Maude ANTOINE

Master TAL

Rapport d'Alternance

[Contrat d'alternance effectué du 21/09/2020 au 10/09/2021]

Lettria

21 rue de Berri 75008 Paris

Gestion, maintenance et amélioration des jeux de données.

Participation à l'élaboration de la plateforme d'annotation.

Sous la direction de :

M. MARIAN SZCZESNIAK

Soutenu le 25/06/2021 à l'UFR Phillia

Université Paris Ouest Nanterre

200 Avenue de la République 92001 Nanterre cedex

Année Universitaire 2020-2021

Remerciements

J’aimerais remercier les personnes qui m’ont aidée tout au long de mon parcours et lors de cette année d’alternance.

Tout d’abord, je voudrais remercier Serge FLEURY qui durant ma troisième année de licence de sciences du langage m’a donné mon premier cours de Perl et m’a orientée vers ce Master. Je tiens également à remercier toute l’équipe enseignante du Master TAL pour tous les enseignements et l’encadrement malgré les conditions difficiles auxquelles nous avons dû faire face au cours de ces deux années.

Je tiens à remercier le tuteur de mon alternance chez Lettria, Marian SZCZESNIAK pour sa disponibilité, son aide et sa compréhension qui m’ont permis d’avancer énormément, de m’épanouir et de me projeter dans ce domaine.

J’aimerais également remercier l’ensemble de l’équipe de Lettria pour leur superbe accueil et leur esprit d’équipe. Merci à Charles BORDERIE de m’avoir recrutée, me permettant de faire partie de ce magnifique projet et de cette incroyable équipe.

Enfin je tiens à témoigner ma reconnaissance à mon entourage, mes parents et toute ma famille qui m’a soutenue et conseillée tout au long de mes études. Merci à Sophie, Daria et Laurianne sans qui ces deux dernières années n’auraient pas été les mêmes. Elise qui m’a aidée et soutenue depuis tant d’années jusqu’à la rédaction de ce rapport.

TABLE DES MATIÈRES

Introduction

Dans le cadre du Master 2 Traitement Automatique des Langues parcours Ingénierie Linguistique de l'Université Paris Nanterre, j'ai commencé en septembre 2020 un contrat d'apprentissage au sein de la start-up Lettria.

J'ai choisi ce parcours afin de pouvoir continuer les études tout en ayant une première expérience professionnelle réelle dans le milieu du traitement automatique des langues. En effet, avant d'entamer cette année je n'étais pas sûre des possibilités d'emplois après le Master. Le fait de devoir déjà chercher une entreprise m'a fait me rendre compte de l'état du marché du travail dans ce domaine. Cette expérience m'a permis de m'y confronter mais également de me projeter dans des projets de recherches en entreprise pour la suite.

Les missions de cette alternance sont les suivantes : gestion, maintenance, amélioration des jeux de données Lettria et participation à l'élaboration de la plateforme d'annotation. Au cours de cette expérience j'ai pu participer au développement de l'interface de programmation (API) et à plusieurs projets clients qui m'ont permis d'apprendre à gérer de vraies situations de responsabilité et d'engagement auprès d'un client. J'ai beaucoup travaillé pour réunir de la donnée et la catégoriser de la meilleure des manières afin d'enrichir la base de données Lettria.

Un des projets les plus importants auxquels j'ai participé est la désambiguïsation lexicale. En effet c'est un grand enjeu du traitement automatique des langues qui reste à maitriser afin de pouvoir parvenir à ce que les machines comprennent totalement la langue. Cette tâche va permettre d'améliorer énormément d'applications de l'API de Lettria et implémenter un tout nouveau module qui n'existe pas encore.

Dans ce rapport, nous verrons les missions que j'ai eu l'occasion de réaliser au cours de cette alternance. J'aborderai pour chacune d'entre elles la méthodologie adoptée, les difficultés ainsi que les solutions et les résultats finaux. Dans un second temps je répondrai aux questions suivantes : Comment la désambiguïsation intervient dans TAL ? Comment est-elle mise en place chez Lettria ? Je détaillerai ces questionnements et mon rôle dans leur réalisation au sein de l'entreprise.

I. Présentation de Lettria

Lettria est une start-up fondée en 2017, spécialisée dans le Traitement Automatique du Langage Naturel. Le projet a commencé par le constat que les techniques de reconnaissance du langage naturel disponibles sur le marché n'étaient pas suffisamment développées pour appréhender le français en profondeur. Les plus grands projets et les recherches majeures étaient et sont toujours effectués sur l'anglais et ensuite adaptés aux autres langues, dont le français.

Les trois fondateurs de l'entreprise ont donc décidé de créer eux-mêmes un outil pensé pour le français, leur langue maternelle. Tout a été repris à la base afin de mettre en place de nouvelles méthodes complètement centrées sur le français. Aujourd'hui Lettria propose une interface de compréhension du français de haut niveau qui reste en amélioration constante avec de nombreux projets ayant pour but de faire avancer l'analyse profonde du français. Lettria met à disposition cette API qui est facile à prendre en main et à utiliser par les développeurs. Par ailleurs, elle propose également les services de son équipe pour réaliser des projets spécifiques.

La tâche principale actuellement au sein de l'entreprise est la structuration des données. En effet la compréhension du langage était la première étape qui a été développée pendant deux ans. Maintenant que celle-ci est maitrisée, la structuration des données extraites est l'étape suivante. La méthode se base sur des algorithmes qui vont structurer les données et les placer dans des bases de données afin d'en extraire de la valeur. Le but de l'entreprise est de fournir un outil clair et facile à comprendre, les algorithmes sont donc compréhensibles et toutes les informations livrées sont explicables.

Un point important chez Lettria est de respecter les données personnelles en suivant les dispositions du règlement général sur la protection des données (RGPD). Cet aspect est vraiment capital et a fait naitre un projet entier autour du RGPD à partir de l'outil de base. 'Lettria GDPR' est un produit qui a pour but d'analyser les zones de commentaires libres dans les logiciels contenant des données personnelles (CRM, ERP, SIRH…) et de signaler des risques de non-conformités éthiques.

II. Détail des tâches réalisées

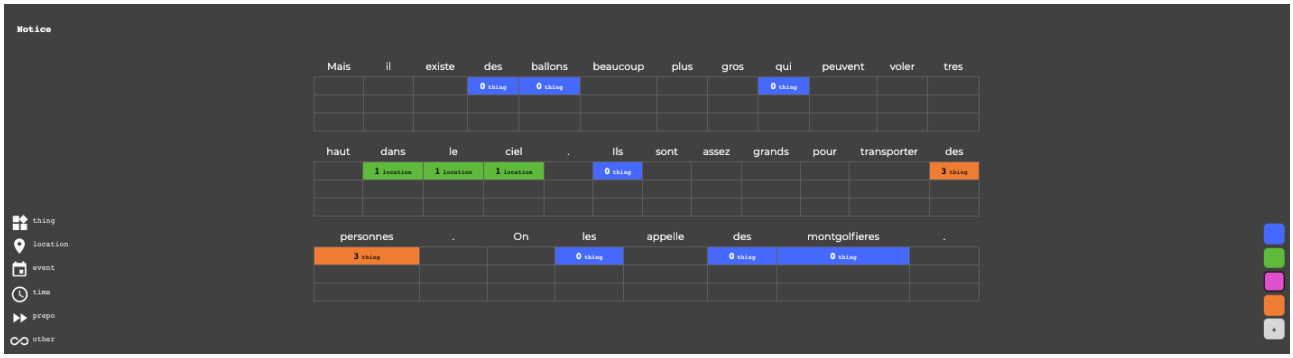
II.1. Missions continues tout au long de l'alternance

II.1.1. Annotation de coréférence

La première tâche qui m'a été confiée et que je continue d'exécuter tout au long de mon alternance, est l'annotation de coréférence. Cette dernière se réalise sur une plateforme qui a été créée à cet effet. Un grand nombre de données sont à annoter, elles sont organisées en paragraphes contenant plusieurs phrases qui sont tokénisées.

La coréférence est la relation entre deux termes ayant le même référent (définition Larousse), c'est-à-dire que plusieurs éléments désignent la même entité qui peuvent être des personnes, des dates, des objets, des lieux ou même des événements.

L'annotation de la coréférence a deux objectifs, dans un premier temps il faut identifier les mentions/références, et pour le faire, on attribue une couleur à un token ou à une suite de tokens. Dans un second temps on cherche les liens de coréférence entre les mentions/références qui font référence à la même entité, on les identifie en leur assignant la même couleur. Une mention est un ensemble continu de tokens, chaque mention correspond à une entité. De façon générale, les noms propres, les pronoms personnels (il, elle, me ...), les pronoms possessifs (notre, le mien ...), les pronoms démonstratifs (ceux-ci, celui ...), les dates et les groupes nominaux indiquent la présence d'une mention.



La deuxième étape est de classifier les groupes annotés avec les types suivants : *thing*, *location*, *event*, *time*, *prepo* ou *other*. *thing* indique que l'élément est une chose concrète (une personne, un animal, un objet…), *location* identifie un lieu introduit par une préposition ou non (un pays, une ville, un lieu-dit…), *event* désigne les éléments référant à une action (nominalisation verbale), *time* réfère à un élément donnant une information temporelle (adverbe temporel, date…), *other* contient tout les éléments que nous n'arriverions pas à classer. La catégorie *prepo* n'a finalement pas été utilisée car, après une première phase de test elle a été jugée trop large. Jusqu'à ce jour je n'ai pas rencontré de difficulté particulière en effectuant cette tâche.

II.1.2. Analyse et identification de problème de postagger

Ensuite je suis passée à l'analyse de données du postagger afin d'identifier les problèmes éventuels et de proposer des solutions en conséquence. On m'a fourni un document tableur contenant quatre colonnes, une première avec les mots annotés en partie du discours, une seconde contenant les annotations de référence, une troisième avec les annotations générées automatiquement par le postagger et une dernière colonne indiquant VRAI si les annotations diffèrent et FAUX si elles sont les mêmes. J'ai donc fait un premier tri en gardant uniquement les annotations différentes afin d'observer si les annotations de références étaient fausses ou celles du postagger ou même les deux.

J'ai identifié plusieurs problèmes :

- le premier était une confusion entre les noms communs et les noms propres, quelques noms communs écrits avec une majuscule sont identifiés comme des noms propres et inversement.
- Il y avait également des difficultés à identifier les participes passés lorsqu'ils sont utilisés en tant qu'adjectifs. Cette question est compliquée, je n'ai pas trouvé de

réponse à mettre en place dans un postagger mais uniquement des conseils pour bien les reconnaître tels que : si l'on peut ajouter un intensifieur comme 'très' avant ou si l'on peut le remplacer par un autre adjectif.

- Les chiffres romains semblaient également poser problème car ils étaient annotés comme des noms propre au lieu de nombre.
- Pour finir j'ai identifié des problèmes de classification de termes spécifiques, 'des' et 'de' n'étaient pas correctement tagués entre déterminant et préposition, 'tout' variait entre adjectif, déterminant, adverbe et nom, et 'que' qui pose souvent soucis entre ses natures de conjonction, pronom et adverbe. Encore une fois je n'ai pas trouvé de réponse à mettre en place pour la machine mais des conseils pour bien les reconnaître : remplacer par un autre terme de la classe, par exemple si 'de' est utilisé en tant que déterminant on peut le remplacer par un autre déterminant comme 'de beaux papillons' au singulier donne 'un beau papillon'. On peut aussi remplacer 'des' par de 'de les' lorsqu'il est utilisé en tant que préposition.

II.1.3. Revue d'erreur de parser et de lemmatizer

Une autre tâche que j'ai réalisée tout au long de mon alternance est la revue des erreurs de *parser* et de *lemmatizer*. Dans un premier temps il y avait quelques phrases à faire d'un coup puis par la suite lorsqu'une erreur qui était récurrente apparaissait, je la corrigeais et je créais des phrases qui pouvaient poser les mêmes soucis. De cette manière les données ajoutées permettent de faire évoluer le parser et de l'améliorer. J'ai dû adapter le format des arbres de dépendance, que j'avais étudiés en cours, au format *conllu* selon les données de *Universal Dependencies*[1] et également les relations de

[1] *Universal Dependencies:* https://universaldependencies.org/en/dep/index.html

dépendances de référence qui n'étaient pas toutes les mêmes puisque j'avais travaillé sur des discours oraux tandis qu'ici ce sont des données écrites.

II.2. Participation aux projets clients

II.2.1. Création de catégories contenant du vocabulaire pour les projets clients Gendarmerie et Doctolib

En parallèle, j'ai également travaillé pour préparer des projets de collaboration avec la gendarmerie et Doctolib. Pour ces deux présentations, j'ai préparé des lexiques ainsi que des synonymes exacts spécifiques aux domaines. J'ai fait des recherches afin de trouver le vocabulaire nécessaire pour remplir les catégories suivantes : l'historique médical, les examens médicaux et traitements, les maladies, l'historique chirurgical, l'historique gynécologique, le style de vie, les infractions, les drogues, les armes à feu, les armes blanches et les armes de contact, les documents officiels, les véhicules et les bâtiments. Pour le médical j'ai également fait des équivalences entre les mots spécifiques et leurs abréviations. Pour ce faire, j'ai eu accès à un document de faux rapport d'antécédents médicaux dans lequel il y avait un grand nombre d'abréviations que j'ai complété avec des recherches personnelles. Le but était de faire de la traduction de bilans médicaux, les médecins écrivant de manière assez complexe et difficilement compréhensible. Lettria proposait de classer les données de manière systématique afin que les médecins entre eux puisse relire les informations des patients et que les patients puissent comprendre les informations qui les concerne. Pour finir ce projet, j'ai réalisé sur le même format une grande liste de termes qui désignent des allergènes importants.

II.2.2. Recherche d'indice pour faire des résumés automatiques de textes médicaux

Ensuite, je suis intervenue sur un projet client qui avait pour but de résumer des textes scientifiques sur des sujets médicaux. Le résumé devait être fait partie par partie, en commençant par l'introduction, les objectifs, puis la méthodologie, le développement, les résultats et la conclusion. Le projet s'est basé sur un modèle de *machine learning* qui repère les mots clés et qui ne conserve que les phrases les plus importantes. Sur la base de mes connaissances et de textes manuellement résumés comme souhaité, j'ai cherché des traits qui pourraient améliorer les résultats du modèle. Après l'observation du corpus j'ai trouvé quelques idées :

- prendre les phrase contenant des mots du champs lexical de l'importance, dont j'ai fait une première liste dans un fichier pour tester et si les résultats étaient positifs aurait pu être enrichi,
- essayer de récupérer les dates et ressortir les phrases dans lesquels elles se trouvent,
- prendre les listes de points, en effet les listes sont souvent un bon moyen de résumer les idées pour les auteurs, ce qui fait d'elles de bons éléments à inclure dans les résumés
- récupérer les First, then (ou second) etc… Ces connecteurs logiques indiquent l'introduction d'une nouvelle partie qui va souvent être inclue dans la phrase puis plus développée par la suite.
- pour les résultats, on peut récupérer les valeurs chiffrées et les mots "result", les résultats sont clairement indiqués par le lexique et les données chiffrées sont de bons indicateurs des résultats qui sont les plus importants.
- je leur ai également conseillé de faire attention entre résultats et conclusion, et entre résultat et méthode principalement car les éléments étaient souvent mélangés entre ces parties.

Ces informations ont aidé à améliorer les résultats, malheureusement le champ lexical de l'importance étant surutilisé dans les textes à résumer, il faisait ressortir énormément de phrases qui ne contenaient pas d'informations réellement capitales.

II.2.3. Projet "avec des mots", améliorations de la lisibilité d'un texte

J'ai travaillé sur un projet client avec l'agence de communication éditoriale avec des mots. Leur projet était de créer une plateforme d'évaluation de la lisibilité d'un document textuel. Ils avaient déjà travaillé sur les exigences à observer pour qu'un texte soit dit 'lisible' et donc sur ce que nous devions observer afin de faire l'évaluation et de pouvoir faire en sorte de donner des conseils d'amélioration à l'utilisateur. Mon rôle durant ce projet était de faire le lien entre l'aspect linguistique et la réalisation technique, la partie développement du projet. Pour ce faire, j'ai participé à plusieurs réunions, tout d'abord pour la présentation de l'idée puis ensuite pour la compréhension et l'explicitation des règles à mettre en place. L'analyse était divisée en quatre niveaux, la phrase, le mot, la structure et la mise en page, pour le premier niveau il n'y avait pas vraiment de difficulté grâce aux outils développés dans l'API Lettria. En ce qui concerne les mots, nous avons échangé sur plusieurs points que je ne comprenais ou pour lesquels je n'avais pas d'idée de mise en place. Notamment pour l'identification de certains termes comme les nominalisations, les anglicismes, les termes techniques et formels, les mots désuets, abstraits et concrets. Pour tous ces éléments nous avons décidé de faire des dictionnaires en utilisant la plateforme de travail de Lettria. J'ai donc scrapé des données textuelles des sites sur lesquels étaient listés les termes que nous cherchions à identifier. J'ai également utilisé des articles de recherches sur la différenciation entre les mots qui dénotent du concret et de l'abstrait afin de construire les dictionnaires de ces deux catégories. Un autre élément important qui a posé problème était le travail sur la fréquence des mots c'est-à-

dire que nous voulions savoir si les mots utilisés dans le texte étaient des mots fréquents. Je ne connaissais pas de ressources nous permettant de réaliser cette opération mais le chercheur qui avait établi la liste de variables a proposé Lexique 3. Cette ressource est une base de données qui fournit diverses informations pour 140 000 mots du français dont les fréquences d'occurrences dans différents corpus. Après avoir inspecter les données de Lexique 3, nous avons décidé de garder la fréquence calculée sur un corpus de livre puisque l'autre possibilité était le corpus de sous-titre, mais la langue écrite est plus appropriée pour notre usage. Nous avons également gardé les lemmes et les catégories grammaticales afin de pouvoir identifier les termes et leurs occurrences selon leurs différentes utilisations. Les deux derniers niveaux ne concernaient pas réellement des aspects linguistiques et ne posaient pas de problème de compréhension. Pour finir ce projet, un développeur a mis en place toutes ces règles et ces calculs à partir d'un document sur lequel j'avais noté les actions à coder et les résultats qui devaient être donnés en sortie. Je suis également restée à sa disposition pour toutes les questions linguistiques qu'il pouvait avoir afin de faciliter la mise en place du projet.

II.3. Passage de l'API à l'anglais

II.2.1. Création d'un script python pour générer les formes des verbes irréguliers

J'ai travaillé sur le passage à l'anglais de l'API, pour commencer j'ai fait les formes fléchies des verbes irréguliers. Je devais rendre un fichier contenant l'infinitif, le présent de l'indicatif, le gérondif, le prétérit, le futur de l'indicatif, le présent conditionnel, l'impératif, le subjonctif présent, et le subjonctif passé. Au départ j'ai essayé de créer des scripts spécifiques pour les groupes de verbes suivants :

- finissant en 'w', (blow, sew)

- ayant toujours la même forme, (cost, spread, hurt)
- en ‘ee’ ou ‘ea’, (bleed, leave, feel, meet)
- en ‘ear’, (bear, swear, wear, tear)
- dont la forme fléchie se finit en ‘ght’, (buy, bring, fight, catch)
- dont les formes fléchies sont en ‘i’ ‘a’ et ‘u’ (sing, sang, sung ; swim, stink, begin)
- dont l’infinitif et le présent simple sont les mêmes, (find, abide, bend, pay)
- dont les formes fléchies finissent en ’t’, (burn, spend, learn, lose)
- l’infinitif et le participe passé ont la même forme (become, run, come)

Pour finir j’ai fait un script utilisant plusieurs csv contenants les informations regroupées mais sans offrir la possibilité d’ajouter de nouveaux verbes. Pour l’usage de la reconnaissance des participes passés, j’ai fait une liste des verbes à l’infinitif et de leurs participes passés.

II.2.2. Réunion de data pour l’anglais

Ensuite j’ai réuni de la data pour la base de données de l’anglais. J’ai commencé par faire des listes sous la forme suivante : forme fléchie, catégorie, lemme, informations spécifiques. J’ai commencé par les noms communs irréguliers en précisant leurs genres (féminin, masculin ou neutre) et leurs nombre (singulier, pluriel). Ces derniers étaient des noms qui avaient des formes plurielles particulières par exemple les noms qui ne s’emploient qu’au pluriel comme ‘trousers’ ou ‘scissors’ ou des noms qui ne marquent pas le pluriel comme ‘tuna’ ou ‘advice’. J’ai ensuite listé les adjectifs en notant la catégorie et pour certains en les évaluant sur une échelle de un à dix. J’ai également réalisé ce travail pour les adverbes et les prépositions.

La deuxième étape pour tous ces éléments a été d’attribuer les sens. Pour les verbes j’ai uniquement listé les verbes de sentiments dont les classes sont les suivantes :

- sentiment_surprise
- sentiment_bad
- sentiment_dislike
- sentiment_like
- sentiment_disgust
- sentiment_disappoint
- sentiment_joke
- sentiment_attract
- sentiment_prostest
- sentiment_suffer
- sentiment_upset
- sentiment_fear
- sentiment_worry
- sentiment_hate
- sentiment_love
- sentiment_bore
- sentiment_enjoy
- sentiment_laugh
- sentiment_anger

Dans le reste des catégories, les verbes étaient beaucoup trop nombreux pour que je puisse remplir toute la base de données. Le plus important était de lister les verbes de sentiments car ils sont les plus difficiles à différencier. Pour les conjonctions et les prépositions, il fallait classer les éléments listés dans les sens c'est-à-dire les équivalents en anglais qui permettent de regrouper les termes désignant les mêmes relations (to, towards, via, next_to…)

II.2.3. Travail sur l'anglais

II.2.3.1 Recherche de pattern pour les verbes impersonnels

Pour poursuivre ce travail sur l'anglais, j'ai aidé à trouver des schémas pour la détection des structures impersonnelles. La première étape a été de faire une liste des verbes qui sont toujours impersonnels. Ils sont peu nombreux en anglais et tout à fait similaires à ceux du français : ce sont principalement les verbes météorologiques comme 'to rain', 'to drizzle' et 'to thunder'. Il y a également des verbes, comme en français, qui peuvent être utilisés, ou pas, à l'impersonnel et ce sont ces derniers qui posent problème et pour lesquels j'ai tenté de trouver des patterns.

J'ai fait des recherches et réuni de nombreux exemples à observer afin de pourvoir noter les façons d'exprimer l'impersonnel et d'observer si des schémas se répétaient et pouvaient donc être utilisés pour reconnaitre la structure. J'ai relevé les patterns suivant :

- it + be (is, will be, was) + Verbe en 'ing' (it is snowing)
- it + Verbe en 's' (it rains)
- it + is going + Verbe à l'infinitif (it is going to hail)
- it + be + nom (it is monday)
- it + takes + groupe nominal (it takes the three of us to do that)
- it + [appears | seems] (it appears to be snowing)
- it + verbe modal + be (it could be cancer)
- there + verbe modal + be (there would be a big cake)
- there + be (there is a problem)

II.2.3.2. Recherche de distribution de l'apostrophe 's'

Ensuite j'ai également cherché des règles de distribution de l'apostrophe 's' en anglais. Cette marque peut noter la possession, un aspect temporal ou encore cela peut également être la contraction de la forme fléchie du verbe 'to be' à la troisième personne du singulier, voire une contraction de la forme fléchie du verbe 'to have' à la troisième personne du singulier. La difficulté pour la machine est donc de savoir identifier dans quelle situation tel ou tel usage est fait de cette apostrophe 's'. Encore une fois, pour trouver des règles potentielles, j'ai décidé de réunir un ensemble de phrases pour chaque cas et d'observer si des schémas se dégageaient. La première chose, la plus facile, était d'identifier les marqueurs temporels. L'apostrophe peut suivre un nombre afin de désigner une décennie ou un nom commun ayant une classification temporelle comme 'hour', 'week' ou bien 'minute'. Ensuite j'ai observé le phénomène de possession, tout d'abord il fallait noter que la marque était toujours attachée à un nom commun ou propre et que si

ce dernier était pluriel la plupart du temps le ’s’ disparaissait pour ne laisser que l’apostrophe (ex: the sailors’ ship). Le pattern est assez simple mais malheureusement il peut être confondu avec celui des verbes puisqu’on a un nom, l’apostrophe ’s’ puis un groupe nominal. Cependant la différenciation la plus compliquée est celle entre les deux formes verbales possible ‘is’ et ‘has’, les deux formes s’attachent aux mêmes partie du discours c’est-à-dire soit un nom soit un pronom. Pour les deux possibilités j’ai trouvé les patterns suivant :

‘IS’

- ’s + adjectif (she’s lovely)
- ’s + pronom objet (she’s mine)
- ’s + adverbe (it’s now)
- ’s + Verbe en ‘ing’ (she’s investing a lot of money)
- ’s + nom (it’s a car)

‘HAS’

- ’s + participe passé

En effet, la forme ‘has’ étant un auxiliaire elle ne peut être suivie que par des verbes au participe passé. Ces patterns ont permis de résoudre certaines confusions mais n’éliminent pas toutes les erreurs puisque, par exemple, les participes passés peuvent être des adjectifs et selon comment ils seront identifiés par le postagger le traitement sera différent. Également, comme je l’ai dit précédemment, le schéma nom + marque + groupe nominal est commun au verbe ‘to be’ et à la possession et peut donc poser problème.

II.4. Autres missions

II.4.1. Synonymes exacts

La seconde tâche qui m'a été confiée était de faire une liste de synonymes exacts, c'est-à-dire de mots désignant exactement la même chose, des signifiants différents ayant le même signifié. Le but de cette liste d'équivalents est de pouvoir lier chaque mot au référent le plus commun afin que l'analyse soit simplifiée pour la machine qui n'a alors plus qu'un seul référent pour plusieurs tokens. Pour réaliser cette liste on m'a donné une très large base de synonymes extraite de différents sites web. Parmi cette liste beaucoup de synonymes étaient très éloignés voire même erronés, ce qui a rendu la tâche assez fastidieuse. J'ai commencé par rédiger les premiers éléments qui me venaient en tête, puis j'ai fait des recherches sur le sujet afin d'essayer de trouver des informations qui pourraient m'aider en terme de méthodologie car cette tâche était compliquée à appréhender. Malheureusement, il n'y avait pas de ressource m'indiquant une manière de faire possible ou même fournissant une base de données déjà rédigée. J'ai donc poursuivi en inspectant le fichier tiré du web afin de trier les données et de les enrichir par des connaissances personnelles ainsi que des recherches sur des sites spécialisés et des dictionnaires.

II.4.2. Catégorisation des locutions conjonctives

Ensuite, j'ai complété la catégorisation des locutions conjonctives déjà réalisée mais pas encore mise en place. Il y a tout d'abord la classification puis on accorde un sens à chaque terme. J'ai eu accès à toutes les conjonctions déjà en base de données ainsi que les catégories dans lesquelles elles sont. J'ai ajouté des conjonctions et des locutions conjonctives et j'ai également revu les catégories et les sens. Les catégories sont basées sur celles existantes pour les

prepositions, il y a des grandes classes et il y a ensuite les sous-catégories qui sont donc les classes finales, voici la solution finale que j'ai proposée en plus d'une liste de conjonctions lemmatisées et classées :

- Manner: Imitation
- Causality: Cause | Intention | => (ce signe représente la conséquence)
- Quantity: Precise
- Accompaniment: Addition | Concordance of Circumstance | Exclusion
- Localisation: Temporal Approximation | Temporal Localisation | Origin | Interval Specification
- Condition
- Without Considering
- Opposition
- Comparaison.

III. Projet de désambiguïsation

La tâche de désambiguïsation lexicale a pour but d'améliorer de nombreuses applications du traitement automatique des langues. Il s'agit de permettre à la machine de comprendre le sens de chaque mot dans une phrase c'est-à-dire qu'il faut pour chaque élément, chaque token, associer un sens sémantique pour toutes les possibilités. Par exemple, le mot 'fraise' a quatre sens, il peut désigner un fruit, un outil utilisé par le dentiste, un vêtement du XVIème siècle et une membrane dans le corps d'un veau. La désambiguïsation intervient dans l'attribution de l'un de ces sens selon le contexte de la phrase donc dans l'exemple suivant : 'J'ai mangé une fraise' on parle du fruit mais dans 'le dentiste a utilisé la fraise' on parle bien de l'outil. Ce que les Hommes font naturellement s'avère être une tâche compliquée pour les machines. Ce travail a un double intérêt pour l'entreprise, d'une part la structuration sera beaucoup plus *rapide* et efficace, les structures de sens sont beaucoup plus *rapide* à chercher que des structures de mots pour lesquels il faudrait avoir une base données contenant tous les mots. Le but est de pouvoir, à partir du sens du verbe, obtenir la bonne structure directement. D'autre part pour le *machine learning* le sens des mots permets de réduire la dimension du vocabulaire, il existe un trop grand nombre de mots et surtout énormément de mots polysémiques, avec la désambiguïsation, on sait directement quel sens porte le mot.

Il existe déjà de nombreuses recherches sur ce sujet qui utilisent des approches différentes. Certaines font appel à des apprentissages supervisés sur la base de corpus annotés et d'autres préfèrent une approche non supervisée car la réunion d'annotations manuelles est fastidieuse surtout quand de telles quantités de données sont demandées. Il y a également eu des essais intermédiaires, avec un apprentissage semi-supervisé qui prend peu de données annotées et qui par exemple accorde un sens par défaut lorsque l'algorithme ne parvient pas à un résultat satisfaisant.

En attendant la mise en place de cette phase de désambiguïsation, Lettria propose tout de même une classification des termes. Les catégories concernent les verbes, les noms, les adjectifs, les adverbes, les prépositions et les interjections et elles sont toutes présentées dans la documentation. Cette catégorisation a été complétée grâce à un assistant d'annotation qui était un graphe de synonyme, dans lequel un mot était relié à tous ses synonymes. Quand dans un groupe des mots étaient fortement reliés entre eux (une grande densité de mots dans le graphe), on pouvait considérer qu'ils portaient un sens commun, ce qui a permis à la liste des catégories de grandement augmenter. Mais pour le moment chaque catégorie dans laquelle un terme peut être identifié apparait dans les résultats. Pour reprendre l'exemple de la 'fraise' il pourrait être classé dans plusieurs catégories, mais aucun choix ne peut être fait par l'outil.

L'approche choisie chez Lettria est de développer des systèmes de catégorisation pour chaque partie du discours afin de pouvoir passer par une grande phase d'annotations qui servira à alimenter un modèle d'apprentissage automatique supervisé. Le modèle devra détecter les catégories et s'il se trompe d'un grand écart dans le graphe (un écart se compte comme les degrés dans les arbres généalogiques c'est-à-dire qu'on compte un degré pour chaque mouvement vers le haut ou le bas) il sera pénalisé. Il est donc très important durant l'annotation d'assigner la bonne catégorie mère avant d'assigner la bonne catégorie fille. Actuellement la première phase d'annotations a déjà démarré sur les verbes et nous pousse à envisager des améliorations de la plateforme et de la méthode en général.

Ma mission, pour aider à la mise en place de cet ambitieux projet, se concentre sur l'aspect linguistique. J'interviens sur la création du système de classification ainsi que sur la longue phase d'annotations manuelles des données.

III.1. Graphe des verbes

La première étape a été la création des catégories de verbes. Ces dernières ont été réalisées sous la forme de graphe avec l'outil en ligne WiseMapping par un lexicographe. Ce travail a pris beaucoup de temps et ne s'est pas fait en une seule fois. Il était très compliqué de réunir tous les sens possibles de tous les verbes. La tentative initiale a été faite par le lexicographe seul mais la tâche était trop dense et il était vraiment difficile de trouver une méthodologie lui permettant de réunir tous les actions possibles. Ce sont donc des développeurs qui ont repris le travail et ont avancé en se basant sur de grandes bases de données et en ayant un point de vue plus technique sur le résultat recherché. Finalement, c'est en réunissant les deux que le graphe des verbes a réussi à voir le jour. Ce graphe a été très long à établir et vérifier qu'il soit assez complet a pris également énormément de temps. Ensuite une phase de test a eu lieu afin de faire les changements nécessaires.

Le but du graphe des sens des verbes est d'uniformiser l'annotation, de faire en sorte qu'un verbe annoté corresponde à la catégorie à laquelle il appartient, mais aussi à toutes les catégories mères. Le graphe part d'une division entre ACT et INTERACT donc entre les verbes d'action et d'interaction. ACT correspond à une action s'appliquant par elle-même et INTERACT correspond à une action s'appliquant sur quelque chose ou quelqu'un. Le système est construit en catégories miroirs des deux côtés du graphe, par exemple on a sous ACT 'exist' et sous INTERACT 'make_exist' d'un côté on trouve les verbes comme vivre, être ou commencer et de l'autre plutôt créer, concevoir ou causer. Toutes les catégories ne peuvent pas être faites de cette façon, de chaque côté du graphe on trouve des catégories qui n'ont pas d'équivalent de l'autre côté. Par exemple 'being' qui est dans 'ACT' avec des catégories finales comme 'be_called' (s'appeler, se nommer) ou 'replace' (remplacer, suppléer).

Aujourd'hui une de mes tâches est l'annotation des verbes et pour cela une plateforme a été créée. Une phrase se présente avec les verbes surlignés. Parfois nous

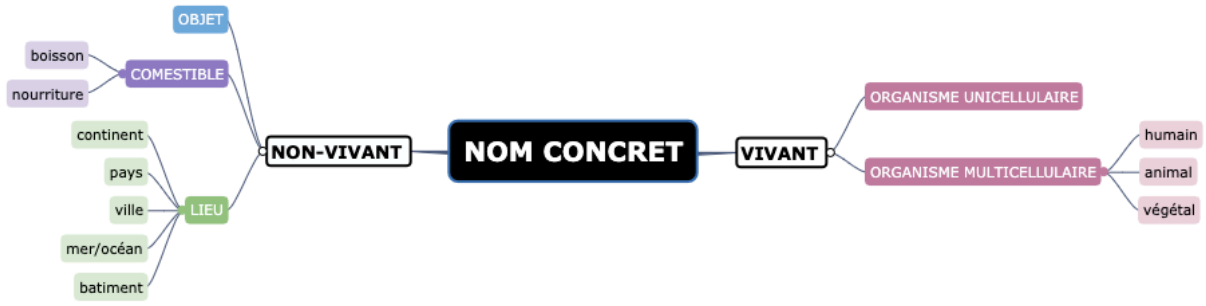
avons des erreurs de pré-traitement comme un verbe mal annoté ou un participe passé qui est surligné alors que son usage est en fait celui d'un adjectif et nous pouvons signaler les phrases mal annotées. La tâche est très compliquée car il faut parcourir le graphe, qui est très grand, pour trouver la bonne catégorie. Pour aider il y a une barre de recherche qui permet de chercher parmi les définitions et les exemples mais malgré cet outil l'annotation prend du temps et une grande réflexion.

III.2. Graphe des noms communs

La tâche suivante était la classification des noms communs et, pour ce faire, j'ai collaboré avec le lexicographe. Les noms communs se divisent en deux grandes catégories, les noms concrets qui réfèrent à des choses du réel et les noms abstraits qui réfèrent à des idées, des concepts. Nous avons donc réparti le travail en faisant chacun un graphe pour l'une de ces grandes catégories. J'ai créé le graphe des noms concrets et mon collègue celui des noms abstraits.

Pour commencer, j'ai fait des recherches sur le sujet de la classification des noms communs, puis j'ai trouvé plusieurs bases de données structurées comme *wikidata*, *DBpedia* ou encore *freebase*. Celles-ci m'ont aidée durant toute mon avancée, lorsque j'avais des hésitations ou des difficultés à réunir certains termes je pouvais comparer mes idées avec ce qui avait déjà été fait. Après ces observations, j'ai décidé de noter mes

premières idées de grandes séparations des catégories. La division de départ, à laquelle j'ai pensé très tôt, était entre le vivant et le non-vivant. J'avais réparti les éléments de cette manière :



En partant de ce premier constat très simple, j'ai développé chaque catégorie. J'en ai créée des nouvelles mais je me trouvais face à de nombreux problèmes. Après la présentation de mon travail, avec mon collègue nous avons décidé de changer la première division entre le 'NATURAL' c'est-à-dire les choses qui sont créées naturellement et celles fabriquées par l'Homme soit le 'FABRICATED'. Du côté 'FABRICATED' le comestible, qui désigne tout ce qui se mange, est devenu la nourriture afin d'exclure les fruits et légumes qui étant des végétaux doivent être dans le côté naturel. Je n'ai ajouté qu'une seule grande catégorie pour les produits chimiques sinon je n'ai fait que des sous catégories. De l'autre côté tout a changé, la première catégorie 'organisme multicellulaire' a été supprimée, j'ai mis les sous-catégories directement en catégories filles du 'VIVANT'. Je me suis basée sur les catégories existantes dans la documentation de Lettria pour créer des catégories mères pouvant les lier. De cette manière j'ai pu établir, au fur et à mesure, un fichier qui montre les équivalences, les regroupements ou les divisions entre les nouvelles et les anciennes catégories. Par exemple, la catégorie 'wine' existait déjà et a été conservée telle quelle dans le graphe, mais les catégories existantes 'drug' et 'psychotropic_drug' ont été réunies sous 'recreational_drug'. La catégorie 'drink' est complètement à revoir et à disperser car il y a un grand mélange de boissons qui

correspondent à de nouvelles catégories. Ce document a pour but de faciliter la création du graphe dans le système de l'API.

III.3. Classification des adjectifs

Ensuite j'ai également revu la catégorisation des adjectifs. Pour cette partie ,je n'ai pas fait de représentation sous forme de graphe mais uniquement une liste de catégories. J'avais déjà travaillé avec certaines de ces catégories lorsque j'avais réuni des données pour l'anglais puisque j'avais trouvé des adjectifs pour les quinze premières catégories. Je savais donc déjà que, par exemple la quinzième catégorie 'Trait' était très large et pouvait poser problème pour la désambiguïsation. Pour réaliser cette tâche je suis partie de la base existante suivante :

1', 'Anterior'	10', 'Shape'	19', 'Density'	28', 'Position'
2', 'Posterior'	11', 'Beauty'	20', 'Flavor'	29', 'Productivity'
3', 'Difficulty'	12', 'Color'	21', 'Hearing'	30', 'Religion'
4', 'Judgement'	13', 'Happiness'	22', 'Importance'	31', 'Sight'
5', 'Speed'	14', 'Feelings'	23', 'Justice'	32', 'Texture'
6', 'Temporality'	15', 'Trait'	24', 'Logic'	33', 'Thermal'
7', 'Time'	16', 'Agriculture'	25', 'Material'	34', 'Validation'
8', 'Quantity'	17', 'Civility'	26', 'Movement'	35', 'Visual'
9', 'Scale'	18', 'Condition'	27', 'Olfactive'	36', 'Weather'

J'ai également eu la liste complète du contenu de toutes ces catégories de manière à pouvoir évaluer lesquels étaient inutiles ou confuses. J'ai pu par exemple voir que la catégorie 'Agriculture' ne contenait que trois termes qui n'avaient aucun rapport avec ce domaine et j'ai considéré qu'elle était trop restreinte pour être conservée dans la

catégorisation finale. J'ai décidé de supprimer les huit catégories suivantes : 'Agriculture', 'Civility', 'Condition', 'Density', 'Justice' , 'Material', 'Sight' et 'Validation', j'ai jugé qu'elles ne contenaient pas assez d'adjectifs ou alors qu'elles étaient trop redondantes par rapport à d'autres catégories que j'ai préféré garder. J'ai également créé de nouvelles catégories qui permettent de diviser des catégories comme 'Trait' ou 'Scale' qui n'étaient pas assez restreintes.

Cette catégorisation des adjectifs est aussi très importante pour la classification des noms communs. J'ai établi des règles de transformation des nominalisations adjectivales afin de pouvoir retrouver à partir l'adjectif à partir du nom afin de classer le nom dans la catégorie de l'adjectif qui lui correspond. Ce type de noms est dispersé dans plusieurs branches du graphe des noms abstraits. J'ai réalisé un travail similaire pour les noms d'actions, j'ai utilisé les nominalisations verbales pour établir des règles de transformations afin que l'on puisse attribuer à ces noms la catégorie du verbe correspondant.

III.4. Suite du projet envisagée

Par la suite j'ai pris en charge une catégorie du graphe des noms communs assez compliquée. C'est la partie des noms de qualités, c'est-à-dire que ce sont les noms qui correspondent à des adjectifs comme 'la célébrité' qui correspond à 'célèbre'. Comme dit précédemment, j'ai établi un certain nombre de règles permettant de remonter à la catégorie de l'adjectif. Nous avons tout de même décidé de classer les noms les plus fréquents. Quelques concepts sont encore assez difficiles à placer dans le graphe, notamment les noms qui désignent des odeurs tels que 'puanteur' ou 'parfum' mais aussi les noms de couleurs. En effet ces derniers peuvent être utilisés comme des qualités de choses concrètes mais peuvent également référer à eux-mêmes. Par exemple on peut dire 'Le bleu de sa robe fait ressortir ses yeux' et dans ce cas on comprend qu'il faut attribuer la couleur au vêtement. On peut également dire 'Le bleu est ma couleur préférée'

dans ce cas la couleur n'est attribuée à aucun élément de la phrase, c'est une référence générale car la plupart des gens savent l'identifier. Nous retrouvons donc la catégorie 'color' sous plusieurs catégories mères. Ceci ne devrait pas être possible et cette question n'est pas encore résolue.

Pour la suite nous allons bientôt entrer dans la phase de test des graphes de noms communs et continuer la classification des derniers éléments. Il faudra passer par une longue phase d'annotation afin de réunir les données nécessaires pour que le modèle fonctionne de manière convenable. Au vu de la longueur de la tâche d'annotation nous cherchons des méthodes qui permettraient de faciliter le processus.

Conclusion

Cette expérience m'a beaucoup apporté malgré un début un peu compliqué car la situation m'a contrainte à être immédiatement en télé-travail. Pour une première immersion dans le milieu, ce n'était pas le départ idéal mais les missions qui m'ont été confiées étaient adaptées et nous avons décidé d'attendre que je puisse venir au bureau pour aborder certaines choses. Finalement, quand j'ai pu rencontrer mes collègues, le fait d'interagir et de pouvoir échanger m'a permis de découvrir complètement l'entreprise, de réellement comprendre sa dynamique et de m'intégrer à l'équipe.

Au cours de mon alternance j'ai acquis de nombreuses connaissances sur l'organisation et la gestion de projet. J'ai participé à de nombreux travaux qui m'ont aidée à prendre confiance en moi tout en me permettant de mettre à profit les apprentissages du master. J'ai eu l'occasion de prendre part à un projet dès son commencement en étant présente activement et en ayant un rôle important pour le bon déroulement du projet. Cela m'a apporté un sens des responsabilités en me plaçant dans une situation réelle de travail, ce que je recherchais en étant en alternance. J'ai également pu participer au projet de la désambiguïsation, principalement sur le graphe des noms concrets. J'ai apprécié bénéficier de la confiance nécessaire pour être en charge d'un pan complet de ce grand travail exploratoire.

Grâce à la spécificité de mon entreprise dans le domaine j'ai découvert, à travers les projets clients, de nombreuses applications du TAL que je n'avais pas envisagées. En effet Lettria étant un API les usages sont variés et mon travail m'a permis d'aider à l'amélioration de l'outil en lui-même ainsi que de participer à des projets très concrets.

J’ai pu voir comment les sujets appris théoriquement en cours s’appliquent concrètement et j’ai approfondi par moi-même, en faisant des recherches personnelles, certains aspects dont j’avais l’usage au cours de mes missions. J’ai pu mettre en application des savoirs autant depuis mes licences que du master puisque la licence d’Anglais m’a donné l’occasion d’aider au passage à l’anglais de l’API.

Cette année m’a beaucoup éclairée sur la suite que j’envisage de donner à mes études, en effet j’aimerais m’engager dans un doctorat. Cela me semblait inenvisageable au début du master 2, mais j’ai découvert la possibilité de faire de la recherche tout en étant dans un environnement d’entreprise et donc de pouvoir diversifier ses activités.

Références bibliographiques

Ferrand, L. et Alario, F.-X. (1998). Normes d'associations verbales pour 366 noms d'objets concrets. *L'année psychologique*, 98(4), 659–709.

Ferrand, L. (2001). Normes d'associations verbales pour 260 mots « abstraits ». *L'année Psychologique*, 101(4), 683–721.

Goriachun, D., et Gala, N. (2020). Identifying Abstract and Concrete Words in French to Better Address Reading Difficulties. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)* (pp. 33-40).

Bonin, P., Méot, A., Aubert, L.-F., Malardier, N., Niedenthal, P. et Capelle-Toczek, M.-C. (2003). Normes de concrétude, de valeur d'imagerie, de fréquence subjective et de valence émotionnelle pour 866 mots. *L'Année psychologique*, 103(4), 655–694.

Navigli, R., & Lapata, M. (2009). An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE transactions on pattern analysis and machine intelligence*, *32*(4), 678-692.

Schwab, D., Goulian, J., & Guillaume, N. (2011). Désambiguïsation lexicale par propagation de mesures sémantiques locales par algorithmes à colonies de fourmis. *Traitement Automatique des Langues Naturelles*, 185.