Projet Scientifique Collectif X2013

Proposition détaillée

24 septembre 2014

Table des matières

Notre groupe

- Fernandez-Pinto-Fachada Sarah, 8^e compagnie, section **équitation** ;
- Schrottenloher André, 8^e compagnie, section **escrime** ;
- Angibault Antonin, 8^e compagnie, section **escrime** ;
- Hufschmitt Théophane, 8^e compagnie, section **escrime** ;
- Cao Zhixing, 9^e compagnie, section **escrime** ;
- Boisseau Guillaume, 6^e compagnie, section **natation** ;

1 Présentation du sujet

Nous cherchons à mettre sur pied un analyseur syntaxique de documents rédigés en langue anglaise.

1.1 Motivation et enjeu

Nous vivons dans un âge d'abondance d'information, qu'il convient de traiter efficacement pour en profiter. Dans ce contexte, les synthétiseurs automatiques de textes ont bénéficiés d'efforts importants de recherche au cours des années passées. Notre travail s'inscrira dans cette démarche.[1]

Contrairement à un certain nombre de projets ayant abouti pour l'instant, nous espérons dépasser la simple sélection de phrases pertinentes dans un corpus plus volumineux (*extractive summarization*) pour donner à notre programme une compréhension des idées présentes dans le texte (*abstractive summarization*).

1.2 Objectif du projet

Nous cherchons à résumer un texte ou un corpus de textes sur un sujet donné. A partir de ces textes, traitant d'un thème commun, le programme devra *in fine* traduire les informations importantes en termes compréhensibles par un être humain.

Pour effectuer cette opération, nous envisageons l'approche suivante :

1. Effectuer une analyse syntaxique permettant de passer du texte écrit à un ensemble d'informations informatiquement exploitables ;
2. Traiter ces informations, par exemple sous forme de réseau sémantique, afin d'en extraire les données pertinentes ;
3. Retraduire cette information pertinente en termes simples du point de vue du langage et de la syntaxe, éventuellement sous forme de graphe.

Nous ciblerons des textes d'actualité, dont la profusion permet d'envisager une analyse statistique, en plus de fournir un moyen d'évaluer notre programme (en comparant ses résumés à des résumés existants).

Dans la mesure où des outils pour la transformation d'un texte en réseau sémantique exploitable par la machine existent, l'accent sera mis sur le traitement de l'information et la sélection de celles qui sont importantes ou sujettes à débat.

Pour cela deux méthodes sont envisageables : la première reposera essentiellement sur une analyse statistique d'un grand nombre de textes pour déterminer où sont les unités de sens et comment elles s'accordent ; la seconde sur un réseau de concepts représentant le "bon sens" de notre programme. Dans la mesure où des outils d'extraction existent déjà, nous espérons aboutir à des résultats concret en choisissant la méthode statistique ; c'est cependant la seconde méthode qui laisse le plus de place à l'innovation et que nous explorerons dans un premier temps, même si elle sera plus difficile à mettre en place.

2 Etat de l'art

2.1 Autres approches du problème

Nous prenons le problème sous l'angle de la compréhension du texte, mais en réalité, un plus grand nombre de travaux s'intéresse à l'extraction de phrases jugées pertinentes sur des critères statistiques.
Cette extraction se fera par exemple sur la base d'une analyse fine des champs lexicaux, de la syntaxe, et des occurrences de certains termes importants.

2.2 Documentation disponible

Différents outils existent et permettent de se concentrer sur les étapes-clés du projet.

- L'analyse syntaxique du texte brut peut être réalisée par un outil *open-source* déjà disponible qui intervient dans certains logiciels de traduction automatique et dont le but est de séparer le texte en unités syntaxiques ;
- Des outils existent pour gérer de manière efficace des réseaux de concepts.

3 Traitement envisagé du sujet

3.1 Etapes-clés

3.1.1 Analyse syntaxique et représentation des informations

Nous nous concentrerons dans un premier temps sur l'analyse du texte à l'aide des outils déjà existants. Le texte sera subdivisé en unités syntaxiques liées par des verbes d'action, d'état ou diverses relations (propriété, caractéristique, nature).

Il s'agira donc, d'une part, d'effectuer des recherches bibliographiques sur les divers outils à notre disposition, puis dans un second temps d'apprendre à les maîtriser.

3.1.2 Traitement des informations

Deux possibilités s'offrent à nous selon l'état d'avancée du projet :

- Traduire immédiatement cette information en réseau sémantique, puis la traiter toujours sous cette forme (cela reviendra à retirer les nœuds les plus faibles en terme de poids ou de relations)
- Utiliser une architecture plus complexe faisant intervenir un réseau de concepts, sur lequel la lecture du texte agit avant de produire un réseau sémantique comme dans le point précédent

La seconde possibilité sera explorée d'abord, s'il s'avère impossible de produire une avancée quelconque dans le temps imparti nous nous rabattrons sur la première.

3.2 Répartition des tâches

- Chef de projet, contact avec l'encadrement : Antonin Angibault
- Contact avec le tuteur : Théophane Hufschmitt
- Établissement de la bibliographie : Antonin Angibault, André Schrottenloher
- Codeurs : ?
- Obtention d'outils : Guillaume Boisseau
- autre ?

4 Références bibliographiques

Références

Références

[1] Michael Elhadad. "Natural Language Processing with Python". English. Dans : *Computational Linguistics* 36.4 (déc. 2010). WOS :000285382400009, p. 767–771. ISSN : 0891-2017. DOI : 10.1162/coli_r_00022.

[2] Mohamed Abdel Fattah et Fuji Ren. "GA, MR, FFNN, PNN and GMM based models for automatic text summarization". English. Dans : *Computer Speech and Language* 23.1 (jan. 2009). WOS :000262688100007, p. 126–144. ISSN : 0885-2308. DOI : 10.1016/j.csl.2008.04.002.

[3] Karen Spaerck JONES. "Automatic summarising : The state of the art". English. Dans : *Information Processing & Management* 43.6 (nov. 2007). WOS :000249742500004, p. 1449–1481. ISSN : 0306-4573. DOI : 10.1016/j.ipm.2007.03.009.

[4] G. SALTON et al. "Automatic text structuring and summarization". English. Dans : *Information Processing & Management* 33.2 (mar. 1997). WOS :A1997WU98800006, p. 193–207. ISSN : 0306-4573. DOI : 10.1016/S0306-4573(96)00062-3.

[5] C. N. SILLA et al. "Automatic text summarization with genetic algorithm-based attribute selection". English. Dans : *Advances in Artificial Intelligence - Iberamia 2004*. Sous la dir. de C. LEMAITRE, C. A. REYES et J. A. GONZALEZ. T. 3315. WOS :000226646200031. Berlin : Springer-Verlag Berlin, 2004, p. 305–314. ISBN : 3-540-23806-9.

A Les annexes