Université de Montréal

IFT6285 – Traitement automatique des langues naturelles

Devoir 10

---

par:

Bassirou Ndao
(0803389)
Eugénie Yockell
(20071932)

Date de remise: 13 décembre 2020

Dans ce devoir on s'intéresse à l'extraction d'information. On créer un extracteur à l'aide de la librairie `Spacy` pour obtenir des extractions de type *s IS A r* sur la première phrase de la page Wikipedia de différents sujets.

On débute par définir un patron à l'aide des analyses en dépendance pour trouver la racine *r* d'une phrase reliée à un verbe qui est relié à un sujet selon une relation *nsubj*. Nous avons étudier des phrases en anglais où les relations sont définie un peu différemment qu'en français. Par exemple, avec les phrases *Jarvis Cocker est un chanteur* et *Jarvis Cocker is a singer*, les modèles `Spacy` anglais (*en_core_web_lg*) et français (*fr_core_news_sm*) retournent différentes relations. On remarque que le modèle français considère la racine comme étant *chanteur* tandis que la racine est *is* avec le modèle anglais. Dans notre cas, on s'intéresse alors aux éléments anglais *nsubj*, *ROOT* et *attr*.

Français		**Anglais**	
Jarvis	nsubj → PROPN	Jarvis	compound → PROPN
Cocker	flat:name → PROPN	Cocker	nsubj → PROPN
est	cop → AUX	is	ROOT → AUX
un	det → DET	a	det → DET
chanteur	ROOT → NOUN	singer	attr → NOUN

Nous avons défini deux façon de faire l'extraction. La première consiste à simplement utiliser les tags du modèle de `Spacy`. On extrait les *coumpound* (noms composés) avant le *nsubj* (sujet nominal), le *nsubj* ainsi que le *attr* (attribut au verbe). On utilise alors des relations très simplistes. Un tel extracteur fonctionne plutôt bien avec des phrases simples. Pour faciliter la tâche, on retire les éléments entre parenthèse. Par contre, cet extracteur perd beaucoup d'informations comme les adjectifs. De plus, il peut extraire des résultats erronés si les tags ne sont pas bien placés dans la phrase.

Un extracteur plus complexe à alors été défini à l'aide de `Matcher` de `Spacy`. On définit des *patterns* qui sont reconnus dans la phrase passée en argument. Ces *pattern* sont plus général qu'avec la méthode précédente. Ce nouvel extracteur permet, à l'aide de différents attributs, d'améliorer nos extractions.

Le premier souci auquel on a fait face, est la possibilité de présence de plusieurs relations `ROOT`. Ainsi dans l'extracteur de base un dilemme se pose quant au choix du bon groupe à extraire. Ce problème est facilement contournée dans notre deuxième extracteur grâce à l'utilisation de l'attribut `LEMMA`. En effet, on précise ainsi que l'on recherche la racine de l'arbre dominant contenant un mot ayant pour lemme *être*.

Initialement, nous utilisions qu'un seul *pattern* à la fois pour identifier le gouvernant de la copule *est* ainsi que l'entité nommée qui lui est associée. Cela avait pour effet de limiter le type d'extraction que nous pouvions faire. En effet, lorsqu'il y avait plusieurs tokens non important pour notre extraction entre le sujet nominal et la copule *est*, notre extracteur devenait obsolète et ne pouvait pas produire l'extraction souhaitée. Nous avons donc pris la décision de configurer un `Matcher` de manière indépendante pour chaque groupe. Cette méthode paraît valide, car on considère que dans une phrase il ne peut y avoir qu'un seul groupe sujet et qu'un seul groupe verbale contenant le lemme de *être*.

Enfin, une autre particularité rencontré dans notre analyse, que ce soit pour le groupe verbal ou pour le groupe verbal, est la difficulté pouvoir récupérer tout les éléments du groupe au complet. En effet le mot gouvernant peut être:

- constitué d'un mot simple,
- constitué d'un mot composé
- constitué d'un chiffre,
- constitué de plusieurs mots séparés d'une virgule ou d'un *and/et*
- constitué de plusieurs mots liés par une préposition *of, in, on, at, etc*

En effet, nous avons remarqué qu'il arrivait souvent qu'à la suite de l'attribut *nsubj*, que l'on retrouve des informations importantes avec la préposition of. Par exemple, *Pterolophia chebana is a species of beetle.* La suite *of beetle* est très pertinente à l'extraction. On ajoute alors des règles conséquentes pour prendre en compte ces cas.

Plus précisément, pour définir les *patterns*, il est nécessairement de prendre en compte que certain mots qui peuvent apparaître aucune fois ou plusieurs fois. Par exemple, un des *pattern* définit est:

Pattern pour le groupe de *nsubj*: [det(*) compound(*) **nsubj(+)** prep (?) pobj(?)]

Pattern pour le groupe verbal: [**BE A/AN** amod(*) compound(*) **attr(+)**
OF(?) det(*) amod(*) compound(*) pobj(?)]

Où on définit (*) un mot présent 0 ou plusieurs fois, (+) un mot présent une ou plusieurs fois, (?) un mot présent 0 ou 1 fois. De plus, nous avons un *det* (déterminant) ainsi qu'un *prep* (préposition) et un *pobj* (objet d'une préposition) avant le sujet nominal (*nsubj*). Tandis qu'à la suite du verbe *être*, on peut avoir un *amod* (adjectif), un *compound* (noms composés) et l'attribut au verbe (*attr*). Comme mentionné plus haut, on ajoute aussi un *prep* (préposition) et un *pobj* (objet d'une préposition) à la suite de l'attribut au verbe. Il est aussi précisé que ce groupe prépositionnel peut contenir des adjectif, et des noms composés.

De plus, le `Matcher` de `Spacy` permet d'extraire tous les *patterns* créer. C'est-à-dire, qu'avec les conditions implémentées, on fait ressortir différents types d'extractions répondants au différentes conditions. Il était alors nécessaire d'utiliser la plus longue extraction faite, car c'est elle qui contient le plus d'informations.

Au tableau 1 on présente les extractions faites par les deux extracteurs. On peut facilement visualiser les améliorations effectuées en vert.

Phrase originale	Extracteur de base	Extracteur amélioré
Ocean Boys Football Club were a Nigerian football club based in Brass, Bayelsa State.	Ocean Boys Football Club were a club	Ocean Boys Football Club were a Nigerian football club
Calla is a genus of flowering plant in the family Araceae.	Calla is a genus	Calla is a genus of flowering plant
Keldysh is a lunar impact crater that is located in the northeastern part of the Moon.	Keldysh is a crater	Keldysh is a lunar impact crater
The BBC Regional Programme was a British radio broadcasting service which was on the air from 9 March 1930 until 1 September 1939.	BBC Regional Programme was a service	The BBC Regional Programme was a British radio broadcasting service

Table 1: Comparaison de l'extracteur de base et de l'extracteur amélioré sur quelques exemples.

À l'aide de cet extracteur, nous avons effectué 100 extractions d'informations sur des sujets aléatoires de Wikipedia. On peut étudier ces extractions pour présenter les faiblesses et forces de notre extracteur.

Lee Gong-Joo (born March 25, 1980) is a South Korean handball player who competed in the [...]	Joo is a South Korean handball player
The Mazda J-family are a range of 60-degree V6 engines featuring a cast-iron cylinder [...]	family are a range of
Joe Kodeih is a writer, actor and director.	Joe Kodeih is a writer
The Dark Tower, first published in 2007, is a series of comic books based on Stephen King's [...]	The Dark Tower is a series of comic books
Joana Hadjithomas and Khalil Joreige are filmmakers and artists.	-

On remarque alors que l'extracteur n'est pas capable de gérer les noms composés contenant un tiret -, ainsi que les nombres. Ces attributs ne sont pas ajoutés dans les *patterns* et ne sont pas reconnus par le `Matcher`. De plus, les *patterns* ne prennent pas en compte le cas où l'attribut au verbe peut être composé de plusieurs informations relié par *and*. Par contre, comme mentionné, dans le quatrième exemple, il est mesure de reconnaître les relations importantes avec le verbe et ignoré les éléments entre les virgules.

En conclusion, un extracteur de base à été amélioré afin de généré des extractions d'informations de type *IS A* pertinentes et contenant beaucoup d'informations. Une étude importante des phrases à été faite pour créer le meilleur extracteur possible. Il aurait été intéressant de prendre en compte les autres problématiques mentionnées comme les mots composés tel que ($Mots1$-$Mots2$) et les chiffres. Les *patterns* se brisaient lorsqu'ils rencontraient de tel cas, car ils n'ont pas été pris en compte. Il aurait aussi été intéressant de prendre en compte les cas où le groupe sujet ou verbal contient un *and* comme par exemple, *Keriya is a patwar circle and village* ou encore *Joana Hadjithomas and Khalil Joreige are filmmakers and artists*. L'extracteur ne prend en compte que *circle* dans le premier cas et ne reconnaît rien dans le deuxième cas. Bref, de telles améliorations produiraient un extracteur presque infaillible sur les type *IS A*.