâce à l'écriture de fichiers .lst et de fichiers .def, d'introduire dans le travail une automatisation que nous n'avions pas précédemment. Ainsi dans le corpus français se retrouvent annotés les lieux (pays, îles), les lèmes correspondant à une description physique (grandeur, forme, beauté/âge, cheveux, force), les références temporelles (moments passés, références temporelles générales), les matières (tissus, couleurs). Pour le corpus anglais se retrouvent annotés d'une part les lèmes se rapportant à la description physique des choses et personnes (couleur, taille, apparence), les espèces (petits mammifères et insectes), d'autre part les verbes de mouvement, de parole, et d'expression (grâce aux dictionnaires flexibles). L'intérêt de ce type d'annotation est grand. On sait par exemple que la notion de grandeur se situe au coeur de l'histoire du texte anglais *Alice's adventures in wonderland* et il est très intéressant de voir avec ce type de dictionnaire en un clin d'oeil la quantité d'adjectifs de grandeur que recèle le texte (voir figure 2.1).

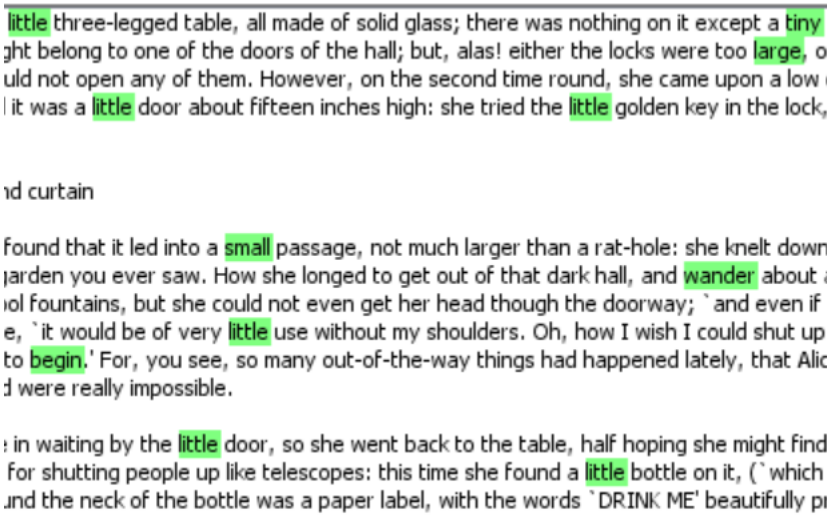

FIGURE 2.1 – De nombreux adjectifs qualifiant la taille

2.2 Les dictionnaires non-flexibles vis-à-vis de leurs homologues flexibles

La préparation est la même. La mise en œuvre seulement diffère (de peu, puisqu'il s'agit d'introduire seulement un "Flexible gazetteer"). Les deux types de dictionnaire sont valables et m'ont paru fournir de riches possibilités. J'aurais aimé pouvoir appliquer un dictionnaire flexible à mon corpus français. Malheureusement, le temps nous a manqué. Ceci dit, j'ai pu entrevoir la qualité du dictionnaire flexible lors de l'annotation de mon corpus anglais.

2.3 Pour aller plus loin

Il me semble que les dictionnaires sont très utiles associés à d'autres types de traitements. On peut penser par exemple à associer un dictionnaire flexible avec différents types de verbes annotés à une grammaire JAPE. Cela n'a pas été fait pour ce travail mais je ne manquerai pas de tester ce type d'usage dans le futur.

Chapitre 3

Utilisation des JAPE

Les grammaires JAPE appartiennent à la dernière étape du travail réalisé. Elles nous permettent également de nous préparer au travail d'annotation qui adviendra lors du second semestre de ce Master1.

3.1 Variétés des grammaires JAPE

Ce qui m'a le plus impacté lors de mes premiers pas avec ce type de grammaire concerne la variété des grammaires JAPE. La consigne était de réaliser des grammaires pour "annoter de façon automatique des passages important dans l'histoire racontée" puis de rafiner cette annotation. Deux chemins se sont ouverts devant moi pour mettre en pratique cette consigne. Le premier se résume à l'écriture d'une grammaire JAPE dans laquelle j'annote les passages désirés ; cette grammaire serait réutilisée par la suite dans une autre grammaire JAPE grâce à l'expression "Token within MAGRAMMAIRE" (MAGRAMMAIRE étant un exemple de nom de grammaire) afin de rafiner les éléments recherchés (voir figure 3.1). Le deuxième chemin et celui choisi consiste à écrire une seule et même grammaire et d'annoter au sein de cette même grammaire les éléments recherchés. J'ai choisi cette deuxième option parce qu'elle me paraissait plus intéressante aux vues du texte concerné. En effet, dans mon corpus anglais, les passages importants ne disposaient pas tous des mêmes éléments et le rafinement était donc compliqué. J'ai préféré faire un traitement plus général.

3.2 Limites des grammaires JAPE

Les grammaires JAPE mettent un certain temps à préparer et la complexité que l'on peut souhaiter introduire dans l'une d'elle entre en conflit

```
Rule:RuleExemple

(
 {Token within MaGrammaireJAPE, Token.category=="NN"}
):tt
-->
:tt.TokenNoun = {rule = "RuleExemple"}
```

FIGURE 3.1 – Exemple de JAPE utilisant une autre JAPE

avec l'interface de GATE. En effet cette dernière m'a créé beaucoup de problèmes. Lorsque ma grammaire était trop compliquée (voir figure 3.2), GATE n'arrivait pas à l'exécuter sur l'ensemble du corpus et se figeait définitivement. Il me fallait alors relancer le programme. Ceci dit, j'ai conscience que le travail en JAPE constitue une étape nécessaire au travail d'annotation et que ce type de problème ne se produira plus lorsque le travail ne sera plus effectué sur l'interface même mais sur le terminal.

```
Phase:  ImportantMoment
Input: Token SpaceToken
Options: control = appelt

Rule:MomentRule
({Token.orth == "upperInitial"}
({SpaceToken}|{Token.kind == punctuation}|{Token.kind == word}|
(({Token.category == VBD}|{Token.category == VBP}|{Token.category == VB})*): verb|
({Token.category == "NNP"}) : noun)*
 (({Token.string == "when"}|{Token.string == "sudden"}|{Token.string == "suddenly"}|
 {Token.string == "out of the sudden"}|{Token.string == "all of a sudden"}|{Token.string == "swiftly"}|
 {Token.string == "abruptly"}|{Token.string == "quickly"}|{Token.string == "shortly"}|
 {Token.string == "all at one"}|{Token.string == "on spur of moment"}|{Token.string == "without warning"}|
 {Token.string == "asudden"}|{Token.string == "forthwith"}|{Token.string == "unexpectedly"})+
({SpaceToken}|{Token.kind == punctuation}|{Token.kind == word}|(({Token.category == VBD}|
{Token.category == VBP}|{Token.category == VB})*): verb|({Token.category == "NNP"}) : noun)*): time
{Token.string == "."})

:verb.Verb= {rule= "MomentRule"},
:noun.Noun= {rule = "MomentRule"},
:time.Action= {rule = "MomentRule"}
```

FIGURE 3.2 – Exemple de JAPE trop compliquée à exécuter sur l'interface

Conclusion

La notion de coût d'entrée permet à coup sûr de conclure de manière adéquate le travail présenté. Dans chacune des trois parties, une méthode d'annotation a été illustrée. Elles ont été classées par ordre de difficulté croissante. L'annotation manuelle qui prend peu de temps à préparer implique pourtant une perte de temps considérable. Les annotations réalisées avec une grammaire JAPE impliquent une préparation en amont plus longue (l'écriture d'un fichier JAPE n'est pas anodine ; il faut ensuite préparer la chaîne de traitement) pourtant le temps total consacré à annoter est moins long. Quant aux dictionnaires, ils impliquent logiquement un temps moyen de travail comparé aux deux autres méthodes. Le résultat est plus précis lors d'annotations manuelles, sans doute, mais moins efficace. Par ailleurs, j'aimerais ajouter que les dictionnaires offrent de riches possibilités si l'on associe leur usage aux grammaires JAPE. Enfin, deux éléments du travail n'ont pas faits l'objets de parties spécifiques mais méritent d'être évoqués : la détection de la langue de chacun des textes avec le plugin TextCat Language Identification et la création d'une application de traitements conditionnels qui a été évoquée dans les deux derniers chapitres.

Remerciements

Le rapport présent n'est qu'un simple rapport de travail d'une introduction au travail d'annotation, mais je suis très heureuse d'avoir pu l'effectuer. Avant de laisser place à ma bibliographie, je souhaite vous remercier Madame Victoria Eyharabide parce que, aux côtés des autres professeurs, vous n'avez pas cessé de nous dispenser vos enseignements alors même que la situation dans laquelle nous nous trouvons actuellement impacte grandement notre qualité de vie.

Bibliographie

[Carroll, 1865] Carroll, L. (1865). *Alice's adventures in Wonderland*. Macmillan and Co.

[Hugo, 1866] Hugo, V. (1866). *Les Travailleurs de le mer*. Albert Lacroix et Cie.