Le génie pour l'industrie

Rapport de laboratoire

Étudiants	José Melancon (MELJ22089209) Samy Lemcelli (LEMS26019103) Philippe Grenier-Vallée (GREP02028906) Alexandre Billot (BILA29099203)
Cours	LOG635
Session	Hiver 2015
Groupe	02
Laboratoire	01
Équipe	03
Chargé de laboratoire	Richard Rail
Date	17 mars 2015

INTRODUCTION

Dans le cadre du cours LOG635 - *Systèmes intelligents et algorithmes*, il est demandé de développer un système de traitement automatique d'une langue naturelle (le Français) avec l'aide de la boîte à outils *NLTK* et le langage *Python*. Le système devra être en mesure d'analyser un texte d'une complexité lexicale et syntaxique suffisante et être capable de décoder de l'information en lien avec le système développé lors du premier laboratoire. Donc, le système doit être capable d'analyser des phrases seulement en lien avec ce contexte.

Finalement, le logiciel doit être en mesure de mettre les informations contenues dans le texte décodées dans le format JESS. Ceci pourrait techniquement faire en sorte qu'il serait possible d'utiliser ces faits dans notre système expert développé lors du premier laboratoire.

PRÉ-TRAITEMENT

1. **Séparation en lignes**
   Chacune des phrases sont séparées du texte à l'aide du point. Chaque phrase est ainsi analysée individuellement par NLTK pour être ensuite traitée.
2. **Remplacement des noms propres à la version présente dans la grammaire**
   Tous les noms propres sous leur forme dans la langue naturelle est remis sous la forme qui est présente dans la grammaire.
3. **Suppression des accents**
   Puisque la langue naturelle choisie est le français, il y a beaucoup d'accents qui ne sont pas facilement gérés par NLTK. Alors, il est nécessaire de remplacer les caractères spéciaux du français par la lettre non accentuée. Par exemple, le "è" devient un "e".
4. **Mise en minuscules**
   Tout le contenu du texte est finalement mis en minuscules pour éviter de devoir spécifier chaque nomenclature possible dans la grammaire.

COMPLEXITÉ

1. **Syntaxique**
   La principale complexité de la création de la grammaire était de limiter sa flexibilité suffisamment pour ne pas avoir d'ambiguïtés. De plus, nous avons tenté au possible de former la grammaire de façon à ce que l'arbre soit facilement traitable en Python par la suite. Par exemple, l'utilisation du verbe ***être*** dans plusieurs cas implique que l'action réelle est le mot qui suit, dans le cas du mariage on dit que "X est marié à Y". Nous avons donc tenté de construire la grammaire de façon a faire ressortir *marié* au lieu du verbe, qui ne représente rien qui nous est utile dans ce cas.

2. **Lexicale**
   La partie lexicale était relativement simple, nous avons ajouté plusieurs synonymes

aux différents emplacements qui sont représentés dans notre scénario. Par exemple, la station d'essence peut être référée par "dépanneur". Cela permet d'éviter une certaine répétition dans nos phrases en variant les mots utilisés.

3. **Sémantique**
   L'analyse sémantique faite dans notre grammaire représentait une autre partie importante du travail. Il fallait, le plus possible, positionner les informations sémantiques de façon à pouvoir facilement les transformer par la suite vers notre format JESS. Dans plusieurs cas il a été relativement simple de créer cette association, principalement quand on a une relation simple d'un sujet vers un nom, par exemple "X est l'amant de Y". Dans d'autres cas, la forme de la phrase était plus complexe et a demandé bien plus d'efforts afin de correctement former l'arbre syntaxique et la sémantique associée afin de pouvoir obtenir les informations dont on avait besoin de façon accessible.

POST-TRAITEMENT

1. **Impression de la sémantique (SEM) du sommet de l'arbre *NLTK***
   Permet de vérifier si le traitement de la ligne s'est bien déroulé. C'est cette information qui sera traitée par la suite.
2. **Remplacement des pronoms par ce qu'ils remplacent**
   Dans le but d'éliminer les pronoms de la phrase et de les remplacer par leur sens voulu, le nom de la personne qu'ils remplacent. Dans le cas d'un pronom normal (i.e.: il, elle, nous, etc..) on recherche les noms des personnes appropriées selon le genre (GEN) et le nombre (NUM), si cela est nécessaire. Dans le cas d'un pronom relatif, le dernier nom ou pronom normal utilisé le remplace.
3. **Séparation des agrégats**
   Il arrive des fois qu'une action est faite par un groupe de personnes dans les phrases. Dans ce cas, on considère l'action comme étant faite par un agrégat de personnes. Cette étape permet de d'appliquer l'action sur chacune des personnes individuellement. *NLTK* retourne dans ce cas un groupe "aggregate" (GN) contenant les différents noms impliqués. On sépare alors ces noms afin de les traiter individuellement dans la phrase.
4. **Transformation en format JESS**
   Finalement, une fois que toute l'information est présente dans la ligne, il est temps de transformer les informations dans un format JESS qui serait possible d'utiliser avec le système expert développé lors du premier laboratoire. Pour y arriver, on vérifie si un des mots-clés est présent dans la ligne et on formate la ligne en conséquence du mot trouvé. Suite à ce traitement final, la phrase de départ peut maintenant être utilisée par ce système expert.

CONCLUSION

Dans le cadre de ce laboratoire, l'équipe a été en mesure de créer un système d'analyse de langage naturel qui est capable d'extraire de l'information pertinente au monde créé lors du premier laboratoire. Tout d'abord, le logiciel prend le texte et en fait un prétraitement pour en faciliter l'analyse. Ensuite, il extrait le sens avec *NLTK* qui crée les arbres syntaxiques. Finalement, un post-traitement a lieu dans le but de mettre ces informations dans un format utilisable pour notre système intelligent, le langage *JESS.*

Malgré le fait que le contenu sémantique possible est très limité car il se limite au monde conçu lors du premier laboratoire, la complexité du système assez grande. En effet, celui-ci est en mesure de considérer plusieurs particularités de la langue: pronoms relatifs, phrases composées, vérification du nombre, phrases négatives et autres.