Nouvelle méthode d'analyse syntactico-sémantique profonde dans la lecture et l'analyse de textes assistées par ordinateur (LATAO)

Ngoc Tan Le[1], Jean-Guy Meunier[1], Louis Chartrand[1], Davide Pulizzotto[1], Jose A Lopez[1], Francis Lareau[1], Jean-François Chartier[1]

[1]LANCI, Université du Québec à Montréal, Montréal, Québec - Canada

Abstract

The conceptual mining aims to analyze in depth the content of a concept in a text. But when this concept is of a philosophical nature, the task becomes difficult because of the complexity of this type of concept. In fact, it involves a wide variety of properties and relationships and which are marked as a linguistic expression. If this mining must be assisted by computer, it will be sensitive to syntactic dimensions.This research hypothesizes that a syntactic-semantic analysis allows better use of statistical analyzers classificatory results have been applied to some contexts of a conceptual term target. The method of operation of these results consist in being applied to segments of these classes of syntactic-semantic analysis. Hence this allows researchers to discover in these sentences revealing important semantic synthetic components having dominant conceptual properties such as definition, inferences, properties, etc. The approach is tested on the concept of MIND in "Collected Papers" of CS Peirce that was re-edited in 1994. The results show that we can significantly reduce the size of a corpus by the presentation of lexical patterns that are sensitive to syntactic-semantic constraints. Such type of synthetic representation allows researchers to guide the reading to revealing specific dimensions of important conceptual properties.

Résumé

Le forage conceptuel vise à analyser en profondeur le contenu d'un concept dans un texte. Mais lorsque ce concept est de nature philosophique, la tâche devient difficile en raison de la complexité de ce type de concept. En effet, celui-ci met en jeu une grande diversité de propriétés et de relations qui se marquent dans son expression linguistique. Et si ce forage doit être assisté par ordinateur, il devra être sensible à certaines dimensions syntaxiques et sémantiques. La présente recherche repose sur l'hypothèse qu'une analyse syntactico-sémantique permet de mieux exploiter des résultats d'analyseurs statistiques classificatoires ayant été appliqués à des contextes d'une expression conceptuelle cible. La méthode d'exploitation de ces résultats applique aux segments de ces classes des analyses syntactico-sémantiques. Ceci permet alors aux chercheurs(-euses) de découvrir dans ces phrases d'importantes organisations sémantiques synthétiques révélatrices de certaines propriétés conceptuelles dominantes qui seraient liées à des définitions, des inférences, des propriétés, etc. L'approche est expérimentée sur le concept de MIND dans les "Collected Papers" de C.S. Peirce qui a été réédité en 1994. Les résultats obtenus montrent qu'on peut réduire significativement la taille de ce corpus par la présentation de patrons lexicaux sensibles à des contraintes syntactico-sémantiques. Un tel type de représentation synthétique permet aux chercheurs d'orienter la lecture vers des dimensions spécifiques révélatrices de propriétés conceptuelles importantes.

Mots clés : analyse conceptuelle; forage conceptuel; analyse syntactico-sémantique; connaissances linguistiques; patrons lexicaux; humanité numérique.

1. Introduction

Le forage conceptuel consiste à trouver un concept particulier en se basant sur des hypothèses empiriques puis à effectuer une analyse profonde des corpus textuels dans différents secteurs d'étude des sciences humaines et sociales. Dans cette stratégie, l'analyse syntactico-

sémantique vient assister l'exploration de segments qui peuvent dans certaines classes se révéler très nombreux. Ainsi la recherche de patrons lexicaux permet d'identifier et de structurer davantage certaines informations pertinentes qui s'y trouvent. Les patrons lexicaux sont des structures représentant des schémas récurrents du langage. Les patrons lexicaux sont définis comme tous les mots ou structures qui sont régulièrement associés à ce mot et qui contribuent à sa signification (Hunston et Francis, 2000). Le repérage des patrons lexicaux permet de dévoiler le sens du mot exprimant en contexte des propriétés d'un concept.

2. Méthodologie

Une chaîne de traitement LATAO (Lecture et Analyse de Textes Assistées par Ordinateur), appliqué plus spécifiquement à un concept, envisage une série d'étapes permettant de lire et d'analyser des concepts spécifiques dans un corpus de texte. Traditionnellement, avant tout, il s'agit d'une classification des textes qui est définie comme une opération, appliquée à des segments de texte, permettant d'identifier des classes d'équivalence entre ces segments à l'égard de leur contenu informationnel tel que mots, n-grammes, etc.

Par ailleurs, nous constatons qu'une langue a ses propres régularités linguistiques et qu'un texte possède un sens global. D'où

> *Hypothèse:* À partir des résultats d'analyseurs statistiques classificatoires ayant été appliqués à des contextes d'une expression conceptuelle cible, nous pouvons trouver des patrons lexicaux qui représentent un sens global des segments de texte par une analyse syntactico-sémantique.

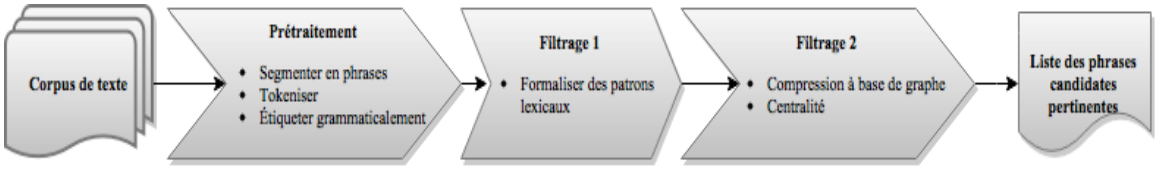

Figure 1 : Chaîne de traitement de notre méthodologie d'analyse syntactico-sémantique

En premier lieu, un prétraitement sur le corpus de texte (Figure 1) est effectué. Le corpus est segmenté en phrases. Après une étape de tokenisation, ces phrases sont étiquetées avec les catégories grammaticales telles que noms, verbes, adjectifs, adverbes, déterminants, etc. Par la suite, un premier filtrage est appliqué. Ce premier filtrage consiste à formaliser un ensemble de patrons lexicaux. Deux catégories grammaticales de noms et de verbes retiennent l'attention. Les noms représentent les agents ou les objets de l'action. Et les verbes représentent les actions possibles. Les couples nom-verbe ainsi formés présentent un intérêt applicatif notamment en recherche d'information (Claveau, 2003; Fabre et Bourigault, 2006). Les patrons lexicaux permettent de représenter sous l'aspect syntaxique tous les constituants d'une structure de phrase donnée.

Les patrons lexicaux sont constitués par deux règles suivantes en se basant sur les expressions régulières:

Règle 1: r'(nom_à_gauche).*(verbe)' *Règle 2:* r'(verbe).*(nom_à_droite)'

Par exemple, si la classe 1 avec 627 phrases dont le traitement produit une liste des termes cibles comportant '*sign, interpretant, object, determination, representation, mean, idea*' (cf. Tableaux 1 et 2), les deux règles auront été appliquées de la manière suivante:

Règle 1	Règle 2
[('**sign**', 1)] **actualize** [('object', 1)]	[('meaning', 1)] **react** [('**sign**', 1)]
[('**sign**', 2)] **address** [('word', 3)]	[('term', 2)] **annoy** [('**sign**', 2)]
[('**Sign**', 5)] **bring** [('relation', 9)]	[('development', 1)] **follow** [('**sign**', 3)]
(…)	(…)

Le deuxième filtrage permet d'extraire des phrases candidates pertinentes à partir des résultats obtenus durant le premier filtrage. Les patrons lexicaux ainsi extraits servent de prototypes pour les entrés du deuxième filtrage en utilisant des techniques du domaine du résumé automatique, la technique de compression et en nous basant sur l'approche à base de graphe (Torres-Moreno, 2014).

3. Expérimentations et Évaluations

3.1. Préparations de corpus

Le corpus philosophique choisi est celui des "Collected Papers" de C.S. Peirce dans sa réédition de 1994. Ce corpus se compose de 940 segments contenant 13 066 phrases et de 1 190 054 mots (Tableau 1).

Tableau 1: Statistiques sur tous les segments avec 18 classes

Classes	#Phrases (#Segments)	Classes	#Phrases (#Segments)
1	627 (40)	10	197 (16)
2	650 (50)	11	185 (5)
3	2 513 (171)	12	158 (15)
4	152 (14)	13	631 (39)
5	268 (15)	14	225 (21)
6	2 419 (188)	15	26 (14)
7	363 (27)	16	466 (37)
8	**3 103 (198)**	17	769 (66)
9	32 (2)	18	282 (22)

3.2. Expérimentations et Résultats

Premièrement, le prétraitement sur le corpus de test est effectué en appliquant une analyse linguistique à l'aide de l'outil de NLTK[1]. Puis un ensemble des termes cibles approprié à

[1] NTLK – Natural Language Toolkit : http://www.nltk.org

chaque classe est extrait en fonction de leurs fréquences maximales. Ces termes cibles sont des points de repères pour les deux filtrages dans la chaîne de traitement. Par exemple dans la classe 1, les termes importants sont "*sign, interpretant, object, ...*". Cette classe aborde *un thème des signes et ses représentations*; dans la classe 2, les termes importants sont "*belief, doubt, opinion, truth, reason,...*". Cette classe discute de *la conviction, du doute à l'encontre des opinions, des vérités et des raisons* ; ou bien dans la classe 17, on aura les termes importants sont "*idea, association, habit,...*". Cette classe aborde *l'idée, de l'association et des habitudes*, etc.

En ce qui concerne le premier filtrage, un ensemble des patrons lexicaux est obtenu à partir du corpus de test et des termes cibles trouvés. Des fonctions sont implémentés en Python afin d'extraire une liste des index des phrases dans le corpus en nous servant du résultat des patrons lexicaux. Par la suite une autre fonction permet également d'extraire toutes les phrases candidates avec leurs index correspondants. Les résultats sont représentés dans le tableau 2 :

Tableau 2 : Statistiques des extractions des phrases candidates pertinentes dans tous les segments de 18 classes du corpus de test Collected Paper de Peirce

Classes	Phrases (Segments)	#Termes cibles	#Phrases pertinentes	Pourcentage de réduction de taille
1	627 (40)	7	24	3.83 %
2	650 (50)	10	70	10.77 %
3	2513 (171)	10	40	1.59 %
4	152 (14)	8	10	6.58 %
5	268 (15)	6	19	7.09 %
6	2419 (188)	7	60	2.48 %
7	363 (27)	7	5	1.38 %
8	3103 (198)	7	51	1.64 %
9	32 (2)	5	0	0 %
10	197 (16)	5	14	7.11 %
11	185 (5)	7	42	22.70 %
12	158 (15)	6	8	5.06 %
13	631 (39)	7	50	7.93 %
14	225 (21)	6	9	4 %
15	26 (14)	9	4	15.38 %
16	466 (37)	9	26	5.58 %
17	769 (66)	8	35	4.55 %
18	282 (22)	10	24	8.51 %

En ce qui concerne le deuxième filtrage, un sous-ensemble des phrases candidates pertinentes est obtenu (Tableau 3). Une réduction est possible jusqu'à un certain nombre très petit de phrases selon la nécessité de lecture des chercheurs, par exemple à 10, à 5 et même à 3 phrases.

Tableau 3 : Résultat du deuxième filtrage avec un exemple de la *classe 1* du corpus de texte

Classe	1
Termes cibles	**sign**,interpretant,object,determination,representation, mean, idea
Thème	Signes et ses représentations
Premier filtrage	24 phrases candidates pertinentes correspondant aux 24 patrons lexicaux

Deuxième filtrage **Choix d'extraction: 5 phrases candidates pertinentes**	From that hypothesis, the rules of the **sign** - may be mathematically deduced. Does not electricity mean more now than it did in the days of Franklin? And this **creature of the sign** is called the **Interpretant**. But in analyzing the **general nature of a sign**, it will be needful, to distinguish radically **different kinds of signs**. Hence, we are led to generalize our **idea of argumentation**, from the perception that one **assertion** has to be admitted because another is admitted, to embrace also that process of **thought** in which we think that though one assertion is true yet another is not thereby necessarily true.

Nous constatons une réduction significative du nombre de segments de chaque classe appartenant au corpus de test. Par exemple dans la classe 1, le premier filtrage donne un résultat contenant 24 phrases candidates pertinentes correspondant aux 24 patrons lexicaux. En sachant qu'il y a 627 phrases dans cette même classe (Tableau 2), la taille est réduite à 3.83% de l'ensemble de 40 segments de classe 1. Le tableau 3 montre qu'au deuxième filtrage, avec un choix d'extraction de 5 phrases candidates pertinentes, nous constatons bien que le texte aborde l'aspect des signes et ses différentes représentations possibles en analysant ses caractérisques générales. En plus les règles du signe peuvent être mathématiquement déduites. Autrement dit, cette réduction de taille devient intéressante à la réduction de dispersion des données dans un espace vectoriel.

4. Conclusion et Perspective

Dans cette recherche, une analyse syntactico-sémantique a été présentée. Elle permet de mieux exploiter des résultats d'analyseurs classificatoires ayant été appliqués à des contextes d'une expression conceptuelle cible. Cette méthode d'exploitation de ces résultats consiste à appliquer aux segments de ces classes des analyses syntactico-sémantiques. Il permet alors aux chercheurs(-euses) de découvrir dans ces phrases d'importantes organisations sémantiques synthétiques révélatrices de propriétés conceptuelles dominantes telles que définition, inférences, propriétés, etc.

En perspective, nous souhaitons appliquer notre approche à d'autres techniques d'analyses thématiques. Nous voudrions analyser de manière plus profonde des constituants dans une structure de segments de phrases pour but de faire ressortir toutes les dépendances syntaxiques susceptibles. La chaîne de traitement pourra être naturellement étendue à d'autres types de corpus textuels.

Références

Hunston, S. et G. Francis. (2000). Pattern Grammar. A corpus-driven approach to the lexical grammar of English. *Amsterdam / Philadelphie, John Benjamins*.

Vincent Claveau. (2003). Extraction de couples nom-verbe sémantiquement liés: une technique symbolique automatique. *TALN 2003, Batz-sur-mer, 11-14 Juin 2003*.

Cécile Fabre et Didier Bourigault. (2006). Extraction de relations sémantiques entre noms et verbes au-delà des liens morphologiques. *TALN 2006, Leuven, 10-13 avril 2006*: 121-129.

Juan-Manuel Torres-Moreno. (2014). Automatic Text Summarization. © ISTE Ltd 2014, ISBN 978-1-84821-668-6.