Rapport de projet final sur GATE

Encadrant : **Victoria Eyharabide**
Etudiante : **Yuyan QIAN, 21108372**

Table des matières

Table des figures

Introduction

Le projet final dans le cadre de l'enseignement de Sémantique computationnelle consiste à annoter manuellement et automatiquement sur GATE deux livres, un en français et un en anglais. Le platforme GATE (General Architecture for Text Engineering) nous propose une fonctionnalité complète concernant l'analyse textuelle. Parmi les diverses pratiques, l'une est primordiale dans le domaine sémantique, c'est l'annotation. Dans ce projet, les travaux d'annotation sont réalisés par ordre de difficulté croissante : on commence par les annotations manuelles ; puis on exécute les annotations plus automatiques à l'aide des dictionnaires non flexibles et flexibles ; enfin, on fait l'appel de JAPE (Java Annotation Patterns Engine) afin de rédiger les grammaires spécifiques et d'extraire les informations représentatives par corpus. Quand au choix des corpus, ce projet a sélectionné un mémoire de l'auteure américaine Tara Westove *Educated* publié en 2018 et un chef-oeuvre de Victor Hugo *Notre-Dame de Paris* paru en 1831. Westover raconte le dépassement de sa famille survivaliste Mormon pour aller à l'université, et souligne l'importance de l'éducation pour alarger son monde. Ce corpus anglais contient beaucoup de contenu narratif et il est parfait pour annoter. Sur les deux corpus, cinq premières chapitres ont été choisies à annoter. En plus, on continuera à voir les avantages et les faiblesses d'utiliser les annotations manuelles, celles avec dictionnaire et celles avec JAPE. À fin de ce rapport, on aura une conclusion globale face à ce travail annoté.

1 Annotations manuelles

La partie des annotations manuelles se base sur des schémas écrits en format xml (Extensible Markup Language). Dans ce projet final, trois shcémas d'annatations sont créés pour les deux corpus. Dans chaque fichier, il y a des sous-catégories : les dates (date, time, date time) ; les ChapitreElements (numéroDeChapitre, titre, subtitre, dialogues) ; les locations (region, airport, city, country, province, other).

1.1 Choix des annotations

Les items qu'on a choisi à annoter dépendent de la répartition assez équilibre entre les deux corpus. Pour les locations et les Dates, il est plus pratique de profiter des résultats d'ANNIE (A Nearly-New Information Extraction System). La manière manuelle a en plus une avantage de vérifier les entités nommées notée par GATE.

Dans ce cas, la méthode regex a été pratiquée dans le but de trouver des exprssions, par exemple, « aujour'hui » (Date) répété dans le corpus français, « the vallay » (Location) dans le corpus anglais. Plus avancé, un patron « ".+ ?" » (dialogues de ChapitreElements) a écrit en vue de trouver les conversations marquées par les guillemets en anglais. Parfois, les petites erreurs se sont glissées puisque cette expression contient aussi les mots ou les locutions mis en relief, comme "head for the hill" ou "Mozart". Néanmoins, ce patron n'est pas compatible avec la formatage du corpus français où se trouve des sauts de ligne définis par la signification de nouvelle ligne.

1.2 Intérêt et limites du travail d'annotation

L'atout de cette méthode manuelle revient à la généralité pratique. Le schéma n'a aucun contraint face aux langues différentes. De plus, moins d'automatisation, plus d'espace libre à opérer d'un côté d'utilisateur.

Cependant, cette opération est laborieuse et épuisante. Les temps d'annotation se sont multipliés par la longueur des textes. Dans ce projet, seuls dix chapitres ont été annotées. Si l'on a un grand data textuel, il serait impossible de finaliser cette étape à la main.

2 Utilisation des dictionnaires

Les gazetters rendent la tâche plus automatisée par rapport à l'étape précédente. En vue de bien différencier les dictionnaires, deux sets sont créés séparément : l'un est intitulé « annotations avec dictionnaires non-flexibles », l'autre est regroupé dans « annotations avec dictionnaires flexibles ».

Les deux corpus partagent cinq dictionnaires non flexibles : festival, lieu, person, title, year. En raison de GATE Morphological Analyser, on utilise seulement le dictionnaire flexible sur le corpus anglais qui comporte les verbes « look » et « say ».

2.1 Application des dictionnaires non-flexibles

Grâce aux fichiers .lst et .def, le travail permet d'introduire des sections automatisées dans lesquelles les mots sont prédéfinis. L'application des dictionnaires non flexibles ont annoté d'une manière satisfaisante la plupart des expressions du corpus.

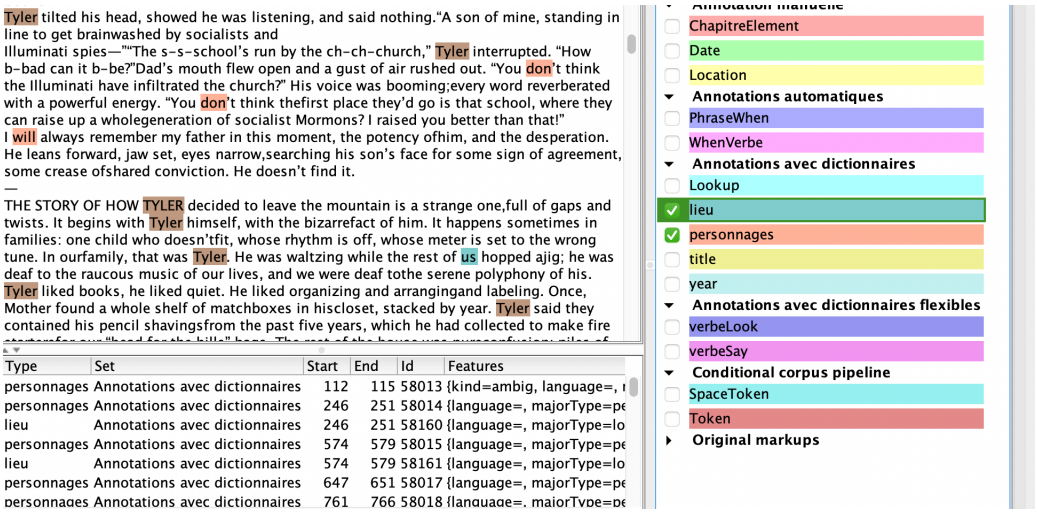

Figure 1 : Les doublons annotés par les dictionnaires Lieux et Personnages

Toutefois, la faiblesse vient du chevauchement entre les listes des mots. Par exemple, « Tyler » pourrait être un nom ou un lieu (voir Figure 1). Ensuite, les mots polysémiques engendrent les annotations fausses : « don » est jugé comme une personne ; l'abréviation « us » des États-Unis devient le pronom personnel.

2.2 Les dictionnaires flexibles par rapport à leurs homologues non flexibles

Les dictionnaires flexibles ont une très bonne performance en traitement les verbes conjugués. Et encore, le trait des mémoires correspond mieux à une fonction qui recherche des verbes au lieu des adjectifs qualitatifs. Et l'application flexible sert évidemment mieux que celle non flexible. Mais sa lacune est assez claire, c'est que la grande partie est construite sur les résultats annotés par les dictionnaires non flexibles. S'il existe déjà des erreurs à la partie non flexible, les variants du mot vont hériter les erreurs automatiquement.

3 Utilisation des JAPE

La grammaire a un grand champs d'applications. Dans ce projet, la fonction est pratiquée afin que les contenus narratifs soient marqués, particulièrement les passages avec le mot-clé « When/While » ou « Quand/Lors ». Après cette étape, les informations plus détaillées sont repérées, comme les personnes qui agissent, les actions qui sont procèdent et enfin la situation où se passe l'histoire.

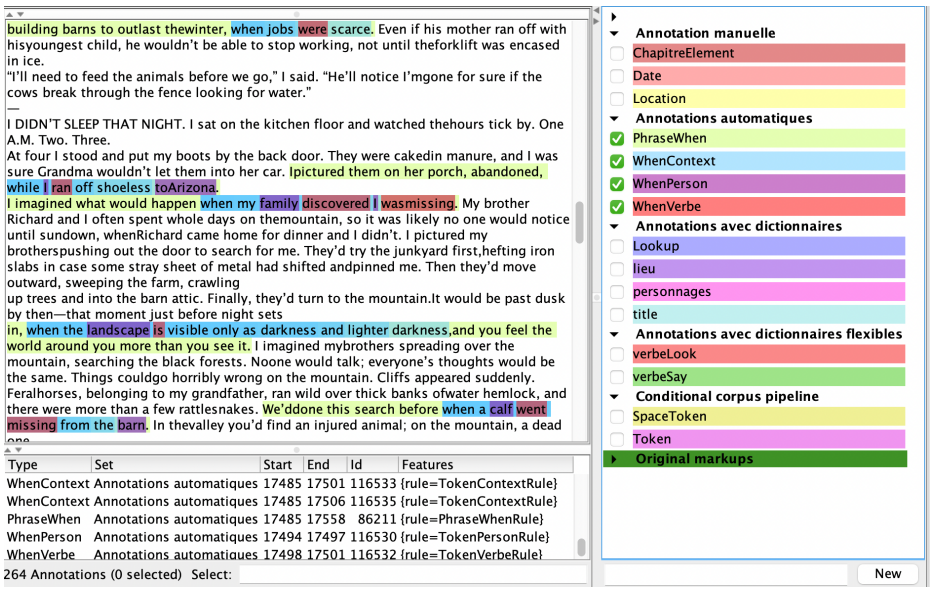

Figure 2 : Les personnages, les verbes et les contextes extraits dans l'histoire racontée

Dans le corpus anglais, la majorité textuelle raconte des histoires. La fréquence des phrases commençantes par « When/While » constate la faisabilité d'être annoté par JAPE. En tout cas, cette moyen reste l'opération le plus efficace et la plus puissante d'extraire des informations

3.1 Limites des grammaires JAPE

La grammaire JAPE est difficile de sorte que son écriture exige de l'utilisateur une bonne compréhension du texte, du logiciel gate et du langage java. L'utilisation de jape à ce seul niveau risque de générer de bien nombreuses contraintes. En outre, la seule façon de vérifier la grammaire JAPE est de l'importer sur GATE et de le traduire en effectuant un jape Transducer, ce qui est trop complexe et prend trop de temps.

Conclusion

Au bout du compte, l'exéctution sur l'application est d'autant plus compétente lorsque la difficulté d'apprentissage augmente. On a analysé trois méthodes dans trois sections. Chacune de ces trois méthodes a ses propres forces et faiblesses. Personnellement, c'est irréalisable qu'une seule de ces méthodes permette d'obtenir les meilleurs résultats en matière de traitement de texte.

Comme je l'ai mentionné dans l'annotation manuelle, les trois méthodes d'annotation devraient être croisées, c'est-à-dire qu'un objectif d'extraction ENR devrait être atteint en utilisant les trois méthodes simultanément, puis les résultats des trois devraient être comparés. Ce n'est que de cette manière que nous pouvons maximiser l'utilité du portail et minimiser les erreurs.

De plus, French NER est loin d'être aussi précis que celui d'ANNIE, et bien que gate prenne en charge les textes multilingues, il laisse encore beaucoup à désirer lorsqu'il s'agit des langues non anglaises.