Projet Scientifique Collectif X2013

Proposition détaillée

24 septembre 2014

Table des matières

Notre groupe

— Fernandes-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 **escalade** ;
— 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 bénéficié d'efforts importants de recherche au cours des années passées. Notre travail s'inscrira dans cette démarche.

Un nombre non négligeable d'outils reposent, essentiellement, sur une extraction des phrases pertinentes *via* une analyse statistique et éventuellement une analyse symbolique plus ou moins poussée (*extractive summarization*). Nous voudrions, à l'instar de certains autres projets, mettre l'accent sur cette analyse de façon à donner une réelle compréhension des corpus à notre programme et, pour se convaincre de son succès, le pousser à la reformulation (*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, telle que présentée dans [3] :

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

Nous ciblerons des textes d'actualité, dont la profusion permet d'envisager une analyse statistique, en plus de fournir un moyen simple 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 concrets 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 État de l'art

D'après [3], des recherches sur la synthèse automatique de documents se sont principalement développées depuis les années 1990, motivées par des analyses statistiques remontant à 1958 []. Les méthodes d'extraction, de par la simplicité de leur mise en œuvre, ont été le plus mises en avant, utilisant dans des travaux plus récents une analyse symbolique plus poussée, menant à la reformulation de la source plutôt que de la simple extraction de son contenu (on parle alors d'abstraction). Dans cette section, nous détaillerons les techniques propres à ces deux paradigmes, puis nous indiquerons les méthodes envisageables d'évaluation de la qualité des résumés produits.

2.1 Extraction de la source (*extractive summarization*)

L'idée de la méthode est d'extraire les phrases jugées importantes dans le texte et de les concaténer dans le résumé (en éliminant les redondances), après quelques modifications de forme de façon à assurer une plus grande cohérence de l'ensemble. Dans sa versions la plus minimaliste, l'interprétation pourra se contenter de transcrire le texte à résumer (la source) en un tableau contenant le numéro de chaque phrase et une valeur quantifiant son importance, la transformation devant alors simplement sélectionner les phrases les mieux notées.

Dans ces méthodes, l'accent est donc plutôt mis sur l'étape de transformation et, dans des systèmes plus développés, l'interprétation de la source.

L'établissement de la notation, au cours de l'interprétation, se fondera généralement sur une analyse statistique (fréquence d'apparition ou d'association de termes) et éventuellement sur une analyse plus fine des champs lexicaux et de la syntaxe. Dans [4], les paragraphes les plus pertinents sont extraits du texte après établissement d'un graphe de similarité entre eux, sur la base de leur vocabulaire (précisément, des occurrences de *termes*, qui peuvent être des mots

ou des unités syntaxiques). La méthode décrite dans [4] applique en fait à l'intérieur d'un texte des techniques de récupération de l'information (*information extraction*) qui sont utilisées pour établir automatiquement des liens hypertextes entre pages web ou articles.

2.2 Méthodes non extractives

Plusieurs raisons expliquent le développement relativement faible des méthodes non-extractives, comparé à celui des premières. Premièrement, les exigences sur la qualité des résumés produits était en général plutôt faibles [3], et les méthodes extractives fournissent le plus souvent des résultats acceptables. La facilité relative de leur mise en œuvre leur a donc profité. D'autre part, les méthodes de résumés automatique se sont principalement développées par un système d'essais et de correction, mettant en avant l'apprentissage des machines et donc les statistiques.

Bien que la limite entre les méthodes extractives et non-extractives soit un peu floue, on peut remarquer que ces dernières poussent l'analyse symbolique de la source nettement plus loin lors de l'interprétation de la source. La représentation machine qui en est faite gardera un contenu sémantique (contrairement à certaines méthodes extractives) et contiendra plus d'informations que la représentation pour une méthode extractive (notamment, dans les systèmes les plus évolués, sur la structure générale de la source et de son argumentation éventuelle). Surtout, la génération sera une étape bien plus importante (au lieu de sélectionner du contenu depuis la source, le programme devra créer du contenu à partir de sa représentation transformée).

Ces méthodes sont, de manière générale, plus sensibles au contenu de la source que les méthodes extractives [3, p.1774], ce qui laisse envisager quelques difficultés à appliquer la méthode à des articles d'actualité (dont le contenu est essentiellement libre, donc varié). Nous la jugeons toutefois plus à même de tenir compte des points de désaccord entre plusieurs documents et donc plus capable de rendre compte des points de polémique dans les corpus étudiés.

2.3 Évaluation

L'évaluation des programmes de résumé automatique n'est pas aussi évidente qu'elle le paraîtrait au premier abord. En effet il s'agirait, dans l'idéal, de qualifier le programme indépendamment du contenu ou de la forme des sources fournies, et de l'objectif du résumé, qui est pourtant sujet à de grandes variations (typiquement, d'un ou plusieurs paragraphes synoptiques à une liste simple d'actions à mener automatiquement en cas d'incendie). Trois méthodes principales ont été retenues par les programmes récents d'évaluation (à partir des années 2000) [3, p.1453-1461] :

— Qualité du texte et qualité du discours. Pour des résumés sous forme de texte en langage naturel, il s'agit d'une part de vérifier la justesse grammaticale du résumé, ainsi que sa cohérence générale à plus grande

échelle. Si les propriétés locales sont assez simplement vérifiables, c'est assez loin d'être le cas pour le discours en général.
— Capture du concept. Il s'agit de vérifier que les informations centrales de la source sont également présentes dans le résumé. Pour cela, une méthode assez prometteuse est de définir un certain nombre de questions sur la source auxquelles on devrait pouvoir répondre en ayant lu seulement le résumé. Il est concevable de définir ces questions à partir de questions de compréhension de texte basiques (du genre de celles qu'on donne en primaire) ou des questions un peu plus poussées sur la structure de l'argumentation. Cependant les questions auxquelles on doit pouvoir répondre restent essentiellement fonction de l'objectif du résumé et ne permettent donc pas d'évaluer le système en toute indépendance.
— Comparaison à un modèle. L'idée ici n'est pas de comparer le résumé à sa source mais à d'autres résumés établis comme bons. Ces modèles sont souvent créés par des humains (dont on suppose qu'ils sont entraînés à produire de bons résumés), même si cela limite la quantité des points de comparaison. Les difficultés de cette méthode résident, pour l'essentiel, à la comparaison d'un texte à l'autre (dès lors que les phrases ne sont plus simplement extraites de la source, il faut juger de l'équivalence ou de la proximité des concepts présents dans les résumés) et au désaccord des juges (ou rédacteurs) sur ce qui est important dans un texte.

Ce dernier point illustre une difficulté fondamentale de toute évaluation : la qualité essentielle du bon résumé est de contenir les informations clés présentes dans la source, une caractérisation vague et impossible à préciser. De plus, des résumés de forme très différente peuvent servir efficacement le même objectif.

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'abord 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'avancement 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, puis 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 : Sarah Fernandes-Pinto-Fachada, Guillaume Boisseau
— Obtention d'outils : Guillaume Boisseau

4 Références bibliographiques

Références

[1] Michael Elhadad. "Natural Language Processing with Python". English. In : *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. In : *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. In : *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. In : *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. In : *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.

[6] Shiren Ye et al. "Document concept lattice for text understanding and summarization". English. In : *Information Processing & Management* 43.6 (nov. 2007). WOS :000249742500015, p. 1643–1662. ISSN : 0306-4573. DOI : 10.1016/j.ipm.2007.03.010.

A Les annexes