

Titre: Annotations sémantiques et analyse de surface pour l'extraction de graphes d'abstraction de débats politiques
Title:

Auteur: Edoukou Philippe Armel N'Techobo
Author:

Date: 2016

Type: Mémoire ou thèse / Dissertation or Thesis

Référence: N'Techobo, E. P. A. (2016). Annotations sémantiques et analyse de surface pour l'extraction de graphes d'abstraction de débats politiques [Mémoire de maîtrise, École Polytechnique de Montréal]. PolyPublie.
Citation: <https://publications.polymtl.ca/2180/>

 **Document en libre accès dans PolyPublie**
Open Access document in PolyPublie

URL de PolyPublie: <https://publications.polymtl.ca/2180/>
PolyPublie URL:

Directeurs de recherche: Michel Gagnon, & Amal Zouaq
Advisors:

Programme: Génie informatique
Program:

UNIVERSITÉ DE MONTRÉAL

ANNOTATIONS SÉMANTIQUES ET ANALYSE DE SURFACE POUR L'EXTRACTION
DE GRAPHES D'ABSTRACTION DE DÉBATS POLITIQUES

EDOUKOU PHILIPPE ARMEL N'TECHOBO

DÉPARTEMENT DE GÉNIE INFORMATIQUE ET GÉNIE LOGICIEL
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION
DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES
(GÉNIE INFORMATIQUE)

JUILLET 2016

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé :

ANNOTATIONS SÉMANTIQUES ET ANALYSE DE SURFACE POUR L'EXTRACTION
DE GRAPHES D'ABSTRACTION DE DÉBATS POLITIQUES

présenté par : N'TECHOBO Edoukou Philippe Armel

en vue de l'obtention du diplôme de : Maîtrise ès sciences appliquées

a été dûment accepté par le jury d'examen constitué de :

M. HURTUT Thomas, Ph. D, président

M. GAGNON Michel, Ph. D, membre et directeur de recherche

Mme ZOUAQ Amal, Ph. D, membre et codirectrice de recherche

M. DESMARAIS Michel C., Ph. D, membre

DÉDICACE

*Je dédie ce mémoire
à mon père N'techobo Assiri Alexandre
et à mes mamans Dedou N'ti Juliette et
Koutouan Loba Janette*

REMERCIEMENTS

Je remercie tout d'abord mes directeurs de recherche Michel Gagnon et Amal Zouaq. Cela a été un immense plaisir de travailler avec eux. Ils ont été très patients, rigoureux, toujours disponibles et très ouverts. J'ai vraiment beaucoup appris avec eux et de simples mots ne suffisent pas pour leur exprimer ma gratitude. Je remercie aussi mes collègues Félix, Ludovic, Lara, Mohamed, Konstantinos et Ludovic, du laboratoire de WeST, pour l'évaluation de mes travaux. Je remercie aussi tous ceux qui ont contribué de près ou de loin à la rédaction de ce mémoire.

RÉSUMÉ

De nombreux gouvernements ont pris des initiatives pour ouvrir et partager leurs données telles que les débats parlementaires. Ce type de corpus est généralement très grand, couvre plusieurs thématiques et n'est pas toujours organisé de manière linéaire. Dans ce travail de recherche, nous proposons une méthode pour extraire automatiquement des représentations abstraites basées sur des graphes qui représentent les sujets discutés durant les débats politiques et les relations qui existent entre ces sujets.

À cette fin, nous exploitons des annotateurs sémantiques basés sur le Linked Data pour l'extraction des entités concernées. De cette façon, nous pouvons représenter les sujets discutés avec des concepts dont la sémantique a déjà été définie dans le Linked Data de manière structurée, à la différence des méthodes existantes qui utilisent de simples mots clés. Aussi, nous extrayons des relations entre les concepts provenant du Linked Data et des relations de haut niveau entre ces entités provenant du corpus de débats. Ces dernières sont extraites avec des patrons d'analyse morphosyntaxiques, définis manuellement, et désambiguïsées en utilisant VerbNet. Avec les concepts et relations extraits, nous construisons un graphe abstrait qui représente le débat. Ce graphe est ensuite réduit en fonction de plusieurs paramètres pour ne conserver que les entités et relations les plus importantes. Le graphe généré, en plus de permettre la recherche sémantique, pourrait être réutilisé par d'autres systèmes pour la génération de résumé abstraitif ou effectuer des systèmes de réponses à des questions (*question answering systems*).

ABSTRACT

Many governments have taken initiatives to open up and share their data such as parliamentary debates. This type of corpus is generally very large, covers several topics and is not always organized in a linear manner. In this research, we propose a method to automatically extract abstract representations based on graphs that represent the topics discussed during political debates and the relationships between these topics.

To this end, we use semantic annotators based on Linked Data for extracting topics. In this way, we can represent the discussed topics with concepts whose semantics has already been defined on the Linked Data cloud in a structured way, unlike existing methods which generally rely on simple keywords. Also, we extract relations between the concepts based on the information available on the Linked Data Cloud and provide high level relations between these entities from the corpus of debates. These relations are extracted with morpho-syntactic patterns defined manually and disambiguated using VerbNet. With the concepts and relationships extracted, we construct an abstract graph representing the debates. This graph is successively reduced based on several parameters to keep only the most important entities and relationships. The generated graph, in addition to enabling semantic search, could be reused by other systems for the generation of abstractive summarization or question answering.

TABLE DES MATIÈRES

DÉDICACE.....	III
REMERCIEMENTS	IV
RÉSUMÉ.....	V
ABSTRACT	VI
TABLE DES MATIÈRES	VII
LISTE DES TABLEAUX.....	IX
LISTE DES FIGURES.....	XIII
LISTE DES SIGLES ET ABRÉVIATIONS	XIV
LISTE DES ANNEXES.....	XV
CHAPITRE 1 INTRODUCTION.....	1
CHAPITRE 2 REVUE DE LITTÉRATURE	4
2.1 Annotations sémantiques.....	4
2.2 Représentation de documents textuels sous forme de graphe	5
2.3 Extraction de relations.....	9
2.4 Analyse de débats politiques	15
2.5 Notre approche	18
CHAPITRE 3 MÉTHODOLOGIE	20
3.2 Architecture générale du système	22
3.3 Identification des entités.....	24
3.3.1 Annotations sémantiques.....	24
3.3.2 Extension des annotations	25
3.3.3 Agrégation d’annotations	30
3.3.4 Calcul des scores d’annotations	31

3.3.5	Extraction de cooccurrences.....	34
3.4	Extraction de relations.....	35
3.4.1	Extraction de relations par analyse de surface	35
3.4.2	Désambiguïsation de relations basée sur VerbNet.....	40
3.4.3	Extraction de relations à partir de DBpedia	44
3.5	Génération de graphes résumés.....	46
CHAPITRE 4	EXPÉRIMENTATION ET ANALYSE.....	50
4.1	Construction du corpus de référence.....	50
4.2	Évaluation des scores associés aux annotations.....	53
4.2.1	Métriques utilisées.....	56
4.2.2	Résultats et analyse	56
4.3	Évaluation des annotations étendues.....	66
4.4	Évaluation des relations verbales	69
4.4.1	Métriques utilisées.....	69
4.4.2	Résultats et analyse	70
4.5	Évaluation du graphe résumé	80
CHAPITRE 5	CONCLUSION	86
BIBLIOGRAPHIE		88
ANNEXES		97

LISTE DES TABLEAUX

Tableau 3.1. Caractéristiques de 1217 débats du corpus Hansard pour les périodes d’avril 2006 à mars 2016.	21
Tableau 3.2. Liste des patrons d’extension d’annotation exprimés sous forme d’expressions régulières. Les catégories grammaticales des exemples sont indiquées en indice.	28
Tableau 3.3. Patrons d’extraction de relations verbales.	38
Tableau 4.1. Caractéristiques du corpus d’évaluation : nombre d’interventions, nombre d’intervenants, nombre d’annotations provenant d’AlchemyAPI, nombre d’annotations étendues et nombre de relations verbales par sujet de travail. Tous les sujets proviennent de débats discutés en 2015.	51
Tableau 4.2. Tableau indiquant la précision des annotations sélectionnées dans les références et l’accord entre les évaluateurs.	54
Tableau 4.3. Tableau indiquant les précisions pour le top-50 calculées avec les références A1r et A2r	57
Tableau 4.4. Tableau indiquant les précisions pour le top-10 calculées avec les références A1r et A2r	58
Tableau 4.5. Tableau indiquant les précisions pour le top-50 calculées avec les références A1r et A2r	59
Tableau 4.6. Tableau indiquant les précisions pour le top-10 calculées avec les références A1r et A2r	60
Tableau 4.7. Tableau indiquant les précisions pour le top-50 calculées avec les références.	62
Tableau 4.8. Tableau indiquant les précisions pour le top-10 calculées avec les références A1r et A2r	63
Tableau 4.9. Récapitulatif des résultats obtenus	64
Tableau 4.10. Tableau indiquant pour chaque sujet de travail la précision d’extension ₁ , la précision d’extension ₂ et l’accord entre les annotateurs.	66

Tableau 4.11. Quelques erreurs dans le processus d'extension d'annotations.....	68
Tableau 4.12. Tableau indiquant les précisions d'extractions de relations verbales.....	70
Tableau 4.13. Tableau présentant quelques erreurs rencontrées dans le processus d'extraction de relations	72
Tableau 4.14. Tableau indiquant la précision de pertinence des relations verbales.....	73
Tableau 4.15. Tableau indiquant la précision VerbNet calculée pour chacun des évaluateurs	75
Tableau 4.16. Tableau indiquant la précision de la désambiguïsation calculée pour chacun des évaluateurs.....	78
Tableau 4.17. Caractéristiques des graphes labélisés et résumés des sujets qui composent le corpus d'évaluation. Pour chaque type de graphe et pour chaque sujet de travail, on présente le nombre de sommets, le nombre d'arêtes, le degré minimum et maximum (entre rochets) et le degré moyen.....	81
Tableau 4.18. Résultats de l'évaluation des graphes résumés. Les indices 1 et 2 représentent les annotateurs.....	82
Tableau C.1. Catégories grammaticales utilisées	102
Tableau E.1. Interprétation des valeurs du Kappa.....	107

Tableau F.1. Évaluations qualitatives du graphe résumé relatif au sujet Public Safety du 4 mai.....	108
Tableau F.2. Évaluations qualitatives du graphe résumé relatif au sujet Taxation du 4 mai....	109
Tableau F.3. Évaluations qualitatives du graphe résumé relatif au sujet Ethics du 5 juin.....	110
Tableau F.4. Évaluations qualitatives du graphe résumé relatif au sujet Anti-terrorism Act du 5 mai	111
Tableau F.5. Évaluations qualitatives du graphe résumé relatif au sujet Taxation du 5 mai....	112
Tableau F.6. Évaluations qualitatives du graphe résumé relatif au sujet Taxation du 6 mai....	113
Tableau F.7. Évaluations qualitatives du graphe résumé relatif au sujet Finance du 7 décembre	114
Tableau F.8. Évaluations qualitatives du graphe résumé relatif au sujet National Defense du 7 décembre.....	115
Tableau F.9. Évaluations qualitatives du graphe résumé relatif au sujet Taxation du 8 mai	116
Tableau F.10. Évaluations qualitatives du graphe résumé relatif au sujet Infrastructure du 8 juin	117
Tableau F.11. Évaluations qualitatives du graphe résumé relatif au sujet Citizen Voting Act du 1 mai.....	118
Tableau F.12. Évaluations qualitatives du graphe résumé relatif au sujet Respecting Families of Murdered and Brutalized Persons Act du 2 juin.....	119
Tableau F.13. Évaluations qualitatives du graphe résumé relatif au sujet Aboriginal Affairs du 3 juin.....	120

Tableau F.14. Évaluations qualitatives du graphe résumé relatif au sujet The economy du 3 juin.....	121
Tableau F.15. Évaluations qualitatives du graphe résumé relatif au sujet National Defence du 4 mai.....	122

LISTE DES FIGURES

Figure 3.1. Architecture générale du système	23
Figure 3.2. Exemple de classe VerbNet	42
Figure 3.3. Graphe non labélisé représentant le sujet de travail <i>Citizen Voting Act</i> pour le débat du 1er mai 2015.....	47
Figure 3.4. Graphe labélisé relatif au sujet de travail <i>Citizen Voting Act</i> pour le débat du 4 mai 2015, généré avec les 15 termes les plus fréquents.....	48
Figure 3.5. Portion d'un graphe labellisé en format RDF	49
Figure A.1. modèle de donnée	98

LISTE DES SIGLES ET ABRÉVIATIONS

LOD	Linked Open Data
NMF	Non negative Matrix Factorization
ORE	Open Relation Extraction
POS	Part of Speech
RDF	Resource Description Framework
TDA	Topological Data Analysis
TF	Term Frequency
TF-IDF	Term Frequency – Inverse Document Frequency
T-VSM	Term – Vector Space Model
URI	Uniform Ressource Identifier
XML	eXtensible Markup Language

LISTE DES ANNEXES

Annexe A – Modèle de donnée	96
Annexe B – Description de l'application	98
Annexe C – Catégories grammaticales	101
Annexe D – Exportation du graphe au format RDF.....	105
Annexe E – Interprétation du kappa de Cohen	107
Annexe F – Evaluations qualitatives des graphes résumés	108

CHAPITRE 1 INTRODUCTION

L'arrivée d'Internet et des technologies du Web ont causé une multiplication de la quantité d'informations disponibles. En particulier, de nombreux organismes gouvernementaux se sont mis à produire et mettre en ligne des données dites ouvertes pour aider le public à mieux comprendre l'action gouvernementale. Dans ce travail de recherche, nous nous intéressons particulièrement à l'analyse des débats politiques. Les corpus de débats politiques sont généralement très larges et traitent d'une variété de sujets et ne sont pas toujours organisés de manière linéaire. Le traitement manuel de ce type de corpus devient une tâche de plus en plus difficile, un défi rencontré par de nombreuses applications ayant trait aux méga-données (*big data*). L'automatisation de ce processus pourrait apporter un gain de temps appréciable en ce qui concerne l'accessibilité et le traitement des informations. Pour cela, des techniques de *recherche d'information* peuvent être utiles pour accéder et récupérer des documents individuels à l'intérieur d'un corpus en fonction de requêtes ou de paramètres spécifiés par l'utilisateur. Dans ce cadre, l'extraction des notions, sujets, et mots clés les plus importants d'un corpus politique permet de créer une représentation de haut niveau du contenu des débats politiques. Certaines techniques existantes [2, 3, 9] de *forage de texte* permettent d'extraire des sujets à partir de documents textuels. Ces techniques, en général, représentent les sujets comme un ensemble de mots clés. Dans ce travail, nous avons décidé d'exploiter plutôt le concept d'*annotations sémantiques*, qui s'appuient sur le réseau *Linked Open Data*. Une annotation sémantique est un lien entre une mention contenue dans un document textuel et un concept défini dans une base de connaissances. L'annotation est une évidence que le concept défini dans la base de connaissances apparaît dans le document textuel. En utilisant ces annotations, nous pouvons bénéficier des enrichissements disponibles dans la base de connaissances. Par ailleurs, le *Linked Open Data* (LOD) est un paradigme de publication des données sur le Web où les informations ne sont pas décrites par des documents isolés, mais forment plutôt un grand réseau global interconnecté. Ce projet, débuté en 2007, a pour but d'identifier les ensembles de données disponibles en licence libre, les convertir selon les standards du *Linked Data* et les publier sur le Web.

Notre travail de recherche se situe dans le contexte d'une tâche d'abstraction des débats politiques canadiens en se basant sur les graphes de connaissances (*knowledge graphs*) disponibles sur le LOD. Dans l'état de l'art, des travaux de recherches se sont aussi penchés sur la tâche d'abstraction

de corpus [31, 17, 18]. Il s'agit, pour la plupart [31], de techniques de résumés extractifs qui extraient du corpus les phrases les plus importantes en fonction d'heuristiques. D'autres travaux [17, 18], exploitent des bases de connaissances spécifiques à certains domaines pour générer une représentation abstraite du corpus basée sur un graphe. Ce mémoire a pour particularité de s'appuyer sur la base de connaissances DBpedia [80], qui représente le contenu de Wikipédia sous forme structurée. La base de connaissances DBpedia est considérée par beaucoup comme le centre (*hub*) du LOD, parce qu'elle décrit plusieurs concepts provenant de différents domaines, ainsi que des relations entre ces concepts. L'exploitation de DBpedia à des fins de compréhension automatique du contenu textuel a connu un essor considérable dans des travaux récents [80], mais n'a pas été, à notre connaissance, appliquée à l'analyse de corpus politiques.

Dans ce travail, l'objectif visé est que la représentation extraite puisse être réutilisée par des systèmes intelligents pour générer des *résumés abstractifs* ou effectuer des systèmes de réponse à des questions (*question answering systems*). Pour cela, cette représentation doit mettre en évidence non seulement les sujets discutés dans les différents débats, mais aussi les relations entre ceux-ci. Nous avons donc décidé d'abstraire le corpus **sous forme d'un graphe** où les sommets représentent les concepts discutés et les arêtes représentent des relations sémantiques entre ces concepts. Dans l'optique d'exploiter au maximum les informations provenant de la base de connaissances, nous extrayons aussi les relations existantes entre les concepts de la base de connaissances qui n'apparaissent pas forcément dans le corpus de débat. Nous extrayons aussi des relations entre les annotations à l'intérieur du corpus de débat en utilisant seulement *l'analyse syntaxique de surface*. Les relations provenant de la base de connaissances et celles extraites à partir du corpus constituent les arêtes de notre graphe. Après avoir généré le graphe représentant le corpus, nous le réduisons pour obtenir un *graphe résumé*.

Plusieurs aspects rendent cette tâche difficile. Le premier est la difficulté d'extraire des annotations pertinentes. En effet, les corpus de débats politiques étant de grande dimension, il faudra trouver des métriques adéquates pour assigner des scores aux annotations et ainsi conserver seulement les annotations les plus importantes. Aussi, l'ensemble des relations qu'il est possible d'extraire du corpus de débats n'est pas connu à priori. Pour que la représentation abstraite que nous générons soit réutilisable par d'autres systèmes, la sémantique de ces relations doit être révélée à travers une phase de désambiguïsation. Être en mesure d'extraire ces annotations et ces relations, les manipuler pour pouvoir obtenir une vision abstraite du corpus est un défi et, à ce jour, il n'existe pas de

solutions satisfaisantes à ce problème. La question de recherche à laquelle nous répondrons dans ce mémoire est la suivante :

Dans quelle mesure, est-il possible d'obtenir un graphe résumant un corpus politique en combinant des annotations sémantiques et des relations extraites suite à une analyse de surface ?

Dans le chapitre 2, nous présentons une revue de littérature sur les annotations sémantiques, les représentations de documents sous forme de graphe et l'analyse de débats politiques. Nous situons aussi dans ce chapitre notre approche par rapport à l'état de l'art. Dans le chapitre 3, nous présentons la méthodologie adoptée. Dans le chapitre 4, nous présentons les résultats de notre expérimentation avant de conclure dans le chapitre 5.

CHAPITRE 2 REVUE DE LITTÉRATURE

Dans cette revue de littérature, nous introduisons d’abord l’annotation sémantique. Ensuite nous présentons quelques travaux effectués dans le cadre de la représentation de documents textuels sous forme de graphe. Après, nous présentons des techniques utilisées pour l’extraction de relations. Finalement, nous présentons des travaux effectués sur des corpus politiques. Nous terminons ce chapitre en situant notre approche par rapport à l’état de l’art.

2.1 Annotations sémantiques

L’annotation sémantique basée sur le *Linked Open Data (LOD)* est une méthode qui a émergé au cours de ces dernières années. Avant de présenter l’annotation sémantique, nous introduisons la notion d’annotation et le Linked Open Data.

Une annotation consiste à associer des métadonnées à un document ou à certaines parties de ce document. Le processus d’annotation peut être effectué manuellement [74] ou de manière automatique [75, 76, 77]. Supposons qu’on ait un document textuel dont le contenu est le suivant :

The winter of 2013-14 saw a transportation crisis that impeded the growth and credibility of our export economy.

Pour ce texte, on pourrait avoir comme annotations :

transportation

crisis

growth

economy

investment [<http://dbpedia.org/resource/Investment>]

On peut noter, dans l’exemple ci-dessus, que les annotations *transportation crisis*, *growth* et *economy* apparaissent dans le texte original. On dira qu’elles sont **ancrées**. L’annotation *investment* n’apparaît pas dans le texte, on dira que c’est une annotation **non ancrée**. Dans cet exemple, les annotations nous permettent de comprendre de quoi parle le texte. Très souvent, les annotations

contiennent de l'information supplémentaire. Par exemple une annotation pourrait indiquer que la portion de texte à laquelle elle réfère représente le nom d'un emplacement ou d'une personne.

Des annotations sont considérées comme *annotations sémantiques* quand elles ne sont pas simplement des mots clés, mais ont un lien vers un thésaurus ou une base de connaissances où leur sémantique est décrite. Par exemple, une annotation sémantique pourrait simplement appartenir à un dictionnaire où son sens est décrit en langage naturel. Aussi, une annotation sémantique pourrait appartenir à une base de connaissances structurée où elle est décrite dans un langage formel comme RDF [81], et aurait des relations sémantiques avec d'autres éléments de la base de connaissances. Dans l'exemple précédent, l'annotation investment possède l'URI suivante :

<http://dbpedia.org/resource/Investment>

Cette URI pointe vers une description du concept Investment, défini dans la base de connaissances DBpedia [80] qui appartient au *Linked Open Data*.

Les outils utilisés pour identifier à l'intérieur de textes les concepts existants dans une base de connaissances du LOD sont appelés *annotateurs sémantiques basés sur le LOD*. Ces outils représentent donc une ressource importante pour relier les corpus politiques aux concepts encyclopédiques. L'avantage d'utiliser cette méthode est de bénéficier de concepts ayant une sémantique et une taxonomie déjà définies dans le LOD plutôt qu'utiliser de simples mots clés.

Ceci conclut notre aperçu sur les annotations sémantiques. Dans la section qui suit, nous présentons des techniques de représentation de texte sous forme de graphe.

2.2 Représentation de documents textuels sous forme de graphe

Dans le domaine de l'analyse de texte et du traitement automatique des langues, le modèle vectoriel, appelé T-VSM [23] (Term – Vector Space Model), est un des plus populaires pour la représentation de contenu textuel. Ce modèle représente chaque document comme un vecteur dans un espace à t dimensions, où t est le nombre de termes distincts ou la taille du vocabulaire utilisé. Dans sa forme la plus simple, ce vecteur indique, pour chaque terme, le nombre de fois qu'il apparaît dans le document. L'hypothèse derrière ce modèle est que la sémantique d'un document

peut être représentée par un ensemble de termes. Ceci est une simplification, car le modèle vectoriel a les limitations suivantes [32, 36, 78] :

- Le modèle considère que tous les mots sont indépendants et ignore la séquence dans laquelle les mots apparaissent ainsi que les relations entre ces mots.
- Si deux documents ont des contenus similaires, mais exprimés au travers des mots différents, il est difficile de capturer la similarité entre ces documents.

Des représentations théoriques de textes sous forme de graphe ont été proposées [32, 36] pour combler les limitations du modèle vectoriel et représenter au mieux les documents textuels en capturant les relations entre les termes. Les auteurs en [32] proposent trois types de graphes pour représenter un document et capturer les relations entre les termes qu'il contient. La première représentation se focalise sur l'ordre dans lequel les termes apparaissent dans le document. Dans cette représentation, un document est représenté par un graphe qui a pour sommets les termes du document et des arrêtes pondérées dont les poids correspondent au nombre de fois où deux termes apparaissent dans une fenêtre mobile de dimension prédéfinie. La deuxième représentation capture les relations de cooccurrence entre les termes. Dans cette représentation, on a un graphe où les arêtes indiquent le nombre de fois où deux termes apparaissent dans un même groupe. Un groupe peut être une phrase, un paragraphe ou toute autre unité lexicale. La différence principale entre ces deux approches est que la première approche impose une distance maximale entre les termes et les arêtes sont orientées. Pendant que pour la deuxième approche, il suffit que les termes apparaissent dans une même unité lexicale. Pour cette approche, l'ordre des termes n'a pas d'importance ; les arêtes ne sont donc pas orientées. La troisième représentation a pour but de capturer le contexte entre les termes. Le graphe associé à cette représentation a des arêtes orientées qui indiquent la probabilité de transition d'un terme à un autre. Cette probabilité de transition est calculée comme la similarité des paragraphes autour des termes, ou avec la formule suivante basée sur la fréquence

de cooccurrence :
$$P(t_j|t_i) = \frac{n(t_i, t_j)}{\sum_{k=1}^{|T|} n(t_i, t_k)}$$
 où $P(t_j|t_i)$ est la probabilité de transition du

terme t_i au terme t_j , $n(t_i, t_j)$ la fréquence de cooccurrence entre les termes t_i et t_j et $|T|$ est le nombre de termes du graphe.

Dans notre travail de recherche, nous avons appliqué le deuxième modèle basé sur les relations de cooccurrence pour représenter notre corpus. En partant de l'intuition que les concepts cooccurents

dans plusieurs documents sont liés, nous avons représenté notre corpus sous forme de graphe et avons utilisé les relations de cooccurrence comme indicateur de la présence de relations entre deux concepts.

Les auteurs en [36] proposent un modèle très similaire à celui en [32] pour représenter les documents textuels sous forme de graphe. L'idée de cooccurrence revient à ceci : il existe une arête entre deux sommets du graphe si les mots qu'ils représentent cooccurrent dans au moins un paragraphe. Par contre, dans ce modèle, il ne s'agit pas d'un score de cooccurrence, mais plutôt d'une distance entre les mots. Soit deux mots A et B apparaissant dans le document textuel. $num(A)$ et $num(B)$ représentent respectivement l'ordre dans lequel A et B apparaissent dans le texte. $(num(B) - num(A))$ indique donc la distance entre les deux mots. Dans le graphe, si A et B cooccurrent dans au moins un paragraphe, il existe un arc pondéré entre ces mots dont le poids W_{AB} est calculé comme suit :

$$W_{AB} = 1 / num(A) - num(B)$$

Si A et B apparaissent plusieurs fois dans le document, W_{AB} sera calculé en fonction de la plus petite distance entre A et B. La formule sera la suivante : $W_{AB} = 1 / \min\{num(A) - num(B)\}$

Ces modèles théoriques sont utilisés dans plusieurs travaux de recherche pour effectuer des tâches d'extraction de mots clés [34, 38, 40, 35], de résumés automatiques [14, 15, 16, 17, 18], ou de représentation conceptuelle de texte [33, 39]. Alors que certains travaux utilisent les graphes comme structure intermédiaire pour effectuer leurs tâches [34, 38, 40, 35, 14, 15, 16], d'autres travaux présentent leurs résultats sous forme de graphe [33, 39, 17, 18].

Un des avantages d'utiliser les représentations basées sur graphes est de pouvoir utiliser le concept de centralité [14] pour pouvoir extraire les éléments les plus importants. Ce concept, largement utilisé dans l'analyse des réseaux sociaux, est fondé sur la simple idée que, dans un graphe, les éléments les plus centraux et connectés aux autres sont les plus importants. Cette idée a inspiré par exemple les créateurs du moteur de recherche Google avec leur algorithme PageRank, basé sur un graphe qui classifie les sites web et par conséquent trouve les sites web les plus importants, ou mieux, les plus populaires. D'autres représentations inspirées de PageRank ont été introduites, comme TextRank [34] pour l'extraction des mots clés et LexRank [14] pour la génération automatique de résumé.

TextRank [34] et d'autres techniques d'extraction de mots clés basées sur un graphe [38, 40, 35] construisent un graphe où les sommets représentent les mots du document et les arêtes, en général pondérées, représentent les distances relatives entre les mots. Les mots représentés sur le graphe sont en général filtrés en utilisant les catégories grammaticales ou une liste des mots à ignorer. L'approche en [40] se distingue par l'utilisation de concepts provenant d'une base de connaissances, plutôt que de simples mots clés. TextRank utilise des arêtes qui indiquent le nombre de fois où deux mots cooccurrent dans une fenêtre mobile de dimension allant de deux à dix. Après avoir construit le graphe, l'algorithme assigne à chacun des mots du graphe, un score en fonction de sa centralité.

Une approche similaire est utilisée par les auteurs en [35] où des arêtes orientées et pondérées sont insérées entre les mots si seulement ceux-ci se suivent dans le texte. Les auteurs en [38] comparent les approches précédentes utilisées en [40] et [34] de mots successifs et fenêtre mobile pour extraire les mots clés. Ils comparent aussi plusieurs méthodes d'assignation de poids aux arêtes. Avec un algorithme de centralité qui considère que les mots clés sont ceux globalement plus proches de tous les autres mots, ils obtiennent les meilleurs résultats. Dans une autre approche basée sur TextRank [40], les auteurs extraient les mots clés d'un document textuel en construisant un graphe conceptuel et en y appliquant l'algorithme TextRank. La différence est que les sommets du graphe sont des concepts provenant de la base de connaissance WordNet. Les relations entre les nœuds dans le graphe sont les relations préexistantes dans WordNet (relations d'hyponymie et d'hypernymie). Après avoir associé des scores aux arêtes en fonction du type de relation qu'elles décrivent, une variante de l'algorithme TextRank est proposée pour extraire les concepts les plus importants.

Les structures de graphes sont aussi exploitées pour générer des résumés [14, 15, 16, 17, 18], une application connexe au travail présenté dans ce mémoire. LexRank [14] utilise un algorithme similaire à PageRank et TextRank pour résumer des documents textuels. Il utilise un graphe où chaque sommet représente une phrase du texte à résumer et les arêtes pondérées représentent des mesures de similarité entre les phrases. Ensuite en appliquant un algorithme, des scores sont associés aux phrases en fonction de leur centralité. Les phrases avec les plus grands scores constituent le résumé. LexRank est une méthode de résumé extractive, c'est-à-dire que le résumé généré ne peut contenir que des phrases provenant du texte original. Cette limitation ne s'applique pas aux techniques de résumé abstractives. Certaines d'entre elles [15, 16] s'appuient sur des structures et algorithmes de graphes pour générer le résumé. Ces approches, en général,

représentent tout le texte original sous forme de graphe. Ensuite, alors que certaines d'entre elles [15] utilisent des algorithmes de traversée de graphe pour générer leur résumé, d'autres [16] utilisent des algorithmes d'extraction de sous-graphes optimaux. Les travaux de génération de résumé présentés jusque-là utilisent les graphes comme structure intermédiaire pour produire le résumé. En [17, 18], le résumé final est présenté sous forme de graphe. Le but des auteurs dans ces travaux est de résumer des articles cliniques. En [17] les auteurs extraient des articles cliniques les concepts et les relations provenant de la base de connaissances UMLS. Ces concepts et relations sont filtrés avec plusieurs heuristiques pour ne conserver que les plus importants. Le résultat final est alors présenté sous forme de graphe. Ce travail est étendu en [18] en utilisant des mesures de centralité sur le graphe produit par [17] pour extraire un graphe résumé.

Dans notre travail, nous ne produisons pas de résumé textuel, mais générons une représentation abstraite du corpus qui pourrait servir non seulement à la production de résumé, mais aussi à d'autres tâches d'analyse de texte, comme la recherche d'information, les systèmes de réponse automatique aux questions (*question answering* en anglais), le calcul de similarité et d'autres encore. Plusieurs travaux, par exemple [17,18], utilisent des représentations basées sur les graphes pour abstraire un corpus. À la différence de ces travaux, nous avons essayé d'utiliser une approche indépendante du domaine ou du corpus d'application.

Dans cette section nous avons fait le tour des représentations basées sur un graphe pour l'abstraction des documents textuels. Une des motivations dans la construction d'un graphe est la mise en évidence des relations entre les concepts. Dans notre travail de recherche, nous avons mis en place un tel processus d'extraction de relations entre les concepts. Dans la section qui suit, nous présentons les techniques généralement utilisées pour cette tâche.

2.3 Extraction de relations

Pour construire notre graphe conceptuel, une phase d'extraction de relations est nécessaire. La tâche d'extraction de relations consiste à découvrir des liaisons sémantiques entre des entités dans un texte. Dans cette section, nous présentons plusieurs approches utilisées pour cette tâche. Nous classifions les différentes approches selon les trois catégories suivantes : supervisées [54, 55, 57, 56, 53, 44, 45, 46, 47], semi-supervisées [49, 50, 51, 52] et non supervisées [58, 59, 60, 41, 61,

62]. Nous présentons dans les paragraphes qui suivent chacune des catégories et nous terminons avec un paragraphe sur l'extraction des relations non contraintes (open relation extraction).

L'extraction des relations peut être considérée comme une tâche de classification où, à partir d'exemples de relations étiquetées positivement ou négativement, il est possible de construire un classificateur en sélectionnant des caractéristiques (*features* en anglais) obtenues par analyse de texte, par exemple des catégories grammaticales, des relations syntaxiques de dépendances entre les mots, etc. Les méthodes supervisées utilisent des caractéristiques sémantiques et syntaxiques comme indices pour décider si deux entités sont en relation ou pas. Ces caractéristiques incluent les entités elles-mêmes, le type des entités, la séquence de mots entre les entités, la séquence des mots dans l'arbre syntaxique qui contient les entités, et le chemin entre les entités dans l'arbre de dépendances syntaxiques. [54] et [55] extraient des relations entre les entités en utilisant ces caractéristiques et un classificateur d'entropie maximale [57] ou un classificateur SVM (Support Machine Vector) [56].

Il est important de sélectionner seulement les caractéristiques importantes, c'est-à-dire celles qui sont de bons indicateurs de la présence de relations entre deux entités. Vu que l'extraction des relations utilise habituellement des représentations structurées des données en entrée (par exemple des arbres de dépendances syntaxiques, ou des graphes qui représentent les données textuelles), et que les caractéristiques sont sélectionnées de manière empirique, il est parfois difficile d'arriver à un sous-ensemble optimal de caractéristiques importantes. Pour remédier à ce problème, des **méthodes à noyau** (*kernel methods* en anglais) spécialisées ont été conçues pour exploiter la représentation des données en entrée et explorer de manière exhaustive et implicite ces représentations dans un espace dimensionnel de plus haut niveau [71].

Les méthodes à noyau sont basées sur la fonction noyau des chaînes de caractères [53] (*string kernel* en anglais). Cette fonction, que nous appellerons K_s , permet le calcul de similarité entre deux chaînes de caractères en fonction de toutes les sous-séquences communes de caractères. Ce concept est généralisé sur des objets génériques. On définit le noyau $K_s(x, y)$ entre deux objets comme les traits communs dans leur structure. Supposons que r^+ et r^- soient respectivement des exemples positifs et négatifs de relations extraites. Supposons que r' soit une relation de test, candidate à l'extraction. Si $K_s(r^+, r') > K_s(r^-, r')$ cela signifie que r' est une relation qui

devrait être extraite. Pour l'extraction de relations, les objets r^+ , r^- et r' peuvent représenter des séquences de mots autour des entités ou des arbres de dépendance contenant les entités.

Les deux méthodes à noyau majoritairement utilisées, que nous présentons dans la suite, sont les *Bag of features Kernel* [44] et les *Tree Kernel* [45, 46, 47].

Considérons la phrase :

“The headquarters of Google are situated in Mountain View.”

Étant donné que *Google* et *Mountain View* sont des entités, les mots *situated* et *headquarters* indiquent une relation de type *organisation - location* entre les deux entités. Dans ce cas, le contexte autour des entités est utilisé pour détecter la relation. Similairement, [44] utilise le contexte des mots autour des entités pour extraire des interactions de protéines. Ils décomposent une phrase s en $s_b e_1 s_m e_2 s_a$, où e_1 et e_2 sont deux entités nommées et s_b, s_m et s_a sont des portions du contexte avant, au milieu, et après les entités nommées, respectivement. Le noyau utilisé ici est une variation de la fonction K_s , qui considère des mots à la place des caractères.

Les auteurs en [45] extraient des relations de types *personne-affiliation* et *organisation-localisation* entre les entités. Ils extraient les arbres syntaxiques de surface des phrases qui contiennent les entités. Ils conçoivent ensuite une fonction noyau pour calculer la similarité entre deux arbres syntaxiques qu'ils implantent dans un classificateur SVM. Étant donné donc des exemples de phrases contenant des relations, qui ont été utilisées pour l'entraînement, et une phrase test y candidate à l'extraction, une relation sera extraite à partir de y si l'arbre syntaxique de y est similaire à ceux des exemples d'entraînement. La similarité est calculée avec la fonction noyau définie. Les auteurs en [46] utilisent une méthode similaire. La seule différence avec [45] est que les auteurs utilisent des arbres de dépendance syntaxique plutôt que des arbres syntaxiques de surface. Les auteurs incluent aussi plusieurs informations à l'intérieur des nœuds de l'arbre de dépendance, comme les catégories grammaticales, les types des entités, etc. Les auteurs en [46] argumentent que l'utilisation d'une structure représentative plus riche pourrait mener à de meilleures performances. Au contraire, les auteurs en [47] soutiennent que le plus court chemin entre deux entités dans un arbre de dépendance est suffisant pour l'extraction de relation. En utilisant seulement cette caractéristique, ils obtiennent des résultats supérieurs à [46]. Les auteurs en [48] représentent des articles scientifiques du domaine biomédical sous forme d'arbres de

dépendance syntaxique et en considérant l'ordre des mots. Au contraire des auteurs en [47], les auteurs en [48] considèrent tous les chemins du graphe dans la définition de leur fonction noyau.

Les méthodes supervisées définies plus tôt nécessitent un entraînement sur une quantité suffisante d'exemples de relations pour atteindre un certain niveau de précision. Lorsqu'on dispose de peu d'exemples pour l'entraînement, les méthodes semi-supervisées [49, 50, 51, 52] peuvent servir à inférer automatiquement des règles ou des patrons d'extraction de relations.

En [49] les auteurs extraient des relations de type *auteur-livre* à partir du Web. À partir d'un petit ensemble d'exemples, leur algorithme explore le Web à la découverte de patrons. Les patrons découverts sont ensuite utilisés pour l'extraction d'autres relations. Supposons, par exemple, que comme exemple de départ, on ait seulement la relation *auteur* (*Arthur Conan Doyle, The Adventures of Sherlock Holmes*), qui indique qu'Arthur Conan Doyle est l'auteur de « The Adventures of Sherlock Holmes ». L'algorithme extrait alors plusieurs phrases du Web qui contiennent les entités *Arthur Conan Doyle* et *The Adventures of Sherlock Holmes*. Chaque phrase extraite est sauvegardée dans un tuple qui considère les mots avant, entre et après les entités. Les tuples sont ensuite regroupés et généralisés en fonction des informations qu'ils ont en commun. La généralisation des tuples est ensuite utilisée pour la découverte d'autres relations. [50] utilise une technique très similaire. Les auteurs extraient des tuples et les regroupent avec une métrique de similarité, contrairement à [49] qui fait de l'appariement exact entre les mots. Les tuples regroupés sont successivement généralisés pour l'extraction de relations. Les auteurs en [51] extraient des relations d'instanciation en utilisant des patrons de type *NP1 such as NP2*. Ce patron indique que *NP2* est une instance de la classe *NP1*. Par exemple "Cities such as Paris" indique que l'entité *Paris* est de type *City*. Les auteurs infèrent aussi d'autres patrons en utilisant le contexte autour des instances trouvées. Par exemple, en analysant les mots qui apparaissent autour de Paris dans le Web, il est possible d'inférer le patron "[City] is the capital of [Country]", qui servirait à trouver de nouvelles entités de type ville.

Toutes ces techniques exigent que le type de relation à extraire soit défini à l'avance. Il n'est pas possible de "découvrir" de nouvelles relations ou d'extraire des relations non spécifiées. Cette limitation est comblée par les auteurs de TextRunner [52] et plus généralement par le paradigme d'extraction de relations non contraintes. Les auteurs définissent manuellement un ensemble de

règles pour extraire des relations entre les syntagmes nominaux à l'intérieur d'un corpus. Ces règles sont basées principalement sur les catégories grammaticales des mots.

Les approches non supervisées sont utilisées très fréquemment pour l'extraction d'hyponymie [58,59, 60, 41]. Elles se basent en général sur des règles ou des patrons définis manuellement. À partir de certains mots, les auteurs en [58] utilisent des heuristiques pour extraire des patrons nécessaires pour trouver d'autres mots de la même catégorie. Par exemple avec les mots *city*, *Colombia* et *Bolivia*, les auteurs extraient à partir d'heuristiques le patron pour l'extraction de la relation de localisation entre une ville et un pays. Les auteurs en [59, 60] utilisent une approche similaire pour extraire et catégoriser des entités nommées présentes dans un corpus. Ils utilisent des patrons lexicaux de type *X such as Y* et *X including Y* pour extraire les entités *X* et *Y*. Ils utilisent aussi la base de connaissances WordNet pour catégoriser les entités. Par exemple à partir de la forme du segment *programming languages such as Java*, ils déduisent que *Java* appartient à la catégorie *programming language*. *Programming language* est ensuite associé au concept *language* appartenant à la base de connaissance WordNet. Les auteurs en [41] tentent la même approche en utilisant uniquement le patron suivant : *[parent] such as [member] and [concept]*

Par exemple on aura *countries such as China and Japan*. Dans cet exemple le concept *China*, dont on connaît déjà la classe *Country*, est utilisé pour découvrir le concept *Japan* de la même classe. *Japan* pourra servir ensuite à la découverte d'autres concepts. Les auteurs définissent alors un graphe où les sommets sont les concepts extraits, et il existe une arête orientée pondérée entre deux sommets *A* et *B* si le concept *A* est utilisé pour découvrir *B*. Le poids de l'arc est le nombre de fois qu'*A* génère *B*. Les auteurs définissent les notions de popularité et productivité des nœuds basés respectivement sur leurs degrés entrants et sortants. Seuls les concepts ayant les plus grands scores de productivité et popularités sont extraits.

Certaines techniques [61], basées sur des relations verbales sont utilisées pour extraire des relations de causalité. Par exemple en [61] les auteurs, en utilisant un analyseur syntaxique, extraient d'abord toutes les relations de type *[NP] [V] [NP]*, où *NP* est un syntagme nominal et *V* est un verbe. Ensuite ils filtrent ces relations pour ne garder que celles qui indiquent des relations de causalité en utilisant des heuristiques. Par exemple, comme heuristique, ils s'assurent que le verbe extrait est un verbe de causalité. Un exemple de relations extraites avec leur système est *[heavy rain] [caused] [the flooding]*.

Dans le système TexRunner [52] présenté plus tôt, les types des relations extraites ne sont pas définis a priori. Il s'agit donc ici d'une extraction de relation non contrainte. Cette approche ne présume pas d'un ensemble prédéfini de relations et cible toutes les relations possibles. On peut identifier trois types d'extraction de relations non contrainte, selon la technique utilisée pour l'extraction. Le premier type utilise seulement l'analyse de surface [52, 63, 64], le deuxième utilise les arbres de dépendances syntaxiques [65, 66, 67] et le troisième, plus récent, utilise l'analyse sémantique [68, 69]. TexRunner [52] et son successeur ReVerb [63] sont basés sur l'idée que la plupart des relations peuvent être extraites avec quelques patrons basés sur l'analyse syntaxique. Par exemple en [63], seulement trois types de relations sont définis (*verbe*, *verbe + préposition*, *verbe + nom + préposition*). Le travail en [63] est étendu en [64] avec l'ajout de l'apposition et des formes possessives. [65] extrait des relations entre des entités en fonction du plus petit chemin qui les relie dans l'arbre de dépendances. [66] extrait des relations entre deux entités en appliquant des patrons dans le sous-graphe de dépendances qui contient les entités. [67] définit une fonction noyau pour détecter si les chemins dans des arbres de dépendances indiquent une relation. Une approche d'extraction de relation basée sur l'étiquetage de rôles sémantiques [70] a été démontrée comme étant efficace pour l'extraction de relation non contrainte. Ces approches [68, 69] ont la particularité d'utiliser, en plus des caractéristiques de dépendances syntaxiques, des caractéristiques sémantiques provenant des corpus annotés *PropBank* et *NomBank*.

La technique d'extraction de relations mise en place dans notre travail est non contrainte. En suivant le principe défini en [52, 64], dans lequel la plupart des relations peuvent être extraites avec quelques patrons grammaticaux et lexicaux, nous avons défini des patrons basés sur l'étiquetage morphosyntaxique. Contrairement à d'autres travaux en extraction de relations non contrainte [52, 63, 64, 65, 66, 67, 68, 69] nous nous sommes limité aux relations entre des concepts extraits lors de la phase d'annotation sémantique. Un des inconvénients de l'extraction de relations non contrainte est que les relations extraites, étant arbitraires, sont difficilement réutilisables par d'autres systèmes. On peut savoir qu'il y a une relation entre deux entités, mais la nature de cette relation est difficilement interprétable par un outil de traitement automatique. Nous avons résolu ce problème en liant les relations à la base de connaissances VerbNet et en les désambiguïsant. Nous obtenons donc des relations abstraites de plus haut niveau dont la sémantique est clairement définie en VerbNet.

Dans notre travail, nous avons choisi comme corpus d'application celui des débats politiques canadiens. La raison de ce choix est que les débats politiques sont vastes, structurés de manière non linéaire et traitent de plusieurs thématiques. Notre hypothèse est que ce type de corpus est idéal pour présenter une approche générique, en plus de présenter un intérêt en soi. Dans la section qui suit, nous présentons les travaux de recherche effectués sur les corpus de débats politiques.

2.4 Analyse de débats politiques

La plupart des travaux de recherche effectués sur des corpus politiques concernent l'extraction de sujets (*topics modeling*) [10, 11, 12] et la prédiction automatique dans les débats [6, 8, 7] des vainqueurs d'une élection ou d'un débat. L'extraction de sujets est une forme d'abstraction de corpus; elle s'inscrit donc dans le cadre de notre recherche. Ce n'est pas le cas pour la prédiction dans les débats. Dans les paragraphes qui suivent, nous survolons donc les techniques utilisées pour la prédiction dans les débats politiques. Nous décrivons ensuite quelques techniques utilisées pour l'extraction de sujets dans les débats politiques.

Certains travaux essaient d'anticiper les vainqueurs des élections en analysant les débats qui opposent les candidats [6]. D'autres travaux essaient de déterminer le résultat des votes parlementaires sur l'acceptation ou le refus de projets de loi en analysant les transcriptions des discours [8, 7]. Les techniques utilisées pour la prédiction sont en général la classification avec des algorithmes d'apprentissage machine et l'analyse de sentiment. On peut définir deux catégories de caractéristiques utilisées pour la classification. La première catégorie est celle des caractéristiques linguistiques qui sont basées sur le contenu textuel du débat, comme le vocabulaire utilisé, l'usage de contenu affectif, etc. La deuxième catégorie contient les caractéristiques structurelles. Celles-ci considèrent exclusivement la structure du débat et non son contenu textuel. Dans [6], les auteurs essaient de prédire le vainqueur des débats pour les élections vice-présidentielles et présidentielles aux États-Unis. Ils identifient donc plusieurs caractéristiques dont ils calculent les valeurs relatives pour trouver les plus importantes. Ils les utilisent ensuite pour effectuer la prédiction. Dans [8], l'auteur essaie de prédire les résultats de débats parlementaires. Il se focalise sur le parlement anglais et prédit l'acceptation ou le rejet des amendements. Il prédit aussi la manière dont le résultat sera déterminé, c'est-à-dire soit par un consensus ou par un vote. Les prédictions sont basées

exclusivement sur les caractéristiques structurelles du débat plutôt que son contenu textuel. En [7], les auteurs essaient de prédire les résultats des votes parlementaires avec l'analyse de sentiments, précisément en utilisant un lexique où des scores de sentiment sont associés aux termes. Les auteurs utilisent une structure de donnée basée sur un graphe. Dans cette représentation, les sommets représentent les intervenants. Ceux-ci sont étiquetés comme positifs ou négatifs en fonction de la prévision de leur vote. Les arêtes du graphe représentent la similarité des discours entre deux intervenants. Si les attitudes de deux sommets connectés sont les mêmes, l'arête associée est étiquetée comme *adhérente*. D'autre part, si les attitudes diffèrent, l'arête est étiquetée comme *opposante*.

Ceci conclut notre aperçu sur les techniques de prédiction utilisées sur les corpus politiques. Dans les paragraphes qui suivent, nous introduisons brièvement l'extraction de sujets et présentons comment elle est utilisée dans les corpus de débats politiques.

« *Un topic model ou modèle de sujet est un modèle probabiliste permettant de déterminer des sujets ou des thèmes abstraits dans un document.* »¹

Les modèles de sujet sont basés sur l'idée générale qu'un document contient une multitude de sujets, où chaque sujet représente une distribution probabiliste de mots qui lui sont reliés. Par exemple en lisant un document politique, on pourrait se rendre compte qu'il traite de la manière dont l'utilisation de l'informatique pourrait améliorer les procédures administratives. Cette conclusion serait liée au fait que nous avons découvert dans le texte les mots *data analysis*, *computer*, *fast*, qui caractérisent l'*informatique* et *administration*, *procedures*, *legal*, qui caractérisent les *procédures administratives*. Nous venons ainsi d'identifier deux sujets, *informatique* et *procédures administratives*. Les modèles de sujets sont des algorithmes qui permettent d'automatiser ce processus pour extraire la structure thématique d'un document. Cette structure thématique est présentée généralement comme un ensemble de sujets, où chaque sujet est constitué d'un ensemble de mots. On a aussi la distribution des sujets dans le document, c'est-à-dire le pourcentage de contenu dédié à chaque sujet. Dans les travaux sur les corpus politiques, le but de l'extraction de sujets est la découverte de patrons dans les activités gouvernementales

¹ https://fr.wikipedia.org/wiki/Topic_model

[10,12] ou l'identification des centres d'intérêt des politiciens [11]. On veut savoir, par exemple, s'il existe des groupes de politiciens qui discutent de certains sujets [10] ou identifier le changement des sujets discutés dans le temps [11, 12]. Les approches utilisées se basent sur des techniques d'analyse de réseaux sociaux [26], sur des métriques traditionnelles en recherche d'information telles que TF-IDF [22], ou encore sur des méthodes telles que le *Dynamic Topic Modeling* [27] ou le *Non negative Matrix Factorization* (NMF [29]) pour identifier l'évolution temporelle des sujets discutés.

Par exemple, en [10], des groupes de politiciens qui débattent de certains sujets sont identifiés à partir des débats du parlement allemand. L'approche construit un graphe où les sommets représentent les politiciens impliqués dans le débat et les arêtes orientées représentent le nombre d'interruptions entre deux politiciens. Après la construction du graphe, des *communautés* sont ensuite extraites. En analyse de réseaux sociaux, une communauté est un groupe de sommets densément connectés entre eux, mais faiblement connectés aux autres sommets. Chaque communauté extraite est donc une évidence d'un regroupement de politiciens qui débattent de certains sujets particuliers. Une autre approche d'extraction de sujets à partir de débats politiques est proposée en [11]. Les auteurs veulent démontrer que chaque membre du parlement possède un ensemble de sujets politiques qu'il préfère aborder. Pour l'identification des intérêts des membres, les auteurs utilisent un modèle de sujet appelé *modèle dynamique de sujet* (*Dynamic Topic Modeling* [27] ou *DMT*). Étant donné un nombre T de sujets déterminés par l'utilisateur ou par entraînement sur un grand corpus, ce type de modèle associe à chaque mot la probabilité d'appartenir à chacun des sujets. Les auteurs utilisent aussi des techniques de visualisation et d'analyse de données appelées TDA (*Topological Data Analysis* [28]) pour évaluer l'évolution temporelle des sujets discutés. Pour mesurer de manière quantitative l'intérêt d'un sujet pour un politicien, les auteurs introduisent plusieurs métriques qui capturent entre autres le temps passé par un membre sur un sujet par rapport aux temps passé par les autres membres sur le même sujet, et la diversité des sujets abordés par chaque membre.

L'identification des sujets discutés par chaque politicien ou chaque parti politique est aussi un des aspects comblés dans notre recherche. Toutefois, nous utilisons des approches différentes de celles utilisées en [11]. Alors que les auteurs en [11] extraient une liste de sujets où la seule relation entre les mots est le fait d'appartenir au même sujet, nous générons un graphe conceptuel avec des relations sémantiques riches entre les différents concepts. Aussi, nous ne nous contentons pas

d'utiliser de simples mots clés provenant du document, mais extrayons en utilisant des annotateurs sémantiques des concepts provenant de la base de connaissances DBpedia. Comme résultat de notre approche, nous générons un graphe conceptuel représentant les débats. Même si d'autres structures de graphes ont été utilisées dans les travaux politiques [7, 10], une approche qui considère des concepts comme sommets du graphe n'a pas été considérée à notre connaissance. Ce genre de représentation est nécessaire pour lier le corpus au Linked Data et ainsi bénéficier des enrichissements qu'il contient. Dans le prochain chapitre, nous présentons notre approche pour la construction du graphe.

2.5 Notre approche

Certains travaux [33, 39] ont abordé (comme nous) la représentation de documents textuels sous forme de graphes conceptuels, pour abstraire le contenu des documents et le lier à une base de connaissances. [39] utilisent les bases de connaissances VerbNet et Wordnet pour extraire les **rôles sémantiques** dans les phrases. Extraire les rôles sémantiques consiste à identifier les acteurs qui participent à une action et le rôle de chacun des acteurs dans l'action. Par exemple dans la phrase *John hits Bill*, les acteurs sont *John* et *Bill*, *John* a le rôle d'**agent**, *Bill* a le rôle de **patient** et l'action est *Hit*. Les auteurs produisent un graphe avec les acteurs comme nœuds et les rôles sémantiques comme arc. En [33], les auteurs extraient des concepts DBpedia provenant d'un document textuel en utilisant l'outil *Wikifier*. Les auteurs utilisent ensuite l'analyseur syntaxique *C&C* et l'outil *Boxer* pour extraire les relations entre ces concepts. Les relations sont successivement abstraites avec un jumelage à WordNet et désambiguïsées. Finalement, les auteurs définissent un modèle basé sur graphe RDF pour représenter les entités et les relations extraites.

La construction de graphe conceptuel et la création de liens entre un document textuel à une base de connaissances structurée sont des aspects que nous traitons dans notre travail. Par contre, à différence de [33] et [39] nous nous concentrons sur l'abstraction de contenu en appliquant à nos graphes conceptuels des techniques de filtrage. Nous nous attaquons aussi à de grands corpus composés de plusieurs documents, alors que les travaux présentés traitent un document à la fois. Les représentations proposées par les auteurs, bien que structurées, sont très verbeuses et difficilement utilisables par des humains. Elles seraient adéquates si le graphe produit devait être

utilisé seulement pour le traitement automatique. Notre travail est similaire aux précédents en ce qui concerne la technique utilisée pour extraire les concepts. Nous faisons usage d'annoteurs sémantiques pour extraire les concepts et les lier au Linked Open Data. Par contre, pour l'extraction de relations, nous avons opté pour l'analyse de surface, à la différence des auteurs qui utilisent l'analyse syntaxique ; nous justifions ce choix par le fait que notre corpus est formé de transcriptions de discours oraux, ce qui pourrait brouter l'analyse syntaxique. En plus, l'analyse de surface est beaucoup moins coûteuse en ressources et donc idéale pour traiter plusieurs documents dans un grand corpus. Nous effectuons aussi l'extraction de relations provenant du LOD. Cet aspect n'est pas pris en compte dans les travaux [33] et [39].

CHAPITRE 3 MÉTHODOLOGIE

Dans ce chapitre, nous présentons la méthodologie de recherche utilisée pour la construction du graphe résumé abstrait. Nous décrivons d'abord notre corpus d'application. Ensuite, après avoir présenté l'architecture générale adoptée, nous expliquons en détail chacun de ses différents modules.

3.1 Le corpus Hansard

Le cas d'utilisation choisi est le Hansard, un recueil des débats du Parlement canadien. Celui-ci rend disponibles sur son site internet les débats de la chambre des communes de 1994 à aujourd'hui. Les débats sont organisés en fonction des sujets traités. Pour chacune des dates où a eu lieu un débat parlementaire, un fichier XML correspondant est fourni sur le site du parlement. La structure de chaque fichier XML est la suivante :

- Le fichier XML est composé de plusieurs *ordres de travail*. Les ordres de travail sont les différentes phases de débats. En général, elles se composent de :
 - Initiatives parlementaires
 - Initiatives ministérielles
 - Déclarations des députés
 - Questions orales
 - Affaires courantes
 - Motions d'ajournement
- Un *ordre de travail* est composé de plusieurs *sujets de travail*. Un sujet de travail représente un des sujets débattus à une date donnée. Il peut être lié à un projet de loi ou traiter simplement de problèmes courants. Quelques exemples de sujets de travail sont : Loi sur la protection du Canada contre les terroristes, les affaires étrangères, le logement, etc.
- Chaque sujet de travail comprend plusieurs *interventions de députés*.

Pour chaque intervention du corpus, on extrait le contenu de l'intervention (ce que la personne dit), ainsi que de l'auteur de l'intervention, son rôle au parlement et le parti politique auquel il est affilié. Les caractéristiques du corpus sont fournies au tableau 3.1.

Tableau 3.1. Caractéristiques de 1217 débats du corpus Hansard pour les périodes d'avril 2006 à mars 2016.

Caractéristique	Minimum	Maximum	Moyenne
Nombre d'interventions par date	19	2004	254
Nombre d'auteurs par date	6	253	89
Nombre de sujets par date	5	88	47
Nombre de phrases par intervention	1	787	11
Nombre de mots par intervention	2	16880	245

Les débats sont disponibles sur le site du parlement en format XML en français et en anglais. Pour ce travail, nous avons traité uniquement les débats en anglais, essentiellement à cause de la disponibilité d'outils pour le traitement automatique de la langue naturelle dans cette langue.

L'extraction de ces données nous permet de faire des analyses à plusieurs niveaux différents. On peut analyser le contenu des débats parlementaires sur plusieurs dates, ou se restreindre plutôt à une seule date. On peut aller plus loin en sélectionnant un ordre de travail particulier, un sujet de travail particulier, un parti politique ou un parlementaire particulier. On peut aussi combiner plusieurs critères : par exemple, on pourrait choisir plusieurs dates, et parmi celles-ci restreindre l'analyse à certains sujets ou certaines affiliations. En général, on veut pouvoir combiner plusieurs critères et générer un graphe résumé correspondant.

Après avoir extrait les interventions et les métadonnées associées (date de l'intervention, nom de l'intervenant, son rôle dans le parlement, son affiliation politique), nous construisons un graphe abstrait qui représente le corpus. Dans la section 3.2, nous présentons l'architecture générale du système mis en place pour la construction du graphe. Dans les sections 3.3 et 3.4 nous détaillons les modules de notre système qui s'occupent d'extraire les concepts provenant du corpus et les relations entre ces concepts. Finalement, dans la section 3.5, nous expliquons comment le graphe final est construit.

3.2 Architecture générale du système

Plusieurs phases de traitements sont nécessaires pour la génération du graphe résumé à partir des débats parlementaires. La figure 3.1 représente l'architecture adoptée, qui est composée de deux modules principaux : *Identification des entités* et *Extraction de relations*.

Le premier module, *Identification des entités*, reçoit en entrée le corpus à résumer et des configurations de l'utilisateur sur le type de graphe qu'il souhaite générer (par exemple: graphe relatif à une affiliation politique, un député particulier, un sujet, etc. ou une combinaison de ces éléments). Ce module extrait des annotations sémantiques du corpus. Ces annotations consisteront les sommets du graphe. Le module *Identification des entités*, après avoir reçu le corpus brut des débats, extrait d'abord toutes les interventions et les métadonnées nécessaires pour le traitement. Après, il extrait les annotations sémantiques provenant du corpus avec un annotateur sémantique. Certaines des annotations retournées par l'annotateur ne représentent pas la forme de surface correcte du concept extrait. Par exemple pour la phrase suivante:

The winter of 2013-14 saw a transportation crisis that impeded the growth and credibility of our export economy.

L'annotateur retourne, entre autres, l'annotation *export economy*. L'annotation correcte dans ce cas serait *credibility of our export economy*. Les annotations sont donc corrigées par un processus d'extension. Toutes les annotations (étendues et non étendues), sont agrégées afin d'abstraire davantage le graphe que nous construisons. Par exemple, si nous avons deux annotations *child* et *children*, elles seront représentées par la même annotation agrégée. Ce processus permettra à notre système d'avoir de meilleures performances. Des matrices de cooccurrences sont finalement

générées avec les annotations agrégées. Les cooccurrences entre les annotations sont utilisées comme indice pour détecter la présence de relations sémantiques.

Le deuxième module, *Extraction de relations*, extrait des relations entre les concepts à partir des matrices de cooccurrence obtenues précédemment. Ces relations représenteront les arêtes dans le graphe résumé. Ce module, à partir des matrices de cooccurrences, extrait d'abord des relations (de type sujet – relation – objet) entre les annotations. Ces relations sont ensuite abstraites et désambiguïsées avec une association aux classes de la base de connaissances VerbNet [43]. Finalement, des relations sémantiques provenant de DBpedia sont extraites entre les annotations.

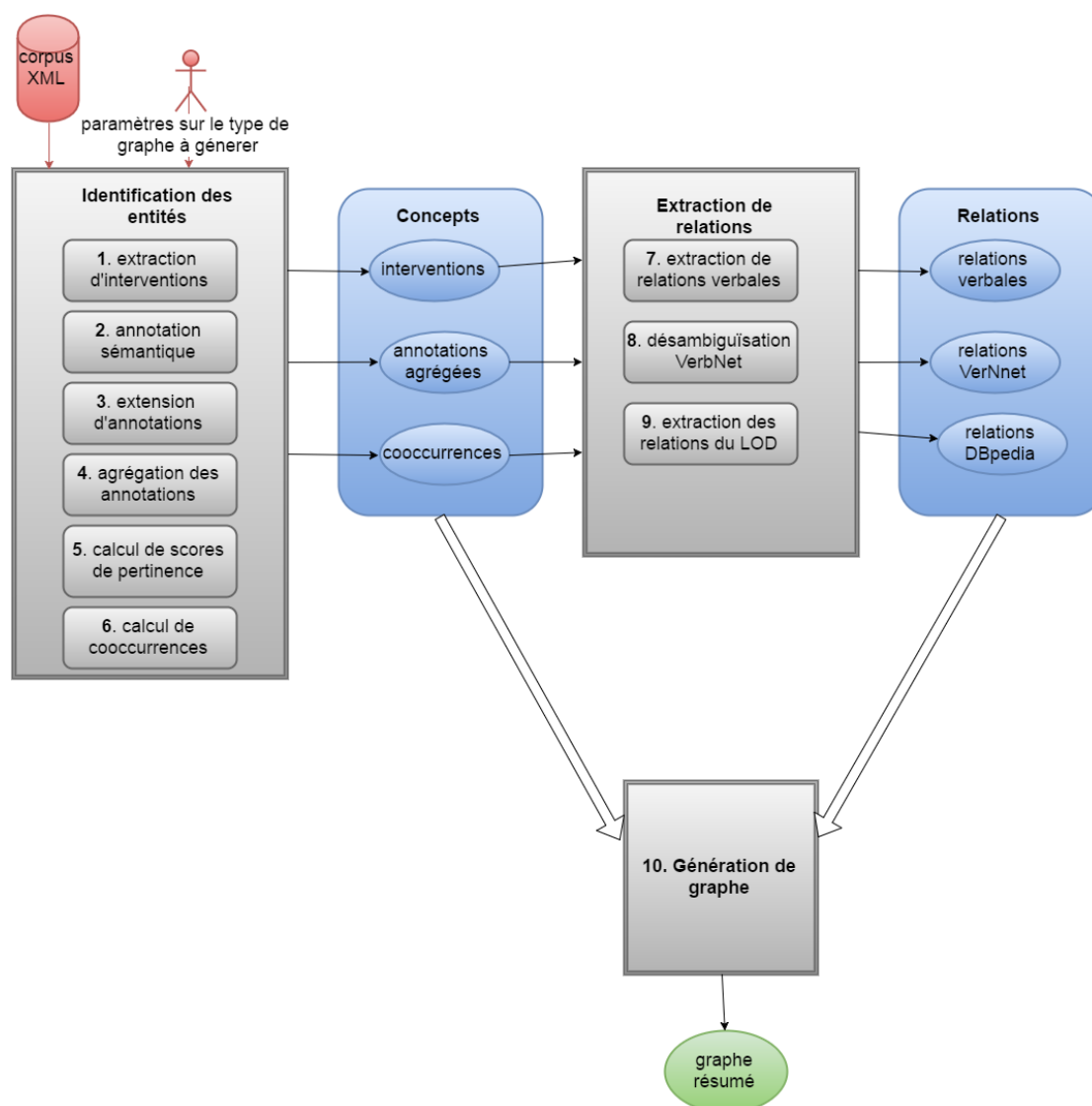


Figure 3.1. Architecture générale du système

Au terme de tous ces traitements, on récupère les résultats des diverses phases et on génère un graphe résumé. Les différents modules et sous-modules sont détaillés dans les sections qui suivent.

3.3 Identification des entités

L’objectif de cette phase est d’identifier les concepts les plus importants dans le texte. Pour cela, nous nous appuyons sur un annotateur sémantique.

3.3.1 Annotations sémantiques

Dans la revue de littérature, nous avons décrit l’annotation sémantique. Rappelons que l’annotation consiste à lier des portions d’un document à une ressource existante dans une base de connaissances. En général, les annotations nous permettent de comprendre de quoi parle le texte et y ajoutent de l’information complémentaire (contenu dans la base de connaissances).

L’analyse effectuée dans le cadre de ce projet est réalisée à partir d’annotations sémantiques retournées par l’annotateur *AlchemyAPI*². Il existe de nombreuses autres alternatives, telles que *DBpedia Spotlight*³ et *Open Calais*⁴. Le choix d’*AlchemyAPI* repose sur les études d’évaluation de performances des annotateurs sémantiques [82, 83], qui l’indiquent souvent comme étant parmi les meilleurs annotateurs actuels.

On envoie le contenu textuel de chaque intervention à *AlchemyAPI*, qui retourne un ensemble d’annotations regroupées dans les catégories suivantes :

- Des **mots clés** : ils proviennent tous du contenu textuel de l’intervention à laquelle ils sont associés (ils sont donc tous ancrés). Ils peuvent être considérés comme les mots ou expressions les plus importants du texte.

² <http://www.alchemyapi.com/>

³ <https://github.com/dbpedia-spotlight/dbpedia-spotlight/>

⁴ <http://http://www.opencalais.com/>

- Des **concepts** : ce sont les idées abstraites ou sujets discutés dans le texte indépendamment du fait qu'ils soient mentionnés explicitement ou pas (ils ne sont donc pas nécessairement ancrés). Les concepts ont tous des liens vers une ressource du LOD (DBpedia, Freebase, etc.).
- Des **entités nommées** : elles identifient les personnes, des organisations, et toute autre entité «typée» discutée dans le texte. Elles sont similaires aux concepts, car elles fournissent (parfois) des liens vers le LOD. Mais à la différence des concepts, les entités sont toujours ancrées dans le texte.
- Des **catégories** : elles fournissent aussi les sujets les plus importants du texte, mais elles sont classées selon une taxonomie prédéfinie, contrairement aux concepts. Par ailleurs, elles ne sont pas ancrées.

L'annotation sémantique est généralement effectuée en deux phases. La première, appelée identification des mentions, repère les expressions du texte qui pourraient indiquer une ressource dans une base de connaissances. Dans la deuxième phase, on associe les expressions repérées aux ressources correspondantes dans la base de connaissances. Bien qu'AlchemyAPI soit considéré comme étant un des meilleurs annotateurs, nous avons identifié certaines lacunes dans le processus d'identification de mentions dans le texte. Notamment, les annotations retournées sont parfois partielles et n'identifient pas toujours les formes de surface correctes. Considérons l'exemple suivant :

*The **Canada Transportation Act review** could not be more timely.*

Pour ce texte, AlchemyAPI retourne les annotations *Canada Transportation* et *Act review* comme deux annotations séparées. Or l'annotation correcte dans ce cas serait *Canada Transportation Act review*. Dans le but de corriger les annotations retournées par AlchemyAPI, nous avons mis en place un mécanisme d'extension d'annotations.

3.3.2 Extension des annotations

L'extension d'annotations corrige les annotations extraites par AlchemyAPI en les étendant au plus long syntagme nominal. Elle est appliquée sur chacune des interventions séparément. Le système d'extension d'annotations que nous avons mis en place prend en entrée les annotations d'AlchemyAPI et produit comme résultat un ensemble d'annotations étendues.

L'extension des annotations est basée sur des règles ou des patrons construits manuellement en utilisant *l'étiquetage morphosyntaxique* du texte en entrée. Ce type d'étiquetage permet d'associer à chaque mot sa catégorie grammaticale, appelée aussi *catégorie morphosyntaxique* ou *part of speech* en anglais (POS). Les catégories grammaticales peuvent être des verbes, des adjectifs, des adverbes, etc. La granularité des catégories est plus ou moins fine selon l'outil utilisé pour l'analyse. Afin de pouvoir effectuer l'extension d'annotations, nous associons donc à chacun des mots des interventions leurs catégories morphosyntaxiques au moyen de l'outil *TreeTagger*⁵. À chacune des annotations ancrées retournées par AlchemyAPI, nous appliquons ensuite le processus d'extension. Ce processus retourne, pour les annotations qui doivent être corrigées, une nouvelle annotation étendue.

Pour implanter notre mécanisme d'extension, nous avons considéré, parmi l'ensemble des catégories grammaticales existantes, les catégories grammaticales suivantes :

- N : nom propre ou nom commun
- IN : préposition
- JJ : adjectif
- PP\$: pronom possessif
- DT : déterminant
- POS : terminaison possessive
- V : catégorie générique qui représente les verbes (peu importe le temps de conjugaison). Cette catégorie inclut : VVN (participe passé), VBG (gérondif), VVD (passé antérieur), VVZ (troisième personne du présent de l'indicatif).
- CD : nombre cardinal

La première étape du processus consiste à rechercher les différentes catégories grammaticales mentionnées à droite et à gauche de l'annotation candidate à l'extension. Pour être candidate à l'extension, une annotation d'AlchemyAPI doit être un syntagme nominal ancré. Pour vérifier cette

⁵ <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

dernière condition, nous contrôlons les catégories grammaticales des mots qui composent l'annotation en question. Pour mieux comprendre, considérons l'exemple suivant :

Severe firearms control is our priority.

Dans cette phrase, l'annotation *control* est retournée par AlchemyAPI. *Control* étant de la catégorie grammaticale N (nom), on a bel et bien un syntagme nominal. L'annotation sera donc étendue vers la gauche. On aura donc comme annotation résultante *firearms control*.

Quand les conditions sont réunies pour que l'annotation soit candidate à l'extension (on a un syntagme nominal ancré), on procède au processus d'extension. Chaque patron d'extension définit les catégories grammaticales de chacun des mots que l'annotation étendue devrait avoir. Les patrons implémentés dans ce travail sont définis sous forme d'expressions régulières au tableau 3.1. Observons le premier patron P1 du tableau. Celui-ci indique que l'annotation étendue correspondante est une succession de noms (communs ou propres). Soit le contenu textuel suivant :

*The **Canada Transportation Act** review could not be more timely.*

Supposons qu'AlchemyAPI ait retourné comme annotation *Canada Transportation*. Cette annotation a la catégorie grammaticale N N parce que les mots *Canada* et *Transportation* sont des noms. Maintenant, si on observe les mots autour de l'annotation, on se rend compte que *Act* et *review* sont aussi des noms. On a donc que *Canada Transportation Act review* a la catégorie grammaticale N N N N, ce qui respecte le patron d'extension P1 du Tableau 3.2. L'annotation *Canada Transportation Act review* sera donc retournée comme annotation étendue.

Tableau 3.2: Liste des patrons d'extension d'annotation exprimés sous forme d'expressions régulières. Les catégories grammaticales des exemples sont indiquées en indice.

ID	Patron d'extension d'annotation	Exemples
P1	CD? (JJ VVN)? N N+	life _N → life _N imprisonment _N
P2	CD? (JJ VVN)? N* IN (JJ DT DT JJ)? (CD N) N*	life _N → sentence _N of _{IN} life _N imprisonment _N life _N → end _N of _{IN} natural _{JJ} life _N officer _N → supervision _N of _{IN} a _{DT} CSC _N parole _N officer _N murder _N → reprehensible _{JJ} forms _N of _{IN} murder _N murder _N → seriousness _N of _{IN} the _{DT} murder _N murder _N → four _{CD} categories _N of _{IN} murder _N murder _N → premeditated _{VVN} murder _N of _{IN} a _{DT} single _{JJ} victim _N
P3	N IN PP\$ JJ ? (CD N) N*	speech _N → end _N of _{IN} my _{PP\$} speech _N process _N → part _N of our _{PP\$} legislative _{JJ} process _N
P4	N POS N	government _N → government _N 's _{POS} record _N safety _N → Canadian _N 's _{POS} safety _N

Dans les paragraphes qui suivent, nous présentons deux exemples d'interventions du Hansard annotées. Dans ces exemples, les annotations d'AlchemyAPI sont marquées en gras et les annotations étendues, entre crochets.

Intervention annotée 1 :

Mr. Speaker, it is truly my pleasure to address the [**House** of Commons] today at [second reading of **Bill C-50**], the [citizen **voting** act]. As was already explained, the bill proposes **electoral reforms** and aims at strengthening [Canada 's **democracy**] by doing two

things. One, it would reinforce the [integrity of the **special ballot voting system**]. Two, it would ensure fairness for resident and **non-resident voters** alike. Despite these **laudable goals**, which should command the [consent of **the House**], we have heard criticism from [members of **the House**] directed at the bill. Some have alleged that this reform could create chaos with **voting** in **Canada**. Nobody understands why we would create chaos in **Canada** with this kind of legislation, but the opposition would allege it. One member suggested that the bill created an **administrative nightmare** for **non-resident voters**. These allegations are fabrications and certainly, overstatements of the facts. The proposed electoral reform was carefully developed and its impacts were appraised. In truth, the reforms are more in **nature procedural adjustments** than a **substantive overhaul** of the special ballot system. I would like to use my time today to review the **new rules** from an **operational perspective**. By this exercise, I wish to demonstrate that the [concerns of **certain members of the House**] should be allayed. I therefore propose to review the steps that **non-resident electors** would have to follow in order to vote in a [general **election**] under the **new rules**. As is known by every member in **the House**, an **election** begins with the issue of the writs. The Governor in **Council** **issues** a proclamation that directs the [Chief Electoral Officer] to issue a writ to the returning officer for each electoral district, fixes the date of the issue of the writ, and then fixes the date for **voting**. This date must be at least **36 days** after the issuance of the writs. After the issuance of the writs, **non-resident voters** need to apply for a special ballot. I wish to point out that this is currently what **resident electors** must do if they plan not to be in their electoral district on **election day**, either because they are vacationing abroad or for another reason. There is no **administrative nightmare** here, I would suggest. I do not see why applying the same requirement to **non-resident voters** would make it one. The [**special ballot** application] should be available on the [website of **Elections Canada**], at any **Canadian embassy, high commission** or consulate, or by calling [**Elections Canada**]. **Non-resident voters** would have to include with their [**application proof** of identity] and proof of residence. [Proof of **prior residence**] in **Canada** is currently not required for **non-resident voters**. I would hope members would agree that it is reasonable to require **non-resident electors** to prove that they have previously resided in **Canada**. I think most Canadians would think that just makes **eminent sense**. Such proof of residence is already required from **resident electors**. We are only asking for the same for **non-resident electors**. **Non-resident voters** would have **various options** to prove their identity and last place of residence in **Canada**. They could provide copies of a single piece of identification issued by a Canadian government or an agency of that government with a photograph, name and address. For example, a driver's licence would be accepted, even if it has expired. A second option is to provide two pieces of identification from the list provided by the [Chief Electoral **Officer**], one with

an address and both with a name. I invite members of the **House** to view the list of the types of ID that have been authorized by the [Chief Electoral **Officer**]. They will see that more than 45 types of identification are currently accepted.

Intervention annotée 2 :

Mr. Speaker, we work with the **municipalities** more than the **NDP** does. The **New Democrats** are making **promises**, but we are keeping promises. I would like to remind my **colleague** that [96 % of **Canada's infrastructure**] belongs to the **municipalities** and **provinces**. We have been their *partner* since 2006. No [other **government**] has invested as much as ours in **infrastructure**, while respecting *jurisdictions*. The **NDP** wants to centralize **decision making** in **Toronto**. We are working with our **partners** while respecting their **jurisdictions**.

Dans les exemples ci-dessus, on peut voir certaines annotations d'AlchemyAPI que nous étendons par la suite. L'intervention annotée 2 sera réutilisée dans les sections suivantes pour exemplifier les autres aspects de notre méthodologie.

La phase d'extension d'annotation est effectuée séparément pour chacune des interventions de chaque débat. Nous avons donc, pour chacune des interventions, en plus des annotations d'AlchemyAPI, un enrichissement avec les annotations étendues. Le but de ce travail étant l'analyse du corpus, les annotations obtenues doivent être aussi analysées dans leur ensemble. Aussi, on voudrait considérer comme équivalentes les diverses formes de surfaces dans lesquelles un concept peut apparaître dans le texte. Nous avons donc mis en place un processus d'agrégation pour pouvoir considérer les annotations au niveau macroscopique et ainsi effectuer l'analyse du corpus dans son ensemble.

3.3.3 Agrégation d'annotations

Dans les sections précédentes, nous avons extrait les annotations et les annotations étendues pour chaque intervention. Le but de cette phase est de créer une classe d'équivalence qui considère comme identiques les annotations qui partagent le même noyau après lemmatisation. En linguistique, le noyau ou la tête d'un syntagme nominal est le mot principal du syntagme. Par exemple, pour le syntagme *premeditated murder*, le noyau est *murder*. Pour le syntagme *vision of a society*, le noyau est *vision*.

La lemmatisation est le processus de transformation d'une forme fléchie d'un mot à sa forme canonique appelé *lemme*. Le lemme est le mot choisi par convention pour représenter toutes les formes d'une flexion. On lemmatise donc toutes les annotations et on regroupe toutes les annotations qui partagent le même noyau après lemmatisation. Une *annotation agrégée* est donc créée pour représenter un regroupement d'annotations qui partagent le même lemme. Alors que la portée d'une annotation est l'intervention d'où elle a été extraite, l'annotation agrégée est relative au débat complet.

Comme exemple, supposons qu'on ait un débat avec deux interventions. La première contient la phrase suivante:

*They would like to see legal protection for **children** both before and after birth.*

La deuxième a pour contenu :

*It is also fair to point out that there have been calls from many parents who are paying **child** support, who are mostly men.*

Supposons que dans la première intervention, AlchemyAPI ait retourné comme annotation *child* et pour la deuxième, il ait retourné *children*. Ces deux annotations, même si elles apparaissent différemment, partagent le même lemme *child*. Dans le processus d'agrégation, une seule annotation agrégée *child* sera créée pour représenter ces deux annotations. L'annotation agrégée aura des références à chacune des formes de surfaces dans les deux interventions.

Une fois les annotations extraites et agrégées, notre objectif est d'identifier les annotations les plus importantes afin d'analyser le débat au complet. On pourra ainsi répondre à plusieurs requêtes telles que : quelles sont les annotations les plus fréquentes ou les plus pertinentes du débat ? Quelles sont leurs catégories ? Sont-elles présentes dans la base de connaissances DBpedia ? On associe donc des scores de pertinence aux annotations agrégées. Les détails sont fournis dans la section qui suit.

3.3.4 Calcul des scores d'annotations

On considère trois métriques différentes dans ce travail pour mesurer l'importance ou la pertinence d'une annotation : la *fréquence d'occurrence*, ou *term frequency* en anglais (*TF*), un score de

pertinence (relevance score, retourné par AlchemyAPI) et le *TF-IDF* (*Term Frequency - Inverse Document Frequency*).

Fréquence d'occurrence :

Ainsi que son nom l'indique, TF_i (Term Frequency) représente la fréquence d'occurrence de l'annotation agrégée i dans le corpus ou le sous-ensemble du corpus que nous considérons pour la construction du graphe conceptuel (par exemple, si on pourrait se limiter à un seul sujet de travail particulier d'une date du Hansard). Une annotation agrégée est composée de plusieurs annotations de surface, chacune d'elles associée à une intervention particulière. Pour chacune des annotations de surface, on calcule sa fréquence d'occurrence dans l'intervention où elle apparaît. On associe donc une fréquence d'occurrence à l'annotation agrégée comme la somme des fréquences d'occurrence de chaque annotation de surface qui la compose. Plus formellement :

Soit $i = \{i_1, i_2, i_n \dots i_n\}$ où i est une annotation agrégée et i_j est une annotation de surface représentée par i . TF_{i_j} représente la fréquence d'occurrence de l'annotation i_j dans l'intervention d'où elle a été extraite par AlchemyAPI. La fréquence d'occurrence de l'annotation agrégée est calculée comme suit :

$$TF_i = \sum_{j=1}^n TF_{i_j}$$

Score de pertinence d'AlchemyAPI

Le score de pertinence est un score retourné par AlchemyAPI pour chaque annotation. Ce score représente la pertinence de l'annotation pour l'intervention d'où elle a été extraite. Un score de pertinence est donc associé à une annotation agrégée en faisant la moyenne des scores de pertinence des annotations associées à l'agrégation. Les annotations étendues, ne provenant pas d'AlchemyAPI, n'ont pas de score de pertinence. Nous leur assignons le score de l'annotation d'AlchemyAPI d'où elles proviennent. Par exemple si AlchemyAPI retourne l'annotation *life*, qui est étendue à *sentences of life imprisonment*, nous associerons à *sentences of life imprisonment* le score de pertinence de *life*. De manière générale, les annotations étendues héritent des scores (TF, pertinence et TF-IDF) des annotations originales (les annotations d'où elles proviennent).

Term Frequency – Inverse Document Frequency (TF-IDF)

TF-IDF est une métrique bien connue de la recherche d'information [79] qui, dans notre contexte, permet de déterminer l'importance d'une annotation agrégée i pour une intervention j à l'intérieur d'un ensemble d'interventions D (corpus de débats). Cette métrique est calculée à partir des éléments TF et IDF :

IDF _{i} (Inverse Document Frequency) représente l'inverse du nombre d'interventions qui contiennent l'annotation agrégée i . Le pouvoir discriminatif d'une annotation agrégée est inversement proportionnel au nombre d'interventions qui la contiennent. Certaines annotations agrégées, comme *Mr Speaker* ou *government*, sont présentes dans presque tous les documents. Elles ne sont donc pas discriminantes.

La formule générale du TF-IDF est la suivante :

$$tfidf_{i,j} = tf_{i,j} * idf_i$$

Où $tf_{i,j}$ représente la fréquence de l'annotation agrégée i dans l'intervention j et idf_i est calculé comme suit :

$$idf_i = \log \frac{|D|}{|\{d_j : i \in d_j, d_j \in D\}|}$$

Où $|D|$ représente le nombre total d'interventions dans le corpus et $|\{d_j : i \in d_j, d_j \in D\}|$ le nombre d'interventions où l'annotation agrégée i apparaît.

Dans notre travail, nous utilisons une variante de TF-IDF, car nous ne voulons pas connaître l'importance d'une annotation agrégée pour une intervention, mais plutôt son importance dans un sous-ensemble du corpus (par exemple un ensemble de débats ou un sujet de travail). Dans cette variante, TF _{i} est la fréquence d'occurrence de l'annotation agrégée i telle que définie plus tôt. L'IDF est calculé comme le logarithme de l'inverse du nombre d'interventions qui contiennent une des annotations de i :

$$idf_i = \log \frac{1}{|\{d_j : i \in d_j, d_j \in D\}|}$$

Dans notre variante, obtenue empiriquement, le IDF est élevé au carré ; nous faisons croître l'IDF de manière quadratique par rapport au TF pour mettre en évidence la composante discriminante.

La formule utilisée est la suivante :

$$tfidf_i = tf_i * idf_i^2$$

Où : tf_i est la fréquence d'occurrence de l'annotation agrégée i sur le sous-ensemble du corpus que nous considérons et idf_i tel que défini dans le paragraphe précédent.

Certaines annotations, comme *government* et *mr. Speaker*, apparaissent très fréquemment dans les débats. Elles ne sont pas discriminantes et on aimerait les faire disparaître avec le TF-IDF. Avec la formule classique du TF-IDF, ces termes ne disparaissent pas, car le IDF ne croit pas assez vite pour rendre dominante la composante discriminative. En élevant le IDF au carré, ces termes disparaissent. Cette formule, trouvée de manière empirique, donne de bons résultats.

Jusque-là, nous nous sommes concentré sur l'extraction des annotations, qui constitueront les sommets de notre graphe résumé. Il nous faut aussi trouver les relations entre les annotations pour former le graphe. Comme première solution, nous avons entrepris la construction d'un réseau de cooccurrence pour évaluer la distribution des annotations dans le corpus. L'intuition derrière cette approche est que les annotations qui apparaissent le plus souvent ensemble ont une relation sous-jacente.

3.3.5 Extraction de cooccurrences

Comme indiqué dans la revue de littérature, les relations de cooccurrence sont largement utilisées pour la représentation de texte sous forme de graphe [32,36]. Les cooccurrences sont considérées comme premier indicateur de la présence de relation entre deux termes. Les cooccurrences sont aussi utilisées pour extraire les sujets discutés à partir de corpus de débat [2]. Aussi, la plupart des techniques d'extraction de relations [45, 46, 61, 62] essaient d'extraire une relation entre deux entités dans la phrase où elles cooccurrent. En partant de ce principe, nous avons construit un réseau de cooccurrence pour étudier les relations (de cooccurrence) entre les annotations. Ce réseau de cooccurrences sera notre premier indicateur de la présence de relations sémantiques entre nos annotations. Nous utiliserons ensuite ces relations de cooccurrence pour extraire des relations labélisées, tel que discuté dans la section 3.4.

Une cooccurrence est une évidence que deux termes apparaissent simultanément dans un même contexte. Dans ce travail, on considère comme contextes possibles les phrases et les interventions.

Plus formellement, on représente la cooccurrence comme un triplet $c = (t_1, t_2, n)$ pour indiquer le fait que les termes t_1 et t_2 apparaissent dans n groupes distincts. n représente la fréquence de cooccurrence. Nous calculons deux types de cooccurrence : la cooccurrence par intervention et la cooccurrence par phrase.

La première matrice de cooccurrences (par intervention) nous sert pour construire un graphe de cooccurrence que nous appelons *graphe non labélisé*. Dans ce graphe, les sommets représentent des annotations agrégées et les arêtes pondérées représentent les relations de cooccurrences entre celles-ci. La deuxième matrice de cooccurrence (par phrase) nous sert pour l'extraction de *relations labélisées* et la construction d'un graphe labélisé. Nous expliquons la nature de ces relations et le processus d'extraction de relations dans la section qui suit. À la fin de cette phase, nous obtenons une matrice de cooccurrence qui représente les termes qui apparaissent ensemble dans une ou plusieurs interventions et une autre matrice de cooccurrences pour représenter les termes qui apparaissent ensemble dans une ou plusieurs phrases.

3.4 Extraction de relations

L'objectif de ce projet étant la production d'un graphe, l'extraction de relations en constitue un aspect très important. Nous nous concentrons ici sur l'extraction de relations dites labélisées entre les annotations agrégées.

3.4.1 Extraction de relations par analyse de surface

Dans la section 3.3.5, nous avons indiqué que nous construisons, en plus de la matrice de cooccurrence par intervention, une matrice de cooccurrence par phrase. En effet, si deux annotations apparaissent très souvent dans une même phrase, notre hypothèse est qu'il doit exister une relation entre celles-ci et l'objectif de cette phase est de trouver une étiquette indiquant la nature de cette relation. Nous utilisons donc la matrice de cooccurrence par phrases pour l'extraction de relations verbales. Pour chaque phrase où deux annotations cooccurrent, nous essayons d'extraire une relation de type sujet – relation – objet où relation est une relation verbale, et où sujet et objet sont des annotations cooccurrentes dans la phrase.

Plusieurs travaux de recherche ont été effectués dans le domaine de l'extraction de relations et il existe une multitude de techniques à ce sujet [54, 55, 57, 56, 53, 44, 45, 46, 47, 49, 50, 51, 52, 58, 59, 60, 41, 61, 62]. Dans notre travail, nous voulions une solution générique, s'appliquant à n'importe quel domaine ou corpus. Nous n'avons donc pas considéré les techniques existantes basées sur l'apprentissage machine, parce que lorsque ces techniques sont appliquées sur des corpus différents des corpus d'entraînement, les résultats pourraient ne pas être satisfaisants. Nous voulions aussi des techniques qui ne nécessitent pas de traitements linguistiques lourds comme l'analyse de dépendance syntaxique. Cette technique, bien qu'ayant de bonnes performances en général, pourrait être bruitée parce que notre corpus est constitué essentiellement de transcriptions de phrases orales. Un autre inconvénient de l'analyse de dépendance syntaxique est qu'elle requiert beaucoup de ressources computationnelles. Nous avons donc voulu vérifier dans quelle mesure l'analyse de surface basée seulement sur les catégories morphosyntaxiques nous permettait d'extraire des relations verbales intéressantes et pertinentes. L'analyse morphosyntaxique est beaucoup moins coûteuse que les autres techniques (arbre de dépendance, analyse syntaxique) et serait donc convenable pour traiter de grands corpus comme celui que nous utilisons.

Le processus d'extraction de relations est appliqué sur chacune des phrases où au moins deux annotations cooccurrent. En utilisant les catégories grammaticales, nous avons modélisé des patrons d'extraction de relations sous forme d'expressions régulières pour l'extraction des relations verbales.

Le Tableau 3.2 résume l'ensemble des patrons implantés, ainsi que des exemples pour chaque patron. Notons que pour chaque patron, la relation qui sera extraite est indiquée en gras. Dans le tableau, on utilise les symboles suivants :

- ANN désigne une annotation d'AlchemyAPI.
- VERB désigne un verbe.
- RP désigne une particule. Par exemple pour *give up*, on a la particule *up*.
- AUX désigne les auxiliaires *être* et *avoir* ainsi que les formes modales de l'anglais comme *could*, *should* et *will*.
- BE et HAVE représentent respectivement les verbes *être* et *avoir*.

- NOT représente tous les mots associés à la négation. Les mots *not*, *never* et *n't* seront associés à NOT.
- ^ indique l'expression régulière de la négation. Donc [[^]X] signifie qu'on peut tout avoir sauf X.
- X* indique 0 ou plusieurs occurrences successives de X.
- X ? indique que X est optionnel. On peut avoir zéro ou une occurrence de X.
- On réutilise les catégories grammaticales déjà définies dans la section 3.3.2 et d'autres catégories définies dans l'annexe C

Tableau 3.3 Patrons d'extraction de relations verbales

ID	Patron POS	Exemple
PR1	ANN [^ANN]* (BE HAVE) NOT (TO IN)? (CD DT JJR JJS PP\$)? ANN	- the <i>[liberal party]</i>_{ANN} was_{BE} not_{NOT} my_{PP\$} <i>[affiliation]</i>_{ANN} - the <i>[NDP]</i>_{ANN} have_{HAVE} not_{NOT} the_{DT} <i>[majority]</i>_{ANN} - the <i>[question]</i>_{ANN} is_{BE} not_{NOT} on_{IN} the_{DT} <i>[motion]</i>_{ANN}
PR2	ANN [^ANN]* AUX? NOT? VERB RP? (TO IN)? (CD DT JJR JJS)? ANN	- the <i>[liberal party]</i>_{ANN} will_{AUX} provide_{VERB} one_{CT} <i>[formal apology]</i>_{ANN} - the <i>[liberal party]</i>_{ANN} provided_{VERB} a_{DT} <i>[formal apology]</i>_{ANN} - the <i>[liberal party]</i>_{ANN} will_{AUX} provide_{VERB} greener_{JJR} <i>[initiaives]</i>_{ANN} - the <i>[liberal party]</i>_{ANN} will_{AUX} not_{NOT} provide_{VERB} a_{DT} <i>[formal apology]</i>_{ANN} - the <i>[liberal party]</i>_{ANN} will_{AUX} not_{NOT} apology_{VERB} in_{IN} their_{PP\$} <i>[discourse]</i>_{ANN} - that <i>[loophole]</i>_{ANN} would_{AUX} open_{VERB} up_{RP} the_{DT} <i>[opportunity]</i>_{ANN} for an abuse of the system

Tableau 3.3 : suite

ID	Patron POS	Exemple
PR3	ANN [^ANN]* AUX? NOT? VERB RP? TO VERB RP? (TO IN)? (CD DT JJR JJS)? ANN	- this [government]_{ANN} can_{AUX} continue_{VERB} to_{TO} protect_{VERB} [Canadians]_{ANN} [NDP]_{ANN} want_{VERB} to_{TO} see_{VERB} this [money]_{ANN}
PR4	ANN [^ANN]* AUX? NOT? VERB RP? [^(ANN VERB)]* TO VERB RP? (CD DT JJR JJS PP\$)? ANN	[force]_{ANN} try_{VERB} very_{RB} hard_{JJ} to_{TO} address_{VERB} [problem of voter apathy]_{ANN} [people]_{ANN} have_{VERB} an_{DT} ability_N to_{TO} change_{VERB} [electoral district]_{ANN} [Canadian]_{ANN} have_{VERB} the_{DT} right_{JJ} to_{IN} vote_{VERB} their _{PP\$} [representatives]_{ANN}
PR5	ANN [^ANN]* AUX? NOT? VERB RP? [^(ANN VERB)]* IN (CD DT JJR JJS PP\$)? ANN	[bill c-50]_{ANN} proposes_{VERB} important_{JJ} reforms_N to_{IN} [Canada 's election act]_{ANN}
PR6	ANN of VERB ANN	[idea]_{ANN} of preventing_{VERB} [crime]_{ANN} [new way]_{ANN} of doing_{VERB} [thing]_{ANN}

Reprenons l'intervention annotée 2 que nous avons considérée plus tôt :

Mr. Speaker, we work with the **municipalities** more than the **NDP** does. The **New Democrats** are making **promises**, but we are keeping promises. I would like to remind my **colleague** that [96 % of **Canada's infrastructure**] belongs to the **municipalities** and **provinces**. We have been their *partner* since 2006. No [other **government**] has invested as much as ours in **infrastructure**, while respecting *jurisdictions*. The **NDP** wants to centralize **decision making** in **Toronto**. We are working with our **partners** while respecting their *jurisdictions*.

En appliquant notre processus d'extraction de relations verbales à cette intervention, nous avons obtenu les quatre relations suivantes entre les concepts :

R1. New Democrats are making promises

R2. 96 % of Canada's infrastructure belongs to the municipalities

R3. government has invested as much as ours in infrastructure

R4. The NDP wants to centralize decision making

3.4.2 Désambiguïsation de relations basée sur VerbNet

En plus d'extraire des relations verbales, tel que décrit dans la section précédente, un des objectifs du projet est d'extraire des relations à un plus haut niveau d'abstraction, soit de leur associer une « sémantique » provenant d'un lexique sémantique. Notre approche d'extraction de relations s'inscrit dans la catégorie des relations non contraintes. Une des limites de ces approches est que les relations sont difficilement réutilisables par d'autres systèmes, vu que les relations pouvant être extraites ne sont pas connues à priori. Pour résoudre ce problème, nous avons utilisé VerbNet, un vocabulaire hiérarchique de verbes [43]. Il permet de regrouper plusieurs verbes en une catégorie commune en fonction de leurs attributs syntaxiques et sémantiques. En rattachant une relation verbale à une classe VerbNet, on l'enrichit en incluant de l'information sémantique contenue dans VerbNet. Dans ce cas, la relation extraite a un sens précis, et pourrait être interprétée par d'autres systèmes. En général, un verbe peut appartenir à une ou plusieurs classes VerbNet. Chacun des verbes extraits dans la phase d'extraction de relations verbales, est associée à la classe VerbNet qui

lui correspond. Dans la section suivante, nous introduisons brièvement la structure de la base de connaissances VerbNet avant de décrire le processus d'association.

VerbNet (VN) [43] est une grande base de connaissances de verbes en anglais. Cette base de connaissances est indépendante du domaine d'application et contient des correspondances vers d'autres bases de connaissances telles que Wordnet⁶, XTag⁷ et FrameNet⁸. Dans la base de connaissances VerbNet, il existe des classes génériques de verbes qui sont représentées avec une structure hiérarchique. Chaque classe VerbNet est définie par des propriétés syntaxiques et sémantiques qui capturent et généralisent le comportement des verbes qu'elle représente. Ces propriétés contiennent aussi des contraintes sur les réalisations de surface possibles qu'une classe VerbNet peut avoir. Ces réalisations sont appelées *member* (membre). Les membres d'une classe VerbNet partagent les mêmes propriétés syntaxiques et sémantiques. Certaines classes VerbNet ont des sous-classes. Les sous-classes héritent de toutes les propriétés des parents, et peuvent avoir des propriétés additionnelles. Par exemple sur la figure 3.2, on peut voir la classe *become-109.1*. Cette classe a huit membres, qui sont *come*, *come_out*, *fall*, *get*, *go*, *grow*, *turn* et *turn up*. Ces membres sont les formes de surface possibles de la classe *become-109.1*. La forme de surface *become* n'est membre directe de la classe *become-109.1*, mais plutôt de la classe *become-109.1.1*, qui est sous-classe de la classe *become-109.1*. Le processus d'abstraction de nos relations verbales à la base de connaissances VerbNet est expliqué dans le paragraphe suivant.

⁶ <https://wordnet.princeton.edu/>

⁷ <http://www.cis.upenn.edu/~xtag/>

⁸ <https://framenet.icsi.berkeley.edu/fndrupal/about>

become-109.1 <i>Members: 8</i>	
MEMBERS	
COME (WN 3; 6; G 3)	GROW (WN 1; G 1)
COME_OUT (WN 4; G 3)	TURN (WN 6; G 2, 3)
FALL (WN 3)	TURN_UP (WN 3; G 7)
GET (WN 2, 15, 7; G 3)	
GO (WN 4; G 4)	

Figure 3.2. Exemple de classe VerbNet

Pour abstraire une relation verbale, nous rattachons son *verbe noyau* aux classes VerbNet qui lui correspondent. La notion de verbe noyau est similaire au noyau utilisé pour les syntagmes nominaux. Nous définissons un verbe noyau d'une relation verbale comme le verbe principal de la relation, celui qui définit l'action. Par exemple pour *[criminal]_{ANN} [can be released on]_{REL} [parole]_{ANN}*, le verbe noyau est *release*. En général, pour associer une relation verbale à VerbNet, nous extrayons d'abord le verbe noyau de la relation. Ensuite nous cherchons les classes VerbNet qui contiennent la forme de surface du verbe de la relation. Pour chaque classe VerbNet trouvée, nous remontons dans la hiérarchie de VerbNet pour trouver la classe de plus haut niveau. Et finalement, on associe à la relation verbale ces classes VerbNet de plus haut niveau. Par exemple, si le verbe de la relation verbale est *become*, nous irons chercher dans VerbNet la classe correspondante qui est *become-109.1.1*. Ensuite nous remontons dans la hiérarchie de VerbNet pour trouver la classe de plus haut niveau dont *become-109.1.1* est une sous-classe. C'est la classe *become-109.1*. Finalement, *become-109.1* est associée à la relation verbale extraite.

La structure hiérarchique de VerbNet ne représente pas un arbre et donc un membre VerbNet peut être rattaché à plusieurs classes VerbNet. Par exemple, considérons la relation suivante :

[Conservatives] [take on] [Bill 51]

Dans ce cas, le verbe noyau *take* a six classes candidates qui sont *bring-11.3*, *characterize-29.2*, *convert-26.6.2*, *cost-54.2*, *fit-54.3*, *hire-13.5.3*, *performance-26.7.2*, *require-103* et *steal-10.5*.

Idéalement, on voudrait obtenir la classe VerbNet qui représente au mieux une relation verbale. Mais cela requiert un processus de désambiguïsation et, plus généralement, la définition de modèles

d'apprentissage complexes. Pour y remédier, nous avons intégré le projet ClearWSD⁹, disponible sur internet, au projet actuel. ClearWSD est un système de désambiguïsation développé par James Gung, un étudiant au doctorat supervisé entre autres par Martha Palmer, une des créatrices de VerbNet. ClearWSD prend en entrée une phrase, et retourne, pour chacun des verbes présents dans la phrase, la catégorie VerbNet qu'il considère comme la plus adéquate. Pour pouvoir désambiguïser les verbes, ClearWSD utilise l'apprentissage machine et des caractéristiques qui incluent des structures syntaxiques par dépendance, les catégories grammaticales et le thésaurus WordNet.

Dans les paragraphes précédents, nous avons expliqué le jumelage de chaque relation verbale r à une liste de classes VerbNet $L = \{v_1, v_2, \dots, v_n\}$ où v_i est une classe VerbNet. Nous effectuons donc un processus de désambiguïsation en passant la relation verbale r et son contexte à ClearWSD qui retourne une classe VerbNet v . Si $v \in L$, on retient v comme la classe VerbNet désambiguïsée de r .

Comme exemple, considérons les relations verbales provenant de l'intervention annotée 2. On applique à chacune de ces relations le processus de désambiguïsation VerbNet. La classe VerbNet retournée par ClearWSD est marquée en indice du verbe correspondant.

R1. New Democrats *are*_[seem-109.1.1] **making promises**_[future_having-13.3]

R2. 96% of Canada's infrastructure *belongs*_[] **to the municipalities**

R3. government *has invested*_[equip-13.4.2] **as much as ours in infrastructure**

R4. The NDP *wants to centralize*_[other_cos-45.4] **decision making**

Pour les relations R3 et R4, le verbe noyau de relation a été correctement désambiguïsé. Par exemple en R3, le verbe noyau *invest* a trois classes candidates, qui sont *equip-13.4.2*, *investigate-35.4* et *sight-30.2*. ClearWSD sélectionne la classe *equip-13.4.2*, dont le sens correspond ici au sens de la relation. Pour la relation R4, le verbe noyau *centralize* a une seule classe candidate, qui est *other_cos-45.4*. La relation R4 n'a donc pas besoin d'être désambiguïsée. Les relations R1 et R2 ont subi des erreurs dans le processus de désambiguïsation. Pour la relation R1, ClearWSD ne

⁹ <https://github.com/jgung/ClearWSD>

reconnait pas le verbe noyau *make*. C'est plutôt l'auxiliaire *be* qui est reconnu et désambiguïsé. On peut voir en R1 que l'annotation *promises* est désambiguïlée à la classe *future_having-13.3*. Ceci est une erreur de ClearWSD, car *promises* dans le texte n'apparaît pas comme verbe et ne devrait donc pas être désambiguïsé. Nous excluons toute désambiguïstation qui n'est pas associée à une relation extraite. Une évaluation détaillée de la désambiguïstation est fournie au chapitre 4. Pour la relation R2, la désambiguïstation n'est pas possible, car le verbe *belong* n'existe pas dans la base de connaissances VerbNet. Jusque-là, nous avons extrait via l'analyse de surface des relations verbales qui proviennent du corpus que nous avons successivement désambiguïsées. À ce niveau, nous voulons enrichir ultérieurement notre graphe résumé en y ajoutant des relations qui proviennent du Linked Open Data, précisément de la base de connaissance DBpedia. La section qui suit décrit notre processus d'extraction des relations provenant de DBpedia.

3.4.3 Extraction de relations à partir de DBpedia

Plusieurs annotations extraites par des annotateurs sémantiques sont des concepts du Linked Open Data. Un autre objectif de notre travail est d'enrichir le corpus Hansard avec des relations définies dans la base de connaissances DBpedia et qui ne sont pas explicites dans le corpus.

En effet, un des objectifs premiers de l'extraction de ces annotations est de pouvoir relier le corpus des débats parlementaires à DBpedia. Comme indiqué dans la section 3.3.1, certaines annotations retournées par AlchemyAPI contiennent des liens vers des entités DBpedia. C'est le cas pour les annotations de type concept et pour certaines des annotations de type entité. Remarquons que, pour l'extraction des relations issues de DBpedia, nous utilisons seulement les URI fournies par l'annotateur sémantique. De cette manière, on évite le processus de désambiguïstation auquel on aurait dû recourir, si nous avions extrait les relations de DBpedia en considérant les formes de surfaces des annotations.

Reconsidérons l'intervention 2:

Mr. Speaker, we work with the municipalities more than the **NDP** does. The **New Democrats** are making **promises**, but we are keeping promises. I would like to remind my **colleague** that [96 % of **Canada's** infrastructure] belongs to the **municipalities** and **provinces**. We have been their **partner** since 2006. No other **government** has invested as much as ours in **infrastructure**, while respecting

jurisdictions. The **NDP** wants to centralize **decision making** in **Toronto**. We are working with our **partners** while respecting their **jurisdictions**.

On a, parmi les annotations, les concepts **Toronto** et **Canada**. Essayons d'extraire les relations DBpedia entre ces deux concepts.

L'extraction des relations du LOD est effectuée à travers des requêtes SPARQL. AlchemyAPI nous retourne les URI de ces concepts qui sont les suivantes :

`<http://dbpedia.org/resource/Canada>` et `<http://dbpedia.org/resource/Toronto>`.

Pour récupérer les relations entre ces concepts dans DBpedia, on exécute la requête SPARQL suivante :

```
Prefix rdfs: http://www.w3.org/2000/01/rdf-schema#
SELECT ?relation
WHERE {
    <http://dbpedia.org/resource/Canada> ?relation
    <http://dbpedia.org/resource/Toronto>
}
```

Cette requête a pour but de chercher et d'extraire toutes les relations existantes dans DBpedia de type *Canada **relation** Toronto*.

On remarque qu'on voudrait aussi extraire les relations de type *Toronto **relation** Canada*. Il faudra pour cela, effectuer une autre requête SPARQL dans l'autre direction. En général, on cherche toutes les relations de type $t_1 \text{ } x \text{ } t_2$, et $t_2 \text{ } y \text{ } t_1$.

En effectuant ce processus d'extraction de relations sur les concepts Canada et Toronto, on obtient les résultats suivants :

- *Canada « largest city » Toronto*
- *Toronto « country » Canada*
- *Toronto « see also » Canada*

Le premier résultat indique que *Toronto* est la ville la plus grande du Canada. La deuxième indique que Toronto fait partie du *Canada*. Et la dernière indique que *Toronto* est en quelque sorte relié à Canada. Ces informations n'existent pas dans le corpus textuel, et ne sont pas non plus retournées

par AlchemyAPI. Avec ces informations, on enrichit aussi bien le corpus initial de débats parlementaires que notre graphe résumé.

3.5 Génération de graphes résumés

Ici, nous nous trouvons dans la dernière phase de notre système. Le but est de générer un graphe qui résume le contenu le plus important d'un corpus de débats. Dans les étapes précédentes, nous avons extrait des annotations agrégées, auxquelles nous avons associé des scores. Nous avons aussi extrait des relations (de cooccurrence, des relations verbales et des relations provenant du LOD) entre ces annotations. Nous générons finalement un graphe représentant tout le corpus où les sommets représentent les annotations, et les arêtes les relations entre celles-ci. On constate que ce graphe abstrait représente tout le corpus analysé sans tenir compte de l'importance des éléments du graphe. Pour la production d'un graphe résumé, nous réduisons successivement ce graphe initial pour ne garder seulement les éléments les plus pertinents en fonction de plusieurs paramètres. Rappelons que les sommets du graphe sont des annotations agrégées (regroupement d'annotations sémantiques par lemme) et les arêtes sont des relations entre ces annotations. Comme indiqué dans la section 3.3, nous avons associé trois métriques aux annotations et nous avons calculé un réseau de cooccurrence entre elles. Les scores associés aux annotations sont la fréquence d'occurrence, le score TF-IDF et le score de pertinence fourni par AlchemyAPI. Nous avons aussi calculé deux types de réseau de cooccurrence : la cooccurrence par interventions et la cooccurrence par phrases. Ces deux types de cooccurrence nous permettent de générer deux types de graphes que nous avons appelés *graphes non labélisés* et *graphes labélisés*. Nous réduisons successivement le graphe abstrait en générant des sous-graphes qui considèrent seulement les sommets ayant les plus grands scores (TF, pertinence, TFIDF) et des arêtes avec les plus grandes fréquences de cooccurrence. La figure 3.4 représente un tel graphe généré à partir du débat du 1er mai 2015, en considérant seulement les 10 annotations avec les fréquences d'occurrence les plus élevées.

précédant. Par exemple la relation *[non-resident voter] receive [ballot]* nous indique que le débat traite du fait que des bulletins de vote soient aussi assignés aux non-résidents.

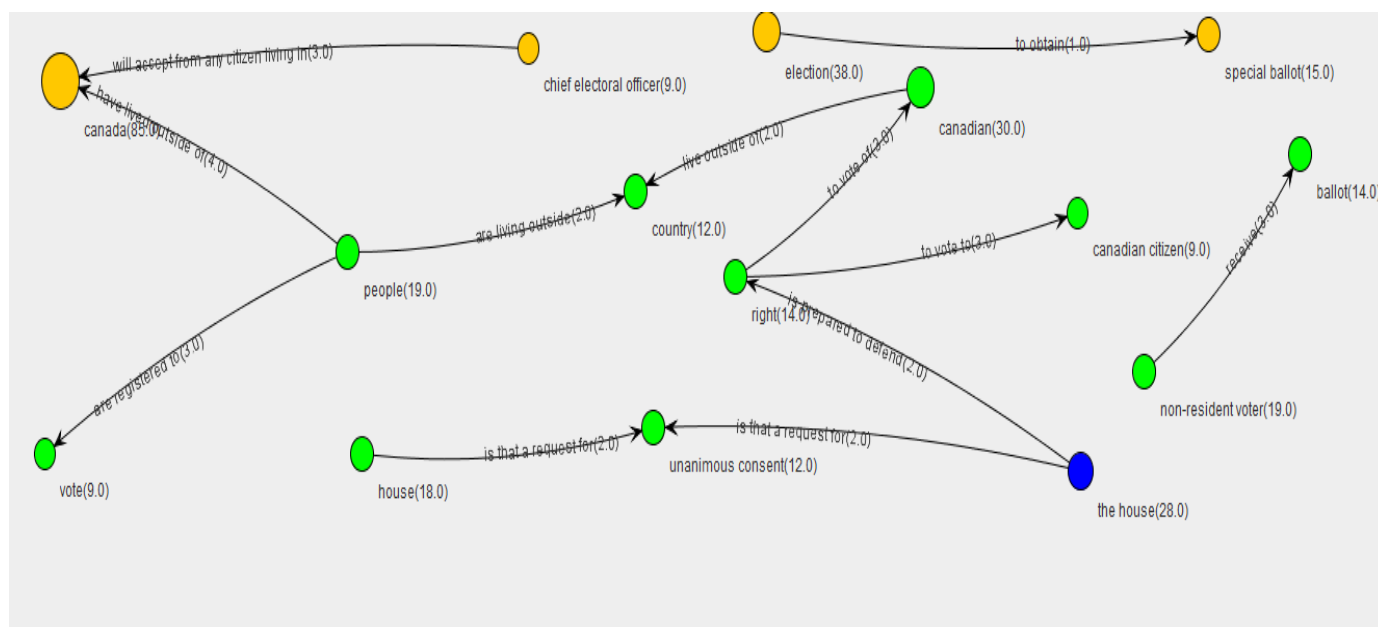


Figure 3.4. Graphe labélisé relatif au sujet de travail *Citizen Voting Act* pour le débat du 4 mai 2015, généré avec les 15 termes les plus fréquents

Le graphe abstrait que nous générons est exporté en format RDF. Cela permet de le publier sur le LOD et de le rendre utilisable par les applications sémantiques. La figure 3.5 présente un échantillon du graphe de la figure 3.4 généré au format RDF. Dans ce graphe, on peut voir l'annotation *right*, ainsi que tous les scores calculés pour cette annotation. On peut aussi voir la relation *to obtain* qui relie les annotations *election* et *special ballot*. Les détails sur l'ontologie utilisée pour l'exportation sont présents à l'annexe D.

```

[ a                "Link" ;
  <clearWSD>       "" ;
  <destination>    "special ballot" ;
  <relation>       "to obtain" ;
  <score>          "1.0" ;
  <source>         "election" ;
  <verbnet>        "[obtain-13.5.2]"
] .

[ a                "Term" ;
  <alchemy type>   "keyword" ;
  <relevanceScore> "0.373791" ;
  <surface>        "right" ;
  <tfScore>        "14" ;
  <tfidfScore>     "97.50473024594426"
] .

```

Figure 3.5. Portion d'un graphe labellisé en format RDF

CHAPITRE 4 EXPÉRIMENTATION ET ANALYSE

Dans ce chapitre, nous évaluons notre système. Nous avons opté pour une évaluation manuelle et extrinsèque, c'est-à-dire que l'évaluation est effectuée par des humains et le graphe résumé est évalué sur sa capacité à permettre la compréhension du corpus. Le but de cette évaluation est de vérifier que le graphe généré contient les entités et les relations les plus importantes du corpus. Dans cette évaluation nous examinons séparément chacune des parties principales de notre système. Nous évaluons premièrement les différentes métriques que nous avons associées aux annotations. Ces métriques sont analysées en regard de leur capacité à filtrer les annotations les plus importantes du corpus. Nous évaluons ensuite l'extension des annotations et aussi l'extraction des relations. Nous terminons par l'évaluation du graphe résumé. Nous avons construit des références (*gold standard* en anglais) en utilisant les évaluations des humains. Ces références sont utilisées pour mesurer la précision des modules de notre système. Dans les sections qui suivent, nous présentons d'abord le protocole d'évaluation utilisé ainsi que les caractéristiques du corpus d'évaluation. Ensuite, nous présentons pour les différentes parties de notre système, les résultats obtenus ainsi que l'analyse de ces résultats.

4.1 Construction du corpus de référence

Le corpus de références est construit en se basant sur les évaluations de six étudiants en maîtrise et au doctorat du laboratoire de Web sémantique et traitement de la langue naturelle de l'École Polytechnique de Montréal. Ces évaluateurs ont tous au moins un diplôme de premier cycle en génie informatique et effectuent des recherches dans le champ du traitement automatique de la langue naturelle. Ils sont donc familiers avec ce genre de tâche.

Notre corpus d'évaluation est constitué de quinze sujets de travail provenant du Hansard, que nous avons choisi de manière aléatoire. L'unique critère de sélection des sujets est le nombre d'interventions qu'ils contiennent. Nous n'avons sélectionné que des sujets dont le nombre d'interventions se situe entre quinze et vingt-cinq. Ce choix se justifie par le fait que nous voulions des sujets qui soient suffisamment longs pour être résumés et suffisamment courts pour pouvoir être analysés manuellement par un évaluateur humain. Sur chacun des sujets, nous avons extrait les annotations sémantiques, les relations verbales et nous avons construit le graphe labélisé correspondant. Les caractéristiques du corpus d'évaluation, ainsi que des statistiques sur les

annotations et les relations verbales extraites à partir de ce corpus sont présentées dans le tableau 4.1. Certains sujets de travail sont discutés sur plusieurs dates. C'est le cas de *taxation*, qui est discuté sur trois dates différentes. Dans le cadre de notre évaluation, on peut considérer ces sujets de travail comme indépendants.

Tableau 4.1. Caractéristiques du corpus d'évaluation : nombre d'interventions, nombre d'intervenants, nombre d'annotations provenant d'AlchemyAPI, nombre d'annotations étendues et nombre de relations verbales par sujet de travail. Tous les sujets proviennent de débats discutés en 2015.

Sujet de travail	Nombre d'interventions	Nombre d'intervenants	Nombre d'annotations d'AlchemyAPI	Nombre d'annotations étendues	Nombre de relations
Public Safety du 4 mai	15	10	321	96	49
Ombre Taxation du 4 mai	17	9	216	50	37
Ethics du 5 juin	16	6	204	47	30
Anti-terrorism Act du 2 mai	25	13	487	167	64
Taxation du 5 mai	18	11	225	58	25
Taxation du 6 mai	22	11	257	69	54
Finance du 7 décembre	16	5	205	38	23
National defense du 7 décembre	19	8	219	47	22
Taxation du 8 décembre	20	9	234	54	30
Infrastructure du 8 juin	22	11	364	102	45
Citizen Voting Act du 1 mai	22	10	188	417	84
Respecting Families of Murdered and Brutalized Person du 2 juin	15	8	200	311	79
Aboriginal Affairs du 3 juin	14	6	50	203	28

Tableau 4.1. Suite

Sujet de travail	Nombre d'interventions	Nombre d'intervenants	Nombre d'annotations d'AlchemyAPI	Nombre d'annotations étendues	Nombre de relations
The economy du 3 juin	17	8	66	270	31
National Defence du 4 mai	18	9	55	248	22
Moyenne	19	9.3	273.2	72.8	37.9

Nous avons demandé à chacun des évaluateurs d'évaluer cinq sujets différents. Nous avons donc obtenu pour chaque sujet, deux évaluations. Ceci nous a aussi permis de mesurer l'accord entre les évaluateurs. Une des métriques les plus utilisées pour calculer l'accord entre deux évaluateurs pour la classification d'éléments en deux classes disjointes est le *Kappa de Cohen* [73]. Des études récentes [86] ont démontré que la métrique *AC1 de Gwet* [87] a l'avantage d'être beaucoup moins sensible au problème de prédominance d'une classe par rapport à une autre, ce qui est le cas avec nos données, où le nombre d'items jugés corrects est en général bien plus élevé que le nombre d'items jugés incorrects. Par exemple au Tableau 4.2, on obtient des valeurs Kappa pratiquement nulles (0,06 et -0,08) pour *Anti-terrorism Act du 5 mai* et *Taxation du 8 décembre* alors que l'accord est nettement évident dans le deuxième cas. Nous avons donc utilisée la métrique AC1 pour mesurer l'accord entre les évaluateurs. La formule d'AC1, quand on a deux évaluateurs A et B et deux classes disjointes C_1 et C_2 est la suivante:

$$AC1 = (p_o - p_h) / (1 - p_h)$$

Où p_o est l'accord observé entre les évaluateurs. Il est calculé comme le pourcentage d'éléments classifiés dans la même catégorie par A et B.

p_h est la probabilité que l'accord soit dû au hasard. p_h est calculé de manière suivante :

$$p_h = 2q * (1 - q) \text{ avec } q = \frac{A1+B1}{2*N}.$$

A1 représente le nombre d'éléments classés dans la catégorie C1 par l'évaluateur A, B1 le nombre d'éléments classés dans la catégorie C1 par l'évaluateur B, et N le nombre total d'éléments à classer. L'interprétation des valeurs du AC1 est la même que celles du Kappa selon [87]. Un tableau indiquant l'interprétation de l'accord en fonction de la valeur Kappa est fournie à l'Annexe E.

Par la suite, nous détaillons les étapes nécessaires pour l'évaluation de chaque sujet. Les étapes principales sont les suivantes :

- Lecture du contenu textuel du sujet à évaluer
- Évaluation des scores de pertinence
- Évaluation des annotations étendues
- Évaluation des relations verbales et désambiguïsées avec VerbNet
- Évaluation qualitative du graphe résumé

Les détails de chacune des étapes de l'évaluation sont présentés dans les sections qui suivent.

4.2 Évaluation des scores associés aux annotations

Le but de cette section est d'identifier laquelle des métriques associées aux annotations est la plus discriminative et permet d'identifier les concepts les plus importants discutés dans le débat. Les métriques évaluées sont la fréquence d'occurrence, le score de pertinence d'AlchemyAPI et le TF-IDF. Pour chacun des sujets de travail du corpus d'évaluation, nous avons choisi les cinquante annotations qui ont les plus grandes valeurs pour la fréquence d'occurrence, le score de pertinence et le TF-IDF. À noter que si une annotation apparaît dans le top-50 des annotations avec les plus grands scores de pertinence et aussi dans le top-50 des annotations avec les plus grands scores de TF-IDF, elle ne sera considérée qu'une seule fois. Après la sélection des annotations, nous les présentons dans un ordre aléatoire aux évaluateurs et nous leur demandons d'indiquer pour chacune d'elles si elles sont pertinentes ou pas. Nous obtenons ainsi une valeur de précision, c'est-à-dire le nombre d'annotations correctes parmi celles retenues par notre système. Un outil avec une interface graphique est fourni pour faciliter la tâche d'évaluation. Les évaluateurs ont seulement à cocher ou décocher une case pour indiquer si l'annotation est pertinente ou non. À la fin du processus d'évaluation, nous disposons, pour chacun des évaluateurs, d'un ensemble d'annotations qui

constituent la référence. Cet ensemble sera appelé A_i^r où $i \in [1,2]$ représente le i ème évaluateur du sujet. Dans le Tableau 4.2 nous présentons, pour chaque sujet, la précision des annotations selon chaque évaluateur. La colonne *Nombre d'annotations* contient pour chaque sujet, le nombre d'annotations présentées aux évaluateurs. La colonne *Nombre d'annotations en commun* contient pour chaque sujet le nombre d'annotations pour lesquelles les évaluateurs donnent la même évaluation

Tableau 4.2. Tableau indiquant la précision des annotations sélectionnées dans les références et l'accord entre les évaluateurs.

Sujet de travail	Précision évaluateur 1	Précision évaluateur 2	Nombre d'annotations	Nombre d'annotations en commun	AC1
Public Safety du 4 mai	60	73	117	90	0.58
Taxation du 4 mai	59	81	116	85	0.54
Ethics du 5 juin	44	83	107	60	0.18
Anti-terrorism Act du 5 mai	41	95	120	55	0.05
Taxation du 5 mai	32	93	113	44	-0.14
Taxation du 6 mai	82	76	113	86	0.64
Finance du 7 décembre	87	82	103	80	0.7
National defense du 7 décembre	83	88	110	89	0.75
Taxation du 8 décembre	85	95	102	82	0.76
Infrastructure du 8 juin	79	91	120	93	0.7

Tableau 4.2. Suite

Sujet de travail	Précision évaluateur 1	Précision évaluateur 2	Nombre d'annotations	Nombre d'annotations en commun	AC1
Citizen Voting Act du 1 mai	44	46	130	111	0.7
Respecting Families of Murdered and Brutalized Person du 2 juin	41	56	123	90	0.46
Aboriginal Affairs du 3 juin	37	53	114	76	0.34
The economy du 3 juin	62	51	115	82	0.44
National Defence du 4 mai	50	31	109	74	0.38
Moyenne	66		114.1	79.8	0.47

Nous obtenons une moyenne de 66% sur le nombre d'annotations sélectionnées par les évaluateurs. En moyenne, les évaluateurs ont trouvé que les annotations qui leur ont été présentées sont généralement pertinentes (66%). L'accord moyen obtenu est de 0.47. Selon le tableau d'interprétation fournie à l'Annexe E, l'accord est modéré ([0.41-0.60]) entre les évaluateurs. Toutefois, sur certains sujets, l'accord entre les évaluateurs est très faible. Si on exclut les trois sujets qui ont les valeurs de AC1 les plus basses, la moyenne de l'accord monte à 0.58, qui rentre presque dans la catégorie d'accord fort ([0.61-0.80]). En excluant ces sujets, la précision baisse légèrement à 65,75%. Les références extraites sont utilisées dans la prochaine section pour évaluer les différentes métriques.

4.2.1 Métriques utilisées

Pour l'évaluation des annotations, nous utilisons la précision qui, rappelons-le, représente la proportion d'items corrects parmi ceux qui sont évalués. Celle-ci est utilisée pour évaluer la sélection des annotations selon chacune des métriques que nous avons implémentées, soit la fréquence d'occurrence, le TF-IDF et le score de pertinence d'AlchemyAPI. La précision est calculée en fonction de la référence, expliquée plus tôt, c'est-à-dire l'ensemble des annotations considérées correctes par un évaluateur (donc nous aurons deux références pour chaque sujet, puisque chacun a été analysé par deux évaluateurs). Par exemple, la précision du TF est calculée comme le nombre d'annotations appartenant au top-50 des annotations ayant les fréquences d'occurrence la plus élevée qui se trouvent dans la référence. Plus formellement soit $A^{TF-50} = \{a_1^{TF}, a_2^{TF}, \dots, a_{50}^{TF}\}$ l'ensemble ordonné des 50 annotations avec les plus grandes fréquences d'occurrence. La précision obtenue en utilisant la fréquence d'occurrence, et en considérant le top-50 des annotations, que nous désignerons P_{TF50} est calculée comme suit :

$$P_{TF50} = \frac{|\{a \in A^{TF-50} : a \in A^r\}|}{50} \times 100$$
 où $|\{a \in A^{TF-50} : a \in A^r\}|$ représente le nombre d'annotations à la fois dans le top-50 du TF et dans A^r . La précision pour l'utilisation des deux autres métriques est calculée de manière similaire, soit :

$$P_{TF-IDF50} = \frac{|\{a \in A^{TFIDF-50} : a \in A^r\}|}{50} \times 100$$

$$P_{Alchemy50} = \frac{|\{a \in A^{Alchemy-50} : a \in A^r\}|}{50} \times 100$$

Pour chaque métrique, on considère le top-50, mais aussi le top-10. Nous aurons donc aussi les valeurs de précision P_{TF10} , $P_{TF-IDF10}$ et $P_{Alchemy10}$.

4.2.2 Résultats et analyse

Comme indiqué dans la section 4.1, chacun des sujets du corpus d'évaluation est évalué par deux évaluateurs humains. Nous avons donc deux ensembles de référence A_1^r et A_2^r et donc deux valeurs de précision calculées. Dans les paragraphes qui suivent, nous présentons les précisions obtenues avec chacun des ensembles de références. Dans les tableaux 4.3 et 4.4, nous présentons pour chacun des sujets de travail, les précisions P_{TF50} et P_{TF10} . Dans les tableaux 4.5 et 4.6, nous présentons pour chacun des sujets de travail, les précisions $P_{TF-IDF50}$ et $P_{TF-IDF10}$. Et finalement dans les

tableaux 4.7 et 4.8, nous présentons pour chacun des sujets de travail, les précisions $P_{Alchemy50}$ et $P_{Alchemy10}$. Ensuite, nous présentons une discussion comparative sur les résultats obtenus.

Précision du TF

Tableau 4.3. Tableau indiquant les précisions pour le top-50 calculées avec les références A_1^r et A_2^r

Sujets	P_{TF50} calculée avec A_1^r	P_{TF50} calculée avec A_2^r	Moyenne par sujet
Public Safety du 4 mai	68	78	73
Taxation du 4 mai	68	92	80
Ethics du 5 juin	46	88	67
Anti-terrorism Act du 5 mai	54	98	76
Taxation du 5 mai	46	96	71
Taxation du 6 mai	80	80	80
Finance du 7 décembre	84	82	83
National defense du 7 décembre	84	88	86
Taxation du 8 décembre	90	96	93
Infrastructure du 8 juin	88	96	92
Citizen Voting Act du 1 mai	50	58	54
Respecting Families of Murdered and Brutalized Person du 2 juin	52	62	57
Aboriginal Affairs du 3 juin	58	68	63

The economy du 3 juin	68	64	66
-----------------------	----	----	----

Tableau 4.3. Suite

Sujets	P_{TF50} calculée avec A_1^r	P_{TF50} calculée avec A_2^r	Moyenne par sujet
National Defence du 4 mai	62	40	51
Moyenne de la précisions du TF_{50}	-	-	72.8

Tableau 4.4. Tableau indiquant les précisions pour le top-10 calculées avec les références A_1^r et A_2^r

Sujets	P_{TF50} calculée avec A_1^r	P_{TF50} calculée avec A_2^r	Moyenne par sujet
Public Safety du 4 mai	70	90	80
Taxation du 4 mai	70	100	85
Ethics du 5 juin	60	80	70
Anti-terrorism Act du 5 mai	70	100	85
Taxation du 5 mai	30	100	65
Taxation du 6 mai	100	80	90
Finance du 7 décembre	80	70	75
National defense du 7 décembre	90	100	95
Taxation du 8 décembre	80	100	90
Infrastructure du 8 juin	80	100	90

Citizen Voting Act du 1 mai	60	80	70
-----------------------------	----	----	----

Tableau 4.4. Suite

Sujets	P_{TF50} calculée avec A_1^r	P_{TF50} calculée avec A_2^r	Moyenne par sujet
Respecting Families of Murdered and Brutalized Person du 2 juin	60	60	60
Aboriginal Affairs du 3 juin	50	80	65
The economy du 3 juin	60	70	65
National Defence du 4 mai	72	50	61
Moyenne	-	-	76.4

Précision du TF-IDF

Tableau 4.5. Tableau indiquant les précisions pour le top-50 calculées avec les références A_1^r et A_2^r

Sujets	$P_{TF-IDF50}$ calculée avec A_1^r	$P_{TF-IDF50}$ calculée avec A_2^r	Moyenne par sujet
Public Safety du 4 mai	66	74	70
Taxation du 4 mai	58	84	71
Ethics du 5 juin	40	90	65
Anti-terrorism Act du 5 mai	46	92	69

Taxation du 5 mai	38	94	66
Taxation du 6 mai	78	82	80
Finance du 7 décembre	92	96	94

Tableau 4.5. Suite

Sujets	$P_{TF-IDF50}$ calculée avec A_1^r	$P_{TF-IDF50}$ calculée avec A_2^r	Moyenne par sujet
National defense du 7 décembre	86	92	89
Taxation du 8 décembre	82	100	91
Infrastructure du 8 juin	56	88	72
Citizen Voting Act du 1 mai	64	68	66
Respecting Families of Murdered and Brutalized Person du 2 juin	54	74	64
Aboriginal Affairs du 3 juin	50	68	59
The economy du 3 juin	60	68	64
National Defence du 4 mai	64	34	49
Moyenne	-	-	71.2

Tableau 4.6. Tableau indiquant les précisions pour le top-10 calculées avec les références A_1^r et A_2^r

Sujets	$P_{TF-IDF10}$ calculée avec A_1^r	$P_{TF-IDF10}$ calculée avec A_2^r	Moyenne par sujet
--------	--------------------------------------	--------------------------------------	-------------------

Public Safety du 4 mai	50	50	50
Taxation du 4 mai	100	40	70
Ethics du 5 juin	100	40	70
Anti-terrorism Act du 5 mai	90	40	65

Tableau 4.6. Suite

Sujets	$P_{TF-IDF10}$ calculée avec A_1^r	$P_{TF-IDF10}$ calculée avec A_2^r	Moyenne par sujet
Taxation du 5 mai	90	30	60
Taxation du 6 mai	70	80	75
Finance du 7 décembre	100	100	100
National defense du 7 décembre	80	100	90
Taxation du 8 décembre	70	100	85
Infrastructure du 8 juin	80	100	90
Citizen Voting Act du 1 mai	60	80	70
Respecting Families of Murdered and Brutalized Person du 2 juin	60	80	70
Aboriginal Affairs du 3 juin	40	80	60
The economy du 3 juin	60	60	60
National Defence du 4 mai	90	40	65

Moyenne	-	-	72
----------------	---	---	-----------

Précision du score d'AlchemyAPI

Tableau 4.7. Tableau indiquant les précisions pour le top-50 calculées avec les références

Sujets	P_{Alchemy50} calculée avec A_1^r	P_{Alchemy50} calculée avec A_2^r	Moyenne par sujet
Public Safety du 4 mai	56	72	64
Taxation du 4 mai	60	78	69
Ethics du 5 juin	52	82	67
Anti-terrorism Act du 5 mai	38	98	68
Taxation du 5 mai	24	94	59
Taxation du 6 mai	88	70	79
Finance du 7 décembre	90	84	87
National defense du 7 décembre	82	86	84
Taxation du 8 décembre	90	92	91
Infrastructure du 8 juin	96	90	93

Citizen Voting Act du 1 mai	32	30	31
Respecting Families of Murdered and Brutalized Person du 2 juin	38	44	41
Aboriginal Affairs du 3 juin	32	50	41
The economy du 3 juin	60	50	55
National Defence du 4 mai	42	30	36
Moyenne	-	-	64.33

Tableau 4.8. Tableau indiquant les précisions pour le top-10 calculées avec les références A_1^r et A_2^r

Sujets	$P_{\text{Alchemy10}}$ calculée avec A_1^r	$P_{\text{Alchemy10}}$ calculée avec A_2^r	Moyenne par sujet
Public Safety du 4 mai	70	90	80
Taxation du 4 mai	80	90	85
Ethics du 5 juin	40	90	65
Anti-terrorism Act du 5 mai	50	100	75
Taxation du 5 mai	20	90	55
Taxation du 6 mai	80	70	75
Finance du 7 décembre	100	80	90
National defense du 7 décembre	80	90	85
Taxation du 8 décembre	100	100	100

Infrastructure du 8 juin	100	90	95
Citizen Voting Act du 1 mai	30	20	25
Respecting Families of Murdered and Brutalized Person du 2 juin	30	40	35
Aboriginal Affairs du 3 juin	30	60	45
The economy du 3 juin	70	60	65
National Defence du 4 mai	60	20	40
Moyenne	-	-	67.66

Tableau 4.9. Récapitulatif des résultats obtenus

	Moyenne P_{TF}	Moyenne P_{TF-IDF}	Moyenne P_{Alchemy}
Top-50	72.8	71.2	64.33
Top-10	76.4	72	67.66

Discussion

Rappelons que le but de cette évaluation est de déterminer laquelle des métriques (TF, TF-IDF et le score de pertinence d'AlchemyAPI) est plus efficace pour l'extraction des annotations les plus importantes de notre corpus. La Tableau 4.9 présente les résultats obtenus pour chacune des métriques de précision. Pour le top-50, le TF se classe en première position avec une précision de 72.8, suivi du TF-IDF avec une précision de 71.1%. et du score de pertinence d'AlchemyAPI avec une précision de 64.33%. Pour le top-10, le TF reste en tête avec 76.4% de précision, toujours suivi du TF-IDF avec 72%. Le score de pertinence d'AlchemyAPI reste constant à la troisième place avec une précision de 67.66%. La fréquence d'occurrence semble donc la métrique la plus adéquate pour extraire les concepts les plus importants d'un sujet dans notre expérimentation. Une des raisons pourrait être que les sujets évalués sont courts. Lors de nos différents tests, nous avons

observé que lorsqu'on génère un graphe relatif à un débat au complet, certaines des annotations les plus fréquentes (par exemple *Mr.speaker, Canada, Government*) ne sont pas toujours pertinentes, car elles se retrouvent dans presque toutes les interventions (de tous les débats). Dans ces situations, le TF-IDF retournait de meilleurs résultats, car les annotations qui se retrouvaient dans les meilleures positions concordaient avec le sujet discuté. Nous nous sommes demandé si le fait que le TF-IDF ait, dans notre évaluation, des performances inférieures à la fréquence d'occurrence ne dépend pas du nombre d'interventions évaluées. L'hypothèse est que pour un nombre suffisamment élevé d'interventions, par exemple un débat au complet, le TF-IDF serait peut-être plus efficace que la fréquence d'occurrence. Malheureusement, les débats sont longs (environ 60000 mots en moyenne) et les faire évaluer manuellement sur toute une date de débat serait très coûteux. Nous n'avons donc pas pu vérifier cette hypothèse. Pour le score de pertinence, rappelons que ce score est retourné par AlchemyAPI pour chaque intervention où une annotation apparaît. Une annotation qui apparaît dans plusieurs interventions aura donc plusieurs scores de pertinence différents. On calcule le score de pertinence globale d'une annotation en faisant la moyenne des scores de pertinence retournés dans chaque intervention où apparaît l'annotation. Ceci dit, considérons le cas où une annotation x apparaît dans une seule intervention, et son score de pertinence est élevé pour cette intervention. Supposons aussi qu'on ait une deuxième annotation y qui apparaît dans plusieurs interventions, mais dont les scores de pertinence sont légèrement inférieurs à celui de x . En faisant la moyenne, x aurait un score de pertinence supérieur à y même s'il est fort probable qu' y soit le plus pertinent, puisque y apparaît dans plusieurs interventions. D'après les résultats de notre évaluation, en faisant seulement la moyenne, on obtient des résultats satisfaisants (précision supérieure à 60%). Nous pensons que cette approche (moyenne des scores de pertinence) pourrait avoir des résultats moins bons si appliqués à un corpus de grande taille. Malheureusement, cette hypothèse n'a pas pu être validée, car il est difficile d'évaluer toute une date de débat. En définitive, on peut affirmer que quand l'analyse se limite à des sujets de travail du Hansard, la fréquence d'occurrence permet d'obtenir de meilleurs résultats si les sujets sont d'une taille relativement petite. Dans les sections suivantes, nous présentons les évaluations de l'extension des annotations et des relations verbales.

4.3 Évaluation des annotations étendues

Le but de cette phase est d'évaluer le processus d'extension d'annotation. On choisit de manière aléatoire cinquante annotations étendues pour chaque sujet évalué et on demande aux évaluateurs d'indiquer si l'annotation étendue produite est sémantiquement correcte. L'évaluateur dispose pour chaque annotation étendue à évaluer de sa forme originale (non étendue) et de la phrase d'où l'annotation provient. Certains sujets ont moins de cinquante annotations étendues. Pour ceux-ci, nous présentons aux évaluateurs toutes les annotations étendues du sujet.

Pour l'évaluation du processus d'extension d'annotation, nous avons également utilisé la *précision* qui représente le pourcentage d'annotations étendues jugées correctes par les évaluateurs. Chacun des sujets est évalué là encore par deux évaluateurs humains. Nous avons donc deux valeurs de précision P_1 et P_2 qui représentent les précisions calculées à partir des évaluations du premier et du deuxième évaluateur humain, respectivement. La Tableau 4.10 présente les résultats obtenus pour l'évaluation du processus d'extension d'annotation. La colonne *Nombre d'annotations évaluées* contient pour chaque sujet, le nombre d'annotations présentées aux évaluateurs. La colonne *Nombre d'annotations en commun* contient pour chaque sujet le nombre d'annotations pour lesquelles les évaluateurs donnent la même évaluation

Tableau 4.10. Tableau indiquant pour chaque sujet de travail la précision d'extension₁, la précision d'extension₂ et l'accord entre les annotateurs.

Sujets	P ₁	P ₂	Moyenne par sujet	Nombre d'annotations étendues évaluées	Nombre d'annotations en commun	AC1
Public Safety du 4 mai	94	98	96	50	48	0.95
Taxation du 4 mai	80	97	88,5	47	37	0.73
Ethics du 5 juin	78	90	84	41	36	0.83
Anti-terrorism Act du 5 mai	80	94	87	50	43	0.81
Taxation du 5 mai	62	85	73.5	48	37	0.62

Taxation du 6 mai	80	62	71	50	39	0.62
Finance du 7 décembre	72	48	60	25	19	0.54
National defense du 7 décembre	70	62	66	40	37	0.86
Taxation du 8 décembre	96	86	91	50	41	0.78
Infrastructure du 8 juin	92	88	90	50	46	0.90
Citizen Voting Act du 1 mai	98	94	96	50	46	0.91
Respecting Families of Murdered and Brutalized Person du 2 juin	100	98	99	50	49	0.97

Tableau 4.12. Suite

Sujets	P ₁	P ₂	Moyenne par sujet	Nombre d'annotations étendues évaluées	Nombre d'annotations en commun	AC1
Aboriginal Affairs du 3 juin	87	96	91.5	33	30	0.89
The economy du 3 juin	74	94	84	50	40	0.72
National Defence du 4 mai	77	80	78.5	45	42	0.9
Moyenne	-	-	83.7	45.1	38.3	0.8

Pour le processus d'extension d'annotation, nous obtenons de bons résultats avec un taux de précision de 83.7%. L'accord des évaluateurs est de 0.8, ce qui est presque parfait. La plupart des erreurs que nous rencontrons dans cette section sont liées aux limites de l'approche que nous

utilisons. Nous utilisons un simple étiquetage morphosyntaxique pour effectuer l'extension d'annotations. Certaines situations d'erreurs sont indétectables avec cette approche, et il faudrait utiliser des outils plus complexes comme des arbres de dépendance syntaxiques. Des exemples d'erreurs sont présentés dans le Tableau 4.11.

Tableau 4.11. Quelques erreurs dans le processus d'extension d'annotations

annotation originale	annotation étendue	phrase de provenance
inspiration	inspiration people	We all know the attacks that have occurred around the world are unusual in that they are not part of an organized and systematic attempt by ISIL to do things, but rather the inspiration people receive through their radicalization
long run	long run Canada	Then there was the third essential element, and perhaps in the long run Canada 's most important contribution, our commitment to the international coalition 's military mission.

Tableau 4.11. Suite

annotation originale	annotation étendue	phrase de provenance
basis	basis of the commentary today	Mr. Speaker, I would neither argue that I purport to understand this particular area nor would I, from the basis of the commentary today , feel that we really do understand this area
military	Russian military land	Instead of bombing Daesh, the vast majority of bombs unleashed by the Russian military land on anti-Assad forces and civilian neighbourhoods

Observons la première annotation étendue. On a l'annotation *inspiration* qui est étendue à *inspiration people*. Le patron utilisé ici est le patron P1 du Tableau 3.2. Lorsqu'on a une succession de noms, ce patron permet de les rassembler en un seul nom composé. On voit bien ici que nous

avons une exception à cette règle. Dans ce cas, on a une forme syntaxique particulière, où *inspiration* et *people* sont l’objet direct et le sujet de *receive*, respectivement. En ne considérant que les catégories grammaticales, il n’est pas possible d’éviter ce genre d’erreurs. Globalement, les résultats obtenus sont quand même satisfaisants. La situation est similaire pour la deuxième annotation, où *long run* et *Canada* appartiennent à des syntagmes différents dans l’analyse syntaxique. Dans la troisième annotation, *today* ne devrait pas être ajouté, puisqu’il joue ici le rôle d’un adverbe et ne devrait donc pas se retrouver dans un syntagme nominal. Finalement, dans la dernière annotation, *land* a été interprété erronément comme un nom, alors qu’il s’agit d’un verbe.

4.4 Évaluation des relations verbales

Dans cette section, nous présentons l’évaluation de notre processus d’extraction de relations. Nous avons choisi de manière aléatoire cinquante relations verbales, pour chacun des sujets de travail de notre corpus d’évaluation que nous avons présentés aux évaluateurs. Nous leur avons ensuite demandé si les relations ont été correctement extraites. Nous avons aussi évalué si les relations extraites sont en général pertinentes et reflètent le contenu du sujet. Les relations verbales ont été désambiguïsées en fonction des classes d’équivalence correspondantes dans la base de connaissances VerbNet. Nous avons donc évalué, pour chaque relation verbale extraite, si parmi les classes candidates il en existe au moins une qui abstrait correctement la relation verbale. Finalement, nous avons évalué si la classe sélectionnée est correcte.

Pour évaluer tous ces aspects, nous avons demandé à chaque évaluateur de répondre par oui ou par non aux questions suivantes, pour chacune relation verbale qui lui est présentée :

- a) La relation est-elle correctement extraite ?
- b) La relation est-elle pertinente ?
- c) La classe VerbNet correspondante est-elle présente ?
- d) La relation est-elle bien désambiguïsée ?

4.4.1 Métriques utilisées

Pour l’évaluation des relations verbales, nous avons à nouveau utilisé la métrique de précision. La *précision d’extraction de relations* permet d’évaluer la capacité de notre technique d’extraction de relation à extraire des relations verbales correctes de la phrase source. Elle représente le

pourcentage de relations considérées correctes par l'évaluateur parmi celles qui lui sont présentées. La *pertinence des relations* indique à quel point les relations extraites permettent la compréhension du débat. Il s'agit du pourcentage de relations verbales jugées pertinentes par l'évaluateur parmi celles qui lui sont présentées. La *précision VerbNet* indique le pourcentage de relations verbales que nous sommes en mesure de désambigüiser. Elle représente le pourcentage de relations verbales pour lesquelles l'évaluateur juge qu'il existe au moins une classe VerbNet correspondante à la relation. La *précision de la désambigüisation* indique la précision du système de désambigüisation mis en place. Elle représente le pourcentage de relations verbales pour lesquelles l'évaluateur juge que la classe sélectionnée par notre système est correcte, dans les cas où il a indiqué qu'il existe au moins une classe parmi la liste des candidates. Pour chacun des sujets, chacune de ces métriques est calculée séparément pour les deux évaluateurs.

4.4.2 Résultats et analyse

Dans cette section, nous présentons les résultats d'évaluation des relations verbales. Nous présentons séparément les résultats et les analyses de chacune des sous-sections, à savoir l'extraction des relations, la pertinence des relations, la précision VerbNet et la désambigüisation des relations.

Précision d'extraction de relations

Dans le tableau 4.12, nous présentons, pour chaque sujet et pour chacun des évaluateurs, la précision du processus d'extraction de relations. La colonne *Nombre de relations* contient, pour chaque sujet, le nombre de relations présentées aux évaluateurs. La colonne *Nombre de relations en commun* contient pour chaque sujet le nombre de relations pour lesquelles les évaluateurs donnent la même évaluation.

Tableau 4.12. Tableau indiquant les précisions d'extractions de relations verbales.

Sujet	Evaluateur ₁	Evaluateur ₂	Moyenne	Nombre de relations	Nombre de relations en commun	AC1
Public Safety du 4 mai	78	50	64	42	26	0.29
Taxation du 4 mai	100	75	87.5	29	22	0.69

Ethics du 5 juin	100	59	79.5	22	13	0.39
Anti-terrorism Act du 5 mai	88	82	85	45	36	0.73
Taxation du 5 mai	85	66	75.5	21	17	0.70
Taxation du 6 mai	70	67	68.5	34	27	0.64
Finance du 7 décembre	80	60	70	15	10	0.42
National defense du 7 décembre	90	85	87.5	21	16	0.70
Taxation du 8 décembre	100	82	91	28	23	0.78
Infrastructure du 8 juin	73	60	66.5	23	14	0.30

Tableau 4.12. Suite

Sujet	Evaluateur ₁	Evaluateur ₂	Moyenne	Nombre de relations	Nombre de relations en commun	AC1
Citizen Voting Act du 1 mai	68	74	71	50	45	0.83
Respecting Families of Murdered and Brutalized Person du 2 juin	80	82	81	50	41	0.74
Aboriginal Affairs du 3 juin	94	88	91	17	16	0.93
The economy du 3 juin	90	80	85	20	18	0.86

National Defence du 4 mai	60	70	65	20	18	0.81
Moyenne	-	-	77.9	28	20.4	0.65

Dans 78% des cas, on réussit à extraire des relations bien formées de type *sujet relation objet* en utilisant les patrons que nous avons définis dans la section 3.3.1. La plupart des erreurs que nous rencontrons sont dues aux limites des outils utilisés, avec lesquels certaines erreurs sont indétectables. Des exemples d'erreur sont présentés au Tableau 4.13. Si on observe la deuxième ligne du tableau, on peut voir qu'il y a erreur sur le sujet de la relation verbale. La relation extraite est *[hon. Colleague]_{ANN} [has to do with]_{REL} [important point]_{ANN}*. En observant la phrase de provenance, on peut noter que le sujet de la relation serait plutôt *my question*. En général, les patrons définis dans la section 3.3.1 considèrent que le dernier syntagme nominal avant la relation est le sujet de la relation. Quand ce n'est pas le cas, la relation extraite est erronée.

Tableau 4.13. Tableau présentant quelques erreurs rencontrées dans le processus d'extraction de relations

annotation	relation	annotation	phrase de provenance
hon. colleague	has to do with	important point	My question to my hon. colleague has to do with the important point she raised about radicalization
local power	try to combat	group	We are working with local powers to try to combat this group.
time	to continue	speech	I am sure that the member will have time to continue his speech in the questions and comments.
need	has in	fact	There are plenty of cases where a western power coming in to help those in need has in fact succeeded in bringing about genuine transformation

Pertinence des relations

Dans le Tableau 4.14, nous présentons, pour chaque sujet et pour chacun des évaluateurs, la *pertinence des relations*. Rappelons que cette métrique indique à quel point les relations extraites permettent la compréhension du débat. La colonne *Nombre de relations* contient pour chaque sujet, le nombre de relations présentées aux évaluateurs. La colonne *Nombre de relations en commun* contient pour chaque sujet le nombre de relations pour lesquelles les évaluateurs donnent la même évaluation.

Tableau 4.14. Tableau indiquant la précision de pertinence des relations verbales.

Sujet	Evaluateur ₁	Evaluateur ₂	Moyenne	Nombre de relations	Nombre de relations en commun	AC1
Public Safety du 4 mai	52	80	66	42	24	0.23
Taxation du 4 mai	58	93	75.5	29	19	0.46
Ethics du 5 juin	59	81	70	22	13	0.30
Anti-terrorism Act du 5 mai	60	93	76.5	45	30	0.48
Taxation du 5 mai	61	90	75.5	21	13	0.40
Taxation du 6 mai	73	76	149	34	27	0.67
Finance du 7 décembre	80	93	86.5	15	13	0.83
National defense du 7 décembre	95	61	78	21	14	0.50

Taxation du 8 décembre	89	64	76.5	28	19	0.50
Infrastructure du 8 juin	73	86	79.5	23	20	0.80
Citizen Voting Act du 1 mai	52	72	62	50	32	0.32
Respecting Families of Murdered and Brutalized Person du 2 juin	60	78	69	50	39	0.61

Tableau 4.14. Suite

Sujet	Evaluateur ₁	Evaluateur ₂	Moyenne	Nombre de relations	Nombre de relations en commun	AC1
Aboriginal Affairs du 3 juin	70	82	76	17	11	0.45
The economy du 3 juin	60	75	67	20	15	0.55
National Defence du 4 mai	30	65	47	20	11	0.10
Moyenne	-	-	72	28	19.2	0.48

Pour ce qui en est de la pertinence des relations extraites, on obtient un taux de précision de 72% en moyenne. Les relations présentées aux évaluateurs ont été sélectionnées de manière aléatoire, sans aucun critère de préférence. L'accord entre les annotations est 0.48, ce qui est modéré selon [73].

Précision VerbNet

Dans le tableau 4.15, nous présentons, pour chaque sujet et pour chacun des évaluateurs, la *précision VerbNet*, qui indique le pourcentage de relations verbales que nous sommes en mesure de désambiguïser. Dans la section 3.3.2, nous avons présenté notre processus de désambiguïsation de nos relations verbales en utilisant la base de connaissance VerbNet. Nous avons indiqué qu’une relation verbale peut avoir plusieurs classes VerbNet candidates. Avec la métrique *précision VerbNet*, on évalue si en général parmi les classes VerbNet candidates, on a au moins une classe VerbNet qui correspond à la relation. La colonne *Nombre de relations* contient, pour chaque sujet, le nombre de relations présentées aux évaluateurs. La colonne *Nombre de relations en commun* contient pour chaque sujet le nombre de relations pour lesquelles les évaluateurs donnent la même évaluation.

Tableau 4.15. Tableau indiquant la précision VerbNet calculée pour chacun des évaluateurs

Sujet	Précision VerbNet ₁	Précision VerbNet ₂	Moyenne par sujet	Nombre de relations	Nombre de relations en commun	AC1

Public Safety du 4 mai	69	66	67.5	42	39	0.87
Taxation du 4 mai	93	86	89.5	29	27	0.91
Ethics du 5 juin	72	63	67.5	22	20	0.84
Anti-terrorism Act du 5 mai	73	80	76.5	45	38	0.76
Taxation du 5 mai	66	90	78	21	16	0.64
Taxation du 6 mai	61	76	68.5	34	29	0.74
Finance du 7 décembre	60	70	65	15	10	0.38
National defense du 7 décembre	28	52	40	21	12	0.17

Tableau 4.15. Suite

Sujet	Précision VerbNet ₁	Précision VerbNet ₂	Moyenne par sujet	Nombre de relations	Nombre de relations en commun	AC1
Taxation du 8 décembre	50	64	57	28	22	0.58
Infrastructure du 8 juin	21	26	23.5	23	22	0.93
Citizen Voting Act du 1 mai	44	36	40	50	44	0.77
Respecting Families of Murdered and Brutalized Person du 2 juin	56	50	53	50	41	0.64

Aboriginal Affairs du 3 juin	35	29	32	17	16	0.89
The economy du 3 juin	50	40	45	20	14	0.40
National Defence du 4 mai	45	40	42	20	15	0.511
Moyenne	-	-	56.4	28	23.5	0.67

Nous obtenons une *précision VerbNet* moyenne d'environ 56%. Cela signifie que 56% des relations extraites ont une classe VerbNet candidate qui correspond au sens du verbe de la relation. Les autres relations qui n'ont aucune classe VerbNet correspondante ne pourront donc pas être désambiguïsées. Nous avons indiqué dans la revue de littérature qu'une des limitations des systèmes d'extraction de relations non contraintes est le fait que les relations extraites par ces systèmes ne peuvent pas être réutilisées par d'autres systèmes parce que le type de relation qu'on peut extraire n'est pas connu à priori. Dans les prochains paragraphes, nous présentons les résultats du processus de désambiguïsation.

Précision de la désambiguïsation

Dans le paragraphe précédent, nous avons trouvé que 56% des relations VerbNet ont au moins une classe VerbNet candidate associée. Dans cette sous-section, nous évaluons la proportion de ces relations que l'on arrive à désambiguïser correctement. Nous calculons donc la *précision de la désambiguïsation* qui indique le pourcentage de relations verbales pour lesquelles l'évaluateur juge que la bonne classe VerbNet a été sélectionnée. Le tableau 4.16 présente les résultats obtenus.

Tableau 4.16. Tableau indiquant la précision de la désambiguïsation calculée pour chacun des évaluateurs

Sujet	Evaluateur ₁	Evaluateur ₂	Moyenne	Nombre de relations	Nombre de relations en commun	AC1
Public Safety du 4 mai	82	78	80	42	34	0.62

Taxation du 4 mai	100	96	98	29	26	0.86
Ethics du 5 juin	87	92	89	22	21	0.91
Anti-terrorism Act du 5 mai	84	80	82	45	34	0.54
Taxation du 5 mai	100	84	92	21	17	0.67
Taxation du 6 mai	95	92	93	34	30	0.78
Finance du 7 décembre	77	70	73	15	11	0.46
National defense du 7 décembre	66	81	73	21	12	0.25
Taxation du 8 décembre	57	55	56	28	18	0.37

Tableau 4.16. Suite

Sujet	Evaluateur ₁	Evaluateur ₂	Moyenne	Nombre de relations	Nombre de relations commun	AC1
Infrastructure du 8 juin	100	100	100	23	22	0.93
Citizen Voting Act du 1 mai	90	94	92	50	45	0.81
Respecting Families of Murdered and Brutalized Person du 2 juin	85	84	84	50	43	0.72

Aboriginal Affairs du 3 juin	83	80	81	17	16	0.90
The economy du 3 juin	90	75	82	20	15	0.53
National Defence du 4 mai	77	75	76	20	15	0.55
Moyenne			83,6	28	22.5	0.66

Nous obtenons un taux de précision d'environ 84% pour la désambiguïsation. À noter que ce pourcentage est calculé seulement sur les relations qui ont un moins un candidat comme indiqué dans le paragraphe précédent. 47% des relations verbales sont correctement désambiguïsées.

4.5 Évaluation du graphe résumé

Finalement, nous avons demandé aux évaluateurs de donner leur point de vue qualitatif sur la sortie finale de notre système, soit le graphe résumé. Pour chacun des sujets de notre corpus d'évaluation, nous avons généré un graphe labélisé associé en considérant toutes les annotations agrégées et toutes les relations verbales provenant du sujet. Le graphe considère seulement les dix annotations agrégées les plus fréquentes du sujet. Le graphe résumé est ensuite étendu, c'est-à-dire que si une annotation du graphe résumé possède une relation verbale avec une annotation qui n'est pas initialement dans le graphe résumé, l'annotation est ajoutée au graphe. Ce processus est itéré jusqu'il n'y ait plus d'annotations à rajouter. Les caractéristiques des graphes générés pour chaque sujet sont présentées au tableau 4.17.

Nous avons demandé aux évaluateurs de répondre aux questions suivantes :

- i. Le graphe résume-t-il le sujet abordé ? Pourquoi ?
- ii. Les concepts du graphe sont-ils pertinents ? Pourquoi ?
- iii. Les relations du graphe sont-elles pertinentes ? Pourquoi ?

Tableau 4.17. Caractéristiques des graphes labélisés et résumés des sujets qui composent le corpus d'évaluation. Pour chaque type de graphe et pour chaque sujet de travail, on présente le nombre de sommets, le nombre d'arêtes, le degré minimum et maximum (entre crochets) et le degré moyen.

Sujets	Graphe labélisé complet			Graphe résumé		
	Nombre de sommets	Nombre d'arêtes	Degré	Nombre de sommets	Nombre d'arêtes	Degré
Public Safety du 4 mai	66	49	[0-4] 1.48	24	19	[0-4] 1.58
Taxation du 4 mai	55	37	[0-4] 1.34	14	11	[0-4] 1.57
Ethics du 5 juin	42	30	[0-3] 1.43	16	13	[0-3] 1.62
Anti-terrorism Act du 5 mai	89	64	[0-8] 1.43	24	20	[0-4] 1.67
Taxation du 5 mai	36	25	[0-3] 1.38	14	11	[0-2] 1.57
Taxation du 6 mai	66	54	[0-4] 1.63	18	15	[0-3] 1.66
Finance du 7 décembre	35	23	[0-3] 1.31	13	10	[0-3] 1.54

Tableau 4.17. Suite

Sujets	Graphe labélisé complet			Graphe résumé		
	Nombre de sommets	Nombre d'arêtes	Degré	Nombre de sommets	Nombre d'arêtes	Degré
National defense du 7 décembre	34	22	[0-5] 1.29	18	13	[0-3] 1.44
Taxation du 8 décembre	44	30	[0-5] 1.36	21	16	[0-4] 1.52
Infrastructure du 8 juin	65	45	[0-3] 1.38	14	10	[0-2] 1.42

Citizen Voting Act du 1 mai	107	84	[0-5]1.57	26	28	[0-4] 2.15
Respecting Families of Murdered and Brutalized Person du 2 juin	102	79	[0-7]1.54	26	22	[0-5] 1.69
Aboriginal Affairs du 3 juin	45	28	[0-4] 1.24	14	9	[0-2] 1.28
The economy du 3 juin	42	31	[0-6] 1.48	14	11	[0-4] 1.57
National Defence du 4 mai	38	22	[0-3] 1.16	10	5	[0-1]1.0
Moyenne	51.55	37.9	[0-3.9] 1.40	17.6	13.8	[0-3.2] 1.56

Pour l'évaluation du graphe résumé, les évaluateurs peuvent répondre par *oui*, ou par *non* pour indiquer que l'élément évalué est pertinent ou pas. Les évaluateurs peuvent répondre aussi par *presque* pour indiquer que l'élément pourrait être amélioré, mais n'est pas totalement inutile.

Le tableau 4.18 présente le résultat des évaluations des graphes résumés. Nous représentons les réponses *oui* par 1, les réponses *non* par 0 et les réponses *presque* par 0.5. Nous évaluons séparément la pertinence des sommets (annotation), la pertinence des arêtes (relations verbales) et la pertinence du graphe au complet.

Tableau 4.18. Résultats de l'évaluation des graphes résumés. Les indices 1 et 2 représentent les annotateurs

Sujet	Pertinence sommets			Pertinence arêtes			Pertinence graphe		
	Ev. 1	Ev. 2	Moy.	Ev. 1	Ev. 2	Moy.	Ev. 1	Ev. 2	Moy.
Public Safety du 4 mai	1	1	1	1	1	1	1	1	1
Taxation du 4 mai	1	1	1	1	1	1	0.5	0.5	0.5

Ethics du 5 juin	1	1	1	0	0.5	0.25	1	1	1
Anti-terrorism Act du 5 mai	1	1	1	1	0.5	0.75	1	1	1
Taxation du 5 mai	0.5	1	0.75	0.5	0.5	0.5	0.5	0.5	0.5
Taxation du 6 mai	0.5	1	0.75	1	0.5	0.75	1	1	1
Finance du 7 décembre	1	1	1	0.5	1	0.75	0.5	0.5	0.5
National defense du 7 décembre	0.5	1	0.75	1	1	1	1	1	1
Taxation du 8 décembre	1	1	1	1	0	0.5	1	1	1
Infrastructure du 8 juin	1	1	1	1	1	1	0	0	0
Citizen Voting Act du 1 mai	1	1	1	1	1	1	0.5	1	0.75
Respecting Families of Murdered and Brutalized Person du 2 juin	1	1	1	1	1	1	1	1	1
Aboriginal Affairs du 3 juin	1	0.5	0.75	1	0.5	0.75	0.5	0	0.25

Tableau 4.18. Suite

Sujet	Pertinence sommets			Pertinence arêtes			Pertinence graphe		
	Ev. 1	Ev. 2	Moy	Ev. 1	Ev. 2	Moy	Ev. 1	Ev. 2	Moy
The economy du 3 juin	1	0	0.5	1	0	0.5	1	0	0.5
National Defence du 4 mai	0.5	1	0.75	1	1	1	0.5	1	0.75
Moyenne	-	-	0.83	-	-	0.78	-	-	0.73

Nous obtenons une moyenne de 0.83 pour la pertinence des sommets, 0.78 pour la pertinence des arêtes et 0.73 pour la pertinence du graphe. Nous avons demandé aux évaluateurs de justifier les résultats en décrivant pourquoi ils trouvaient que certains éléments des graphes ou les graphes eux-mêmes étaient pertinents ou pas. Un exemple de réponse est présenté ci-dessous. Toutes les évaluations qualitatives sont fournies à l'annexe F.

Le graphe résume-t-il le sujet abordé ? Pourquoi ?

« En général, oui. Il y a quelques concepts inutiles, mais on comprend le sujet en regardant le graphe. On voit bien qu'on parle de la motion C-51 et d'une discussion de son impact. »

Les concepts du graphe sont-ils pertinents ? Pourquoi ?

« Oui, les concepts discutés sont présents parce qu'on retrouve la question de liberté fondamentale, de surveillance, de pouvoir additionnel, etc. »

Les relations du graphe sont-elles pertinentes ? Pourquoi ?

« Avec le CSIS oui. Par contre, je ne retrouve pas quelle partie a réagi sur quel concept. »

Plusieurs commentaires indiquaient que certaines idées ne ressortaient pas dans le graphe résumé. Nous pensons que ceci est dû au fait que certaines annotations n'étaient pas présentes dans le graphe résumé. Ces graphes ont été générés avec les dix annotations les plus fréquentes des sujets évalués. Notre hypothèse est que si on ajoutait plus d'annotations au graphe résumé, par exemple les vingt premières annotations plutôt que les dix premières, on aurait un rappel plus grand (au détriment de la précision, car certaines annotations ne seraient pas pertinentes) et donc peut être de meilleures évaluations.

Dans la question de recherche introduite au début de ce mémoire, nous nous demandions s'il était possible de résumer des débats politiques avec un graphe construit avec des annotations sémantiques et des relations extraites suite à une analyse de surface. Pour répondre à cette question, nous avons développé une méthodologie expliquée dans le chapitre 3 et dont nous présentons l'évaluation dans ce chapitre. Nous avons évalué les annotations sémantiques extraites et les scores associés à ces annotations. Il s'est avéré que la fréquence d'occurrence permettait de filtrer le plus d'annotations pertinentes. En considérant le top-50 des annotations avec la plus grande fréquence

d'occurrence, on obtient une précision de 72.8% qui monte à 76.4% quand on considère seulement le top-10. Ce résultat montre qu'il est possible d'extraire des annotations pertinentes. Nous avons évoqué, dans le chapitre 3, les lacunes de l'annotateur sémantique AlchemyAPI sur le fait que les annotations retournées ne correspondent pas toujours à la forme de surface exacte. Le processus de correction que nous avons proposé (appelé extension des annotations) a obtenu une précision moyenne de 83.7%. Nous avons aussi évalué les relations extraites sur plusieurs aspects qui incluent la précision du processus d'extraction et la pertinence des relations extraites. Nous avons obtenu respectivement 77.86% et 72%. On peut observer que les annotations et relations qui composent notre graphe résumé obtiennent individuellement de bons résultats. Nous avons donc évalué si les graphes extraits en combinant ces parties résument correctement les sujets. Nous avons effectué une évaluation qualitative des graphes en demandant aux évaluateurs d'indiquer si les graphes sortis sont pertinents. Nous avons obtenu une moyenne de 83% pour les annotations présentes sur le graphe, 78% pour les relations et 73% pour le graphe au complet. Notre réponse à la question de recherche initiale est que oui il est possible, dans une certaine mesure, d'abstraire un corpus de débat politique avec un graphe construit en combinant des annotations sémantiques et l'analyse de surface.

CHAPITRE 5 CONCLUSION

Dans notre travail, nous avons présenté une approche de représentation et d'abstraction de corpus que nous avons appliquée à un corpus de débats politiques. Notre approche est basée sur un graphe construit à partir d'annotations sémantiques et de relations extraites entre ces annotations. En faisant évaluer manuellement par des humains notre graphe résumé, nous concluons qu'il est possible de générer un graphe résumant un corpus politique en utilisant des annotations sémantiques et de l'analyse de surface pour l'extraction de relations.

Bien que les résultats obtenus soient satisfaisants, plusieurs améliorations restent possibles. La première serait d'étendre notre évaluation. Nous avons évalué notre système sur des sujets qui sont relativement courts (entre quinze et vingt-cinq interventions). Nous devrions évaluer notre système sur des corpus plus grands, constitué d'au moins une date complète ou idéalement de plusieurs dates de débats. Ce genre d'évaluation est toutefois coûteux, car l'évaluation est manuelle et requière beaucoup de temps. Aussi, bien que nous ayons choisi d'appliquer notre travail sur des corpus politiques, il est important de souligner que nous proposons une méthode indépendante du domaine. Toutefois, il serait intéressant de tester notre système sur des domaines spécifiques et de comparer nos résultats aux outils spécialisés dans ces domaines, comme [17] et [18].

Dans notre travail, nous avons extrait des annotations qui provenaient de la base de connaissances DBpedia. Nous avons aussi extrait les relations sémantiques entre ces concepts qui se trouvent sur DBpedia et nous avons utilisé ces données pour enrichir notre graphe résumé. Nous avons noté par contre qu'il existe très peu de relations sémantiques entre les concepts. Bien que DBpedia soit considéré par plusieurs [84] comme une base de connaissances *multi-domaines* (c'est-à-dire que les concepts décrits à l'intérieur de DBpedia proviennent de plusieurs domaines différents), elle ne décrit pas assez les relations entre ces concepts [85]. Pour remédier à ce problème, nous proposons l'enrichissement de notre modèle en y incluant d'autres bases de connaissances pour obtenir plus de relations sémantiques. Une des approches possibles serait de localiser les concepts précédemment extraits dans DBpedia à l'intérieur d'autres bases de connaissances telles que Wikidata [21]. Un des avantages avec DBpedia est qu'elle contient, pour plusieurs de ses concepts, des liens vers d'autres bases de connaissances avec la relation *owl:sameAs*.

Nous ne nous sommes pas limité aux relations sémantiques provenant de DBpedia, mais avons aussi mis en place un système d'extraction de relations verbales. Dans notre évaluation, nous avons rencontré plusieurs cas d'exception où la relation extraite entre les annotations était erronée (voir la section 4.3.2 pour des exemples). Ces erreurs sont dues à l'approche utilisée (l'analyse de surface) qui dans certains cas ne permet pas de déterminer si une annotation est vraiment le sujet ou l'objet d'une relation verbale. En utilisant des techniques plus complexes, comme des arbres de dépendances syntaxiques, ces erreurs pourraient être évitées. L'utilisation de ces techniques pour l'extension des annotations pourraient aussi améliorer les résultats obtenus. Par ailleurs, nous avons appliqué notre processus d'extraction de relations à toutes les annotations quelque soit leur score de pertinence. Une direction future pourrait être d'extraire des relations seulement entre les sujets les plus pertinents plutôt que de faire cette extraction pour toutes les annotations. Il serait aussi utile d'associer des scores de pertinence aux relations. Les évaluations qualitatives des graphes semblent aller dans cette direction puisqu'une observation récurrente des évaluateurs était le manque de pertinence de certaines relations. Nous avons considéré dans le prototype développé une méthode de filtrage qui permettrait d'afficher une relation seulement si les annotations reliées co-occurraient au-dessus d'un certain seuil. Cette heuristique devrait être évaluée sur sa capacité à conserver des relations pertinentes et le seuil de cooccurrence devrait être déterminé. Enfin, même si nous pensons que les verbes permettent de capturer des liens entre les annotations, d'autres éléments linguistiques pourraient apporter un plus au graphe produit. Nous pourrions par exemple ajouter un module d'extraction de relations qui serait basé sur des éléments linguistiques différents des verbes, par exemple les relations émergeant d'expressions nominales.

Dans le but de fournir une approche généraliste et qui ne nécessite pas de traitements linguistiques lourds, nous avons utilisé au cours de ce travail les instruments les plus simples possibles. Cependant, il serait intéressant d'étudier l'impact qu'aurait l'utilisation de techniques plus sophistiquées telles que de l'analyse syntaxique ou spécifique à certains domaines sur nos résultats.

Finalement, un générateur de texte pourrait être branché à notre système pour produire des résumés sous forme textuelle. Nous avons mis en place un modèle de données riche où chaque élément du graphe résumé est directement relié à ses occurrences dans le corpus. Ce modèle pourrait être exploité pour générer du texte, et ainsi comparer notre système à toutes les autres techniques de générations de résumé qui produisent du texte. Le modèle de données est disponible en annexe A.

BIBLIOGRAPHIE

- [1] Daniil, C., Dascalu, M., & Trausan-Matu, S. (2012, June). Automatic forum analysis: a thorough method of assessing the importance of posts, discussion threads and of users' involvement. In *Proceedings of the 2Nd International Conference on Web Intelligence, Mining and Semantics* (p. 37). ACM.
- [2] Jeong, H., Hatori, T., & Kobayashi, K. (2007, October). Discourse analysis of public debates: A corpus-based approach. In *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on* (pp. 1782-1793). IEEE.
- [3] Wu, W., Zhang, B., & Ostendorf, M. (2010, June). Automatic generation of personalized annotation tags for twitter users. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics* (pp. 689-692). Association for Computational Linguistics.
- [4] Mao, W., Xiao, L., & Mercer, R. (2014). The Use of Text Similarity and Sentiment Analysis to Examine Rationales in the Large-Scale Online Deliberations. *ACL 2014*, 147.
- [5] Marx, M., & Aders, N. (2010). From documents to data: linked data at the dutch parliament.
- [6] Kaplan, I., & Rosenberg, A. (2012, December). Analysis of speech transcripts to predict winners of US Presidential and Vice-Presidential debates. In *Spoken Language Technology Workshop (SLT), 2012 IEEE* (pp. 449-454). IEEE.
- [7] Salah, Z., Coenen, F., & Grossi, D. (2013, June). Extracting debate graphs from parliamentary transcripts: A study directed at UK House of Commons debates. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law* (pp. 121-130). ACM.
- [8] Woudenberg, D. (2013). Understanding Arguments-Predicting the outcome of debates using structural features. Bachelor thesis, University of Amsterdam.
- [9] Lin, F. R., Chou, S. Y., Liao, D., & Hao, D. (2015, January). Automatic Content Analysis of Legislative Documents by Text Mining Techniques. In *System Sciences (HICSS), 2015 48th Hawaii International Conference on* (pp. 2199-2208). IEEE.
- [10] van Wees, J., Marx, M., & van Doornik, J. (2011). Applying Social Network Analysis to Parliamentary Proceedings.

- [11] Gurciullo, S., Smallegan, M., Pereda, M., Battiston, F., Patania, A., Poledna, S., Hedblom, D., Oztan, B., T., Herzog, A., John, P. & Mikhaylov, S. (2015). Complex Politics: A Quantitative Semantic and Topological Analysis of UK House of Commons Debates. arXiv preprint arXiv:1510.03797.
- [12] Greene, D., & Cross, J. P. (2015). Unveiling the Political Agenda of the European Parliament Plenary: A Topical Analysis. arXiv preprint arXiv:1505.07302.
- [13] Zhang, J. J., & Fung, P. (2012). Automatic parliamentary meeting minute generation using rhetorical structure modeling. *Audio, Speech, and Language Processing, IEEE Transactions on Audio, Speech and Language Processing*, 20(9), 2492-2504.
- [14] Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 457-479.
- [15] Ganesan, K., Zhai, C., & Han, J. (2010, August). Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd international conference on computational linguistics* (pp. 340-348). Association for Computational Linguistics.
- [16] Liu, F., Flanigan, J., Thomson, S., Sadeh, N., & Smith, N. A. (2015). Toward abstractive summarization using semantic representations.
- [17] Fiszman, M., Rindflesch, T. C., & Kilicoglu, H. (2004, May). Abstraction summarization for managing the biomedical research literature. In *Proceedings of the HLT-NAACL workshop on computational lexical semantics* (pp. 76-83). Association for Computational Linguistics.
- [18] Zhang, H., Fiszman, M., Shin, D., Miller, C. M., Rosembat, G., & Rindflesch, T. C. (2011). Degree centrality for semantic abstraction summarization of therapeutic studies. *Journal of biomedical informatics*, 44(5), 830-838.
- [19] Abraham, A., Hassanien, A. E., & Snášel, V. (Eds.). (2009). *Computational social network analysis: Trends, tools and research advances*. Springer Science & Business Media.
- [20] Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.
- [21] Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10), 78-85.

- [22] Salton, G., & McGill, M. J. (1986). Introduction to modern information retrieval, McGraw-Hill, Inc., New York, NY,
- [23] Becker, J., & Kuropka, D. (2003, July). Topic-based vector space model. In Proceedings of the 6th international conference on business information systems (pp. 7-12).
- [24] Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1), 1-27.
- [25] Mihalcea, R., & Tarau, P. (2004, July). TextRank: Bringing order into texts. Association for Computational Linguistics.
- [26] Derényi, I., Palla, G., & Vicsek, T. (2005). Clique percolation in random networks. *Physical review letters*, 94(16).
- [27] Blei, D. M., & Lafferty, J. D. (2006, June). Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning (pp. 113-12c0). ACM.
- [28] Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, 46(2), 255-308.
- [29] Paatero, P., & Tapper, U. (1994). Positive matrix factorization: A non negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2), 111-126.
- [30] AlchemyAPI | Powering the New AI Economy <http://www.alchemyapi.com/>
- [31] Sipos, R., Swaminathan, A., Shivaswamy, P., & Joachims, T. (2012, October). Temporal corpus summarization using submodular word coverage. In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 754-763). ACM.
- [32] Jin, W., & Srihari, R. K. (2007, March). Graph-based text representation and knowledge discovery. In Proceedings of the 2007 ACM symposium on Applied computing (pp. 807-811). ACM.
- [33] Augenstein, I., Padó, S., & Rudolph, S. (2012). Lodifier: Generating linked data from unstructured text. In *The Semantic Web: Research and Applications* (pp. 210-224). Springer Berlin Heidelberg.
- [34] Mihalcea, R., & Tarau, P. (2004, July). TextRank: Bringing order into texts. Association for Computational Linguistics.

- [35] Litvak, M., & Last, M. (2008, August). Graph-based keyword extraction for single-document summarization. In *Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization* (pp. 17-24). Association for Computational Linguistics.
- [36] Zhou, F., Zhang, F., & Yang, B. (2010, August). Graph-based text representation model and its realization. In *Natural Language Processing and Knowledge Engineering (NLP-KE), 2010 International Conference on* (pp. 1-8). IEEE.
- [37] Mihalcea, R. (2004, July). Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions* (p. 20). Association for Computational Linguistics.
- [38] Abilhoa, W. D., & de Castro, L. N. (2014). A keyword extraction method from twitter messages represented as graphs. *Applied Mathematics and Computation*, 240, 308-325.
- [39] Hensman, S. (2004, May). Construction of conceptual graph representation of texts. In *Proceedings of the Student Research Workshop at HLT-NAACL 2004* (pp. 49-54). Association for Computational Linguistics.
- [40] Ajgalik, M., Barla, M., & Bieliková, M. (2013, August). From ambiguous words to key-concept extraction. In *Database and Expert Systems Applications (DEXA), 2013 24th International Workshop on* (pp. 63-67). IEEE.
- [41] Kozareva, Z., Riloff, E., & Hovy, E. H. (2008, June). Semantic Class Learning from the Web with Hyponym Pattern Linkage Graphs. In *ACL (Vol. 8, pp. 1048-1056)*.
- [42] Mani, I., & Bloedorn, E. (1997). Multi-document summarization by graph search and matching. *arXiv preprint cmp-lg/9712004*.
- [43] Schuler, K. K. (2007). VerbNet: extensions and mappings to other lexical resources. In *2007-01-04*.
- [44] Mooney, R. J., & Bunescu, R. C. (2005). Subsequence kernels for relation extraction. In *Advances in neural information processing systems* (pp. 171-178).
- [45] Zelenko, D., Aone, C., & Richardella, A. (2003). Kernel methods for relation extraction. *The Journal of Machine Learning Research*, 3, 1083-1106.

- [46] Culotta, A., & Sorensen, J. (2004, July). Dependency tree kernels for relation extraction. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (p. 423). Association for Computational Linguistics.
- [47] Bunescu, R. C., & Mooney, R. J. (2005, October). A shortest path dependency kernel for relation extraction. In Proceedings of the conference on human language technology and empirical methods in natural language processing (pp. 724-731). Association for Computational Linguistics.
- [48] Airola, A., Pyysalo, S., Björne, J., Pahikkala, T., Ginter, F., & Salakoski, T. (2008, June). A graph kernel for protein-protein interaction extraction. In Proceedings of the workshop on current trends in biomedical natural language processing (pp. 1-9). Association for Computational Linguistics.
- [49] Brin, S. (1998). Extracting patterns and relations from the world wide web. In *The World Wide Web and Databases* (pp. 172-183). Springer Berlin Heidelberg.
- [50] Agichtein, E., & Gravano, L. (2000, June). Snowball: Extracting relations from large plain-text collections. In Proceedings of the fifth ACM conference on Digital libraries (pp. 85-94). ACM.
- [51] Etzioni, O., Cafarella, M., Downey, D., Popescu, A. M., Shaked, T., Soderland, S., Weld, D., S. & Yates, A. (2005). Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*, 165(1), 91-134.
- [52] Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O. (2007). Open information extraction from the web. *IJCAI '07: Proceedings of the 20th International Joint Conference on Artificial Intelligence*. Hyderabad, India.
- [53] Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., & Watkins, C. (2002). Text classification using string kernels. *The Journal of Machine Learning Research*, 2, 419-444.
- [54] Kambhatla, N. (2004, July). Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In Proceedings of the ACL 2004 on Interactive poster and demonstration sessions (p. 22). Association for Computational Linguistics.
- [55] GuoDong, Z., Jian, S., Jie, Z., & Min, Z. (2005, June). Exploring various knowledge in relation extraction. In Proceedings of the 43rd annual meeting on association for computational linguistics (pp. 427-434). Association for Computational Linguistics.

- [56] Vapnik, V. N., & Vapnik, V. (1998). Statistical learning theory (Vol. 1). New York: Wiley.
- [57] Berger, A. L., Pietra, V. J. D., & Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1), 39-71.
- [58] Riloff, E., & Jones, R. (1999, July). Learning dictionaries for information extraction by multi-level bootstrapping. In *AAAI/IAAI* (pp. 474-479).
- [59] Pasca, M. (2004, November). Acquisition of categorized named entities for web search. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management* (pp. 137-145). ACM.
- [60] Hearst, M. A. (1992, August). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2* (pp. 539-545). Association for Computational Linguistics.
- [61] Cole, S. V., Royal, M. D., Valtorta, M. G., Huhns, M. N., & Bowles, J. B. (2005, March). A lightweight tool for automatically extracting causal relationships from text. In *SoutheastCon, 2006. Proceedings of the IEEE* (pp. 125-129). IEEE.
- [62] Juric, D., Hollink, L., & Houben, G. J. (2012). Bringing parliamentary debates to the Semantic Web. *Detection, Representation, and Exploitation of Events in the Semantic Web*, 51.
- [63] Fader, A., Soderland, S., & Etzioni, O. (2011, July). Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1535-1545). Association for Computational Linguistics.
- [64] Qiu, L., & Zhang, Y. (2014). ZORE: A Syntax-based System for Chinese Open Relation Extraction. In *EMNLP* (pp. 1870-1880).
- [65] Nakashole, N., Weikum, G., & Suchanek, F. (2012, July). PATTY: a taxonomy of relational patterns with semantic types. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 1135-1145). Association for Computational Linguistics.
- [66] Schmitz, M., Bart, R., Soderland, S., & Etzioni, O. (2012, July). Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in*

Natural Language Processing and Computational Natural Language Learning (pp. 523-534). Association for Computational Linguistics.

[67] Xu, Y., Kim, M. Y., Quinn, K., Goebel, R., & Barbosa, D. (2013, June). Open Information Extraction with Tree Kernels. In HLT-NAACL (pp. 868-877).

[68] Johansson, R., & Nugues, P. (2008, October). Dependency-based semantic role labeling of PropBank. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 69-78). Association for Computational Linguistics.

[69] Surdeanu, M., Harabagiu, S., Williams, J., & Aarseth, P. (2003, July). Using predicate-argument structures for information extraction. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1 (pp. 8-15). Association for Computational Linguistics.

[70] Christensen, J., Soderland, S., & Etzioni, O. (2011, June). An analysis of open information extraction based on semantic role labeling. In Proceedings of the sixth international conference on Knowledge capture (pp. 113-120). ACM.

[71] Bach, N., & Badaskar, S. (2007). A review of relation extraction. Literature review for Language and Statistics II.

[72] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychosocial Measurement*, 20, 37-46.

[73] Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

[74] Tallis, M. (2003, October). Semantic word processing for content authors. In Proceedings of the Knowledge Markup & Semantic Annotation Workshop, Florida, USA.

[75] Kogut, P. A., & Holmes III, W. S. (2001, October). AeroDAML: Applying Information Extraction to Generate DAML Annotations from Web Pages. In Semannot@ K-CAP 2001.

[76] Dingli, A., Ciravegna, F., & Wilks, Y. (2003, October). Automatic semantic annotation using unsupervised information extraction and integration. In Proceedings of SemAnnot 2003 Workshop.

- [77] Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D., & Goranov, M. (2003). KIM–semantic annotation platform. In *The Semantic Web-ISWC 2003* (pp. 834-849). Springer Berlin Heidelberg.
- [78] Schenker, A., Last, M., Bunke, H., & Kandel, A. (2003, August). Classification of web documents using a graph model. In *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on* (pp. 240-244). IEEE.
- [79] Joachims, T. (1996). *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization* (No. CMU-CS-96-111118). Carnegie-mellon univ pittsburgh pa dept of computer science.
- [80] Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Kleef, P., Auer, S. & Bizer, C. (2015). DBpedia—a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2), 167-195.
- [81] Klyne, G., & Carroll, J. J. (2006). *Resource description framework (RDF): Concepts and abstract syntax*.
- [82] Gagnon, M., Zouaq, A., & Jean-Louis, L. (2013, May). Can we use linked data semantic annotators for the extraction of domain-relevant expressions?. In *Proceedings of the 22nd international conference on World Wide Web companion* (pp. 1239-1246). International World Wide Web Conferences Steering Committee.
- [83] Jean-Louis, L., Zouaq, A., Gagnon, M., & Ensan, F. (2014). An assessment of online semantic annotators for the keyword extraction task. In *PRICAI 2014: Trends in Artificial Intelligence* (pp. 548-560). Springer International Publishing.
- [84] Mendes, P. N., Jakob, M., & Bizer, C. (2012, May). DBpedia: A Multilingual Cross-domain Knowledge Base. In *LREC* (pp. 1813-1817).
- [85] Font, L., Zouaq, A., & Gagnon, M. (2015, November). Assessing the Quality of Domain Concepts Descriptions in DBpedia. In *2015 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)* (pp. 254-261). IEEE.
- [86] Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). A comparison of Cohen’s Kappa and Gwet’s AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC medical research methodology*, 13(1), 1.

[87] Viswanathan M, Berkman ND. Development of the RTI Item Bank on Risk of Bias and Precision of Observational Studies [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2011 Sep. Appendix A, AC1 Statistic. Available from: <http://www.ncbi.nlm.nih.gov/books/NBK82266/>

[88] Gwet K. Handbook of Inter-Rater Reliability: How to Estimate the Level of Agreement Between Two or Multiple Raters. Gaithersburg, MD: STATAXIS Publishing Company; 2001.

ANNEXE A – MODÈLE DE DONNÉE

La figure 3 représente le modèle de donnée utilisé pour le projet. Dans la conception du modèle, l'accent est mis sur les performances de l'application, vu le nombre de données à traiter.

Dans ce qui suit, nous décrivons les entités du modèle plus en détail :

- **Forum** : un forum représente un ensemble de messages relatifs à une ou plusieurs dates du Hansard. Il contient des tables de hachage pour récupérer rapidement les messages en fonction de leur date, auteur, sujet de travail ou ordre de travail.
- **Post** : un post représente une intervention d'un membre du parlement. Il contient le contenu de l'intervention partitionné en plusieurs phrases (Sentences). Il contient aussi l'ensemble des annotations relatives à l'intervention.
- **Sentence et Token** : Un objet de type Sentence est essentiellement un ensemble de mots (Tokens) où pour chaque token, on a un label (forme originale sous laquelle le token apparaît), son lemme et son étiquette grammaticale (POS Treetagger).
- **Term et Annotation** : Un objet de type Term contient l'ensemble des annotations qui ont le même lemme. Il contient aussi tous les scores associés (Tf, tfidf, pertinence, etc.)
- **Cooccurrence** : Un objet de type Cooccurrence représente l'évidence que deux termes apparaissent ensemble dans au moins un groupe. Un groupe est un concept abstrait et permet de représenter les relations de cooccurrences selon plusieurs granularités. Un groupe peut être une phrase, un message, une date particulière, etc. La fréquence de cooccurrence est le nombre de groupes dans lesquels ces 2 termes apparaissent.
- **PostCooccurrence** : Un objet de type PostCooccurrence représente l'évidence que deux termes x et y apparaissent ensemble dans au moins une intervention. L'attribut *intersection of posts* représente l'ensemble des interventions où x et y apparaissent ensemble.
- **SentenceCooccurrence** : Un objet de type SentenceCooccurrence représente l'évidence que deux termes x et y apparaissent ensemble dans au moins une phrase. L'attribut *intersection of sentences* représente l'ensemble des phrases où x et y apparaissent ensemble. Un objet de type SentenceCooccurrence contiendra l'attribut *List of verbs* qui représente la liste des verbes candidats à la relation verbale entre x et ainsi que l'attribut *List of Verbnet*, qui

représente la liste des catégories VerbNet correspondantes. L'attribut relation contiendra la relation verbale finalement choisie.

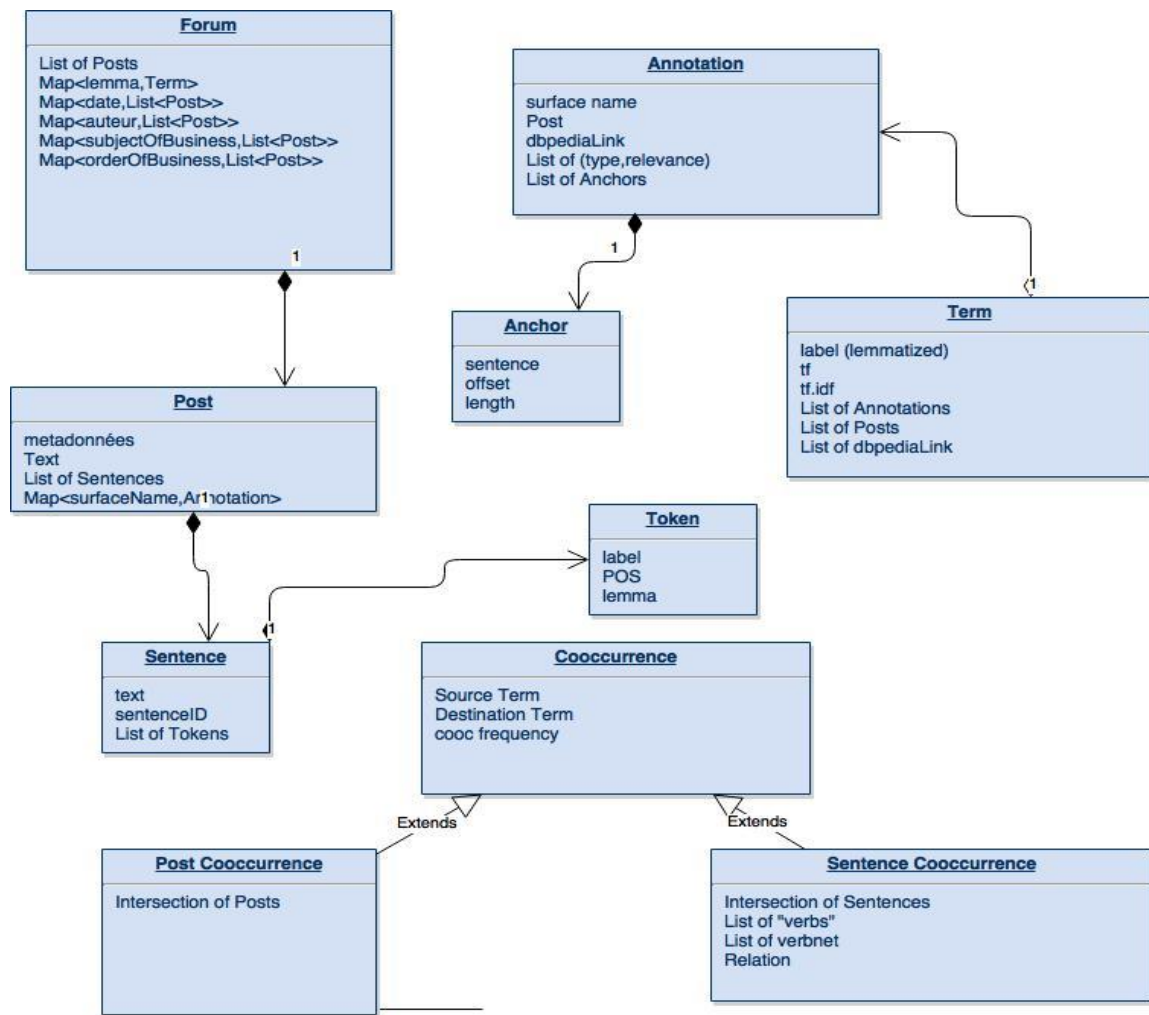


Figure A.1. modèle de donnée

ANNEXE B - DESCRIPTION DE L'APPLICATION

L'application prend en entrée un corpus du Hansard constitué d'un ensemble de fichiers XML relatifs aux dates qui constituent le corpus.

Après lecture des fichiers XML et l'extraction d'information, on représente les graphes extraits du corpus au travers de quatre vues distinctes présentées ci-dessous. Les fonctionnalités représentées dans les vues y sont expliquées.

Cooccurrence Viewer

Dans cette vue, on génère les graphes non labélisés en fonction des paramètres statistiques :

- Par occurrence : on sélectionne x , qui est le nombre de nœuds. Le graphe sera constitué des x nœuds qui apparaissent le plus de fois dans le corpus ainsi que les relations de cooccurrence entre ces nœuds ;
- Par pertinence : on sélectionne x , qui est le nombre de nœuds. Le graphe sera constitué des x nœuds qui ont les scores de pertinence les plus élevés du corpus ainsi que les relations de cooccurrence entre ces nœuds ;
- Par cooccurrence : on sélectionne x , qui est le nombre d'arcs. Le graphe sera constitué des nœuds qui ont les cooccurrences par interventions les plus élevées, telles que le nombre d'arcs soit égal à x .

Relation Viewer

Dans cette vue, on génère les graphes avec des relations verbales en utilisant les paramètres statistiques. On peut générer le graphe de 3 manières différentes :

- Par occurrence : on sélectionne x , qui est le nombre de nœuds. Le graphe sera constitué des x nœuds qui courent le plus de fois dans le corpus ainsi que les relations verbales entre ces nœuds
- Par pertinence : on sélectionne x , qui est le nombre de nœuds. Le graphe sera constitué des x nœuds qui ont les pertinences les plus élevées du corpus ainsi que les relations verbales entre ces nœuds
- Par cooccurrence : On sélectionne x , qui est le nombre d'arcs. Le graphe sera constitué des nœuds qui ont les cooccurrences par phrases les plus élevées si celles-ci contiennent au moins un verbe, tel que le nombre d'arcs soit x .

CMap Viewer

Dans cette vue, on veut analyser un concept à la fois. On génère donc une carte de concepts autour du concept choisi. Les relations ici sont constituées des cooccurrences par interventions. Quand on double- clique sur un des concepts de la liste, un graphe est créé où l'on voit tous les autres concepts liés au concept sélectionné. On peut également continuer d'étendre le graphe en cliquant sur un nœud du graphe. D'autres liens et éventuellement d'autres nœuds s'ajoutent alors. Si on clique de nouveau sur un nœud qui s'est étendu, les nœuds et les relations ajoutés au graphe sont effacés.

CMap Relation

Cette vue est similaire au CMap viewer, mais est basée sur les cooccurrences par phrases et les relations verbales. On sélectionne une entité et on peut voir les relations verbales qu'elle a avec les autres entités.

Utilitaires

Chacune des vues présentées ci-dessus contient les utilitaires suivants :

- **filtrage tf-idf** : on élimine les concepts les moins pertinents à partir de leur score tf-idf. Cela est fait de deux façons : soit on insère un pourcentage x indiquant que les x% concepts les plus bas devront être éliminés soit on spécifie un seuil x, indiquant que tous les concepts ayant un tf-idf plus bas que x sont éliminés.
- **Filtrage par type** : les concepts extraits sont typés. On peut sélectionner l'ensemble des types de concept qu'on voudrait visualiser. Par défaut, les concepts de tous les types sont visualisés.
- **Filtrage du corpus** : Un corpus est constitué d'un ensemble de dates, une date est constituée de plusieurs ordres de travaux et chaque ordre de travail est constitué de plusieurs sujets de travail. On peut générer un graphe relatif à une date particulière, un ordre de travail particulier, un sujet de travail particulier, un auteur particulier ou une combinaison de ceux-ci.
- **Pour chaque terme**, il est possible de voir toutes les interventions dont il est extrait et pour chacune de ces interventions, on peut aussi voir les catégories d'AlchemyAPI auxquelles il est rattaché (exemple : mot-clé).

Pour chaque relation labélisée ou non labélisée, on peut visualiser toutes les interventions (relations non labélisées) ou toutes les phrases (relation labélisée) d'où la relation est extraite.

ANNEXE C - CATÉGORIES GRAMMATICALES

Le Tableau C.1 représente les catégories grammaticales. Ces catégories correspondent à celles fournies par l'outil TreeTagger. Ces données proviennent de : <https://courses.washington.edu/hypertext/csar-v02/penntable.html>

Tableau C.1. Catégories grammaticales utilisées.

Catégories grammaticales	Description	Exemples
CC	coordinating conjunction	<i>and, but, or, &</i>
CD	cardinal number	<i>1, three</i>
DT	determiner	<i>the</i>
EX	existential there	<i>there is</i>
FW	foreign word	<i>d'œuvre</i>
IN	preposition/subord. conj.	<i>in, of, like, after, whether</i>
IN/that	complementizer	<i>that</i>
JJ	adjective	<i>green</i>
JJR	adjective, comparative	<i>greener</i>
JJS	adjective, superlative	<i>greenest</i>
LS	list marker	<i>(1),</i>
MD	modal	<i>could, will</i>
NN	noun, singular or mass	<i>table</i>
NNS	noun plural	<i>tables</i>

Tableau C.1. Suite

Catégories grammaticales	Description	Exemples
NP	proper noun, singular	<i>John</i>
NPS	proper noun, plural	<i>Vikings</i>
N	All noun	
PDT	predeterminer	<i>both the boys</i>
POS	possessive ending	<i>friend's</i>
PP	personal pronoun	<i>I, he, it</i>
PP\$	possessive pronoun	<i>my, his</i>
RB	adverb	<i>however, usually, here, not</i>
RBR	adverb, comparative	<i>better</i>
RBS	adverb, superlative	<i>best</i>
RP	particle	<i>give up</i>
SENT	end punctuation	<i>?, !, .</i>
SYM	symbol	<i>@, +, *, ^, , =</i>
TO	<i>to</i>	<i>to go, to him</i>
UH	interjection	<i>uhhuhhuhh</i>
VB	verb <i>be</i> , base form	<i>be</i>
VBD	verb <i>be</i> , past	<i>was were</i>
VBG	verb <i>be</i> , gerund/participle	<i>being</i>

Tableau C.1. Suite

Catégories grammaticales	Description	Exemples
VBN	verb <i>be</i> , past participle	<i>been</i>
VBZ	verb <i>be</i> , pres, 3rd p. sing	<i>is</i>
VBP	verb <i>be</i> , pres non-3rd p.	<i>am are</i>
VD	verb <i>do</i> , base form	<i>do</i>
VDD	verb <i>do</i> , past	<i>did</i>
VDG	verb <i>do</i> gerund/participle	<i>doing</i>
VDN	verb <i>do</i> , past participle	<i>done</i>
VDZ	verb <i>do</i> , pres, 3rd per.sing	<i>does</i>
VDP	verb <i>do</i> , pres, non-3rd per.	<i>do</i>
VH	verb <i>have</i> , base form	<i>have</i>
VHD	verb <i>have</i> , past	<i>had</i>
VHG	verb <i>have</i> , gerund/participle	<i>having</i>
VHN	verb <i>have</i> , past participle	<i>had</i>
VHZ	verb <i>have</i> , pres 3rd per.sing	<i>has</i>
VHP	verb <i>have</i> , pres non-3rd per.	<i>have</i>
VV	verb, base form	<i>take</i>
VVD	verb, past tense	<i>took</i>

Tableau C.1. Suite

Catégories grammaticales	Description	Exemples
VVG	verb, gerund/participle	<i>taking</i>
VVN	verb, past participle	<i>taken</i>
VVP	verb, present, non-3rd p.	<i>take</i>
VVZ	verb, present 3d p. sing.	<i>takes</i>
VERB	all verbs	
WDT	wh-determiner	<i>which</i>
WP	wh-pronoun	<i>who, what</i>
WP\$	possessive wh-pronoun	<i>whose</i>
WRB	wh-abverb	<i>where, when</i>
:	general joiner	<i>;, -, --</i>
\$	currency symbol	<i>\$, £</i>

ANNEXE D - EXPORTATION DU GRAPHE AU FORMAT RDF

Les graphes générés peuvent être exportés sous forme RDF. La structure diffère légèrement entre les graphes avec des relations labélisées et ceux avec des relations non labélisées parce que pour le cas des relations labélisées, on a de l'information supplémentaire sur les arcs, comme la relation verbale, les classes VerbNet candidates ou encore la classe Verbnets désambiguïsée.

Le fichier exporté représente bien évidemment un graphe ; pour l'explication de la structure, on procèdera d'abord par les nœuds (représentant les termes) et ensuite par les arcs (représentant les relations).

Les termes : un terme représente une annotation agrégée avec les métriques correspondantes. La structure RDF pour un terme est la suivante :

```
[ a          "Term" ;
  <Alchemy type>  [valeur] ;
  <relevanceScore> [valeur] ;
  <surface>       [valeur] ;
  <tfScore>       [valeur] ;
  <tfidfScore>    [valeur]
].
```

Cette structure est indépendante du type de graphe.

Les Relations : en ce qui concerne les relations, la structure du fichier RDF dépend du type de graphe.

Les relations non labélisées

Les relations non labélisées ont la structure suivante :

```
[ a          "Link" ;
  <destination> [valeur] ;
```

<score> [valeur] ;

<source> [valeur] ;

].

Les relations labélisées

Les relations labélisées ont la structure suivante :

[a "Link" ;

<clearWSD> [valeur] ;

<source> [valeur] ;

<destination> [valeur] ;

<relation> [valeur] ;

<score> [valeur] ;

<verbnet [valeur];

]

ANNEXE E - INTERPRÉTATION DU KAPPA DE COHEN

Le Tableau E.1 représente pour différents intervalles de valeurs de kappa, l'interprétation correspondante.

Tableau E.1. Interprétation des valeurs du Kappa.

Valeur du Kappa	Interprétation
<0.00	Désaccord
0.00 – 0.20	Accord très faible
0.21-0.40	Accord faible
0.41 – 0.60	Accord modéré
0.61- 0.80	Accord fort
0.81-1.00	Accord presque parfait

ANNEXE F – EVALUATIONS QUALITATIVES DES GRAPHES RÉSUMÉS

Les tableaux F.1 à F.15 présentent les commentaires des évaluateurs pour différents sujets.

Tableau F.1. Évaluations qualitatives du graphe résumé relatif au sujet Public Safety du 4 mai

	Le graphe résume-t-il le sujet abordé ? Pourquoi ?	Les concepts du graphe sont-ils pertinents ? Pourquoi ?	Les relations du graphe sont-elles pertinentes ? Pourquoi ?
Réponse de l'évaluateur 1	Les idées importantes sont là.	En général les concepts sont pertinents	Les relations sont pertinentes mais ne sont pas toujours correctement extraites. (le problème est souvent dans le sujet si je ne me trompe pas)
Réponse de l'évaluateur 2	En général, oui. Il y a quelques concepts inutiles, mais le comprend le sujet en regardant le graphe. On voit bien qu'on parle de la motion C-51 et d'une discussion de son impact.	Oui, les concepts discutés sont présents parce qu'on retrouve la question de liberté fondamentale, de surveillance, de pouvoir additionnel, etc.	Avec le CSIS oui. Par contre, je ne retrouve pas quel partie a réagi sur quel concept.

Tableau F.2. Évaluations qualitatives du graphe résumé relatif au sujet Taxation du 4 mai

	Le graphe résume-t-il le sujet abordé ? Pourquoi ?	Les concepts du graphe sont-ils pertinents ? Pourquoi ?	Les relations du graphe sont-elles pertinentes ? Pourquoi ?
Réponse de l'évaluateur 1	Il manque des informations, mais ce qui y est est en majorité correct	oui, certains concepts ne sont pas complets. (Par exemple : child care benefit increase -- will help --> 100% of family, l'objet devrait être 100% of families with children)	oui
Réponse de l'évaluateur 2	Plus ou moins. Chaque partie vise des solutions pour un petit ensemble de la population. Le graphe ne montre pas clairement le population visé. Aussi, il a de la misère avec la sémantique des chiffres. Par exemple, on retrouve « child care benefit will give families up to 2000\$ ». Il manque la mesure associée au chiffre. (2000\$/an?)	Oui. Les parties qui discutent sont là. Les programmes discutés sont là.	En général oui. Mais le graphe ne dit pas de qui vient le universal child care

Tableau F.3. Évaluations qualitatives du graphe résumé relatif au sujet Ethics du 5 juin

	Le graphe résume-t-il le sujet abordé ? Pourquoi ?	Les concepts du graphe sont-ils pertinents ? Pourquoi ?	Les relations du graphe sont-elles pertinentes ? Pourquoi ?
Réponse de l'évaluateur 1	Il y'a des informations manquantes	oui	certaines relations ne son pas pertinentes. (canadian -- wake up --> morning))
Réponse de l'évaluateur 2	Oui si on a une connaissance préalable sur le sujet. Sinon, je ne crois pas qu'une personne externe qui n'a pas lu le sujet puisse prendre le graphe. Le graphe n'explique pas réellement le problème en soi. Par exemple, « inappropriate expense » n'est pas relié avec le sénat.	Oui les concepts pertinents sont présents.	Non. Par exemple, « inappropriate expense » n'est pas relié avec le sénat. Aucun lien entre le « appeal process » et les dépenses.

Tableau F.4. Évaluations qualitatives du graphe résumé relatif au sujet Anti-terrorism Act du 5 mai

	Le graphe résume-t-il le sujet abordé ? Pourquoi ?	Les concepts du graphe sont-ils pertinents ? Pourquoi ?	Les relations du graphe sont-elles pertinentes ? Pourquoi ?
Réponse de l'évaluateur 1	oui les gros points sont la	oui	certaines sujets ne sont pas bien choisis
Réponse de l'évaluateur 2	Oui. On voit bien l'opposition au projet de loi Bill C-51. On voit les arguments important qui inquiète les partis d'opposition.	Les concepts discutés sont présents. Quelques concepts inutiles se retrouvent tel que « the house ». En général, ils sont présents.	On voit bien l'opposition au Bill-51. Par contre, il n'y a pas de lien entre Bill-51 et CSIS, ni les parties.

Tableau F.5. Évaluations qualitatives du graphe résumé relatif au sujet Taxation du 5 mai

	Le graphe résume-t-il le sujet abordé ? Pourquoi ?	Les concepts du graphe sont-ils pertinents ? Pourquoi ?	Les relations du graphe sont-elles pertinentes ? Pourquoi ?
Réponse de l'évaluateur 1	Il y'a des information qui manquent	oui	Certaines relations sont inutiles si elles ne sont pas connecté a d'autres. (people -- earning less than --> 6000\$)
Réponse de l'évaluateur 2	Manque de contexte pour les données avec chiffre. (2000\$/mois ou /an?).En général, on peut déduire le sujet de la discussion, mais le graphe n'est pas clair. Certaines phrases nominales (parent) On ne parle pas de plusieurs points abordés dans le débat (le déficit de 2B des Libéraux).	Plusieurs concepts inutiles (government). Mais plusieurs concepts importants sont présents.	Pour un des graphe oui. Pas pour les 2 autres.

Tableau F.6. Évaluations qualitatives du graphe résumé relatif au sujet Taxation du 6 mai

	Le graphe résume-t-il le sujet abordé ? Pourquoi ?	Les concepts du graphe sont-ils pertinents ? Pourquoi ?	Les relations du graphe sont-elles pertinentes ? Pourquoi ?
Réponse de l'évaluateur 1	Non : les relations utilisées ne résument qu'une partie de l'idée. Par exemple, "measure put money" est une relation correcte, pertinente, mais sans intérêt sans son contexte.	a plupart, oui.	Certaines sont trop longues, et donc souvent non pertinentes ("through more high taxes on", "brought in the universal child care benefit for")
Réponse de l'évaluateur 2	oui je pense que le graphe résume bien le sujet, peut être il ne contient pas toutes les informations pertinentes, mais moi quand je lis les différentes relations et concepts ça m'aide à avoir une idée de ce qui a été mentionné	Il y a le concept (plan) qui est inutile je pense, car en lisant la liaison ça n'a pas de sens à mon avis	oui les relations sont bien pertinentes à part la liaison qui est liée au concept plan mentionné précédemment

Tableau F.7. Évaluations qualitatives du graphe résumé relatif au sujet Finance du 7 décembre

	Le graphe résume-t-il le sujet abordé ? Pourquoi ?	Les concepts du graphe sont-ils pertinents ? Pourquoi ?	Les relations du graphe sont-elles pertinentes ? Pourquoi ?
Réponse de l'évaluateur 1	Plus ou moins : il y a de l'information pertinente, mais comme pour le premier graphe, il manque du contexte pour bien comprendre l'information.	Oui, sauf certains (comme "reality check today" ou "today oil")	Oui, sauf certaines ("canadian of balancing budget" ?)
Réponse de l'évaluateur 2	Le graphe ne résume qu'une partie du sujet, il reste encore de l'information à extraire	oui les concepts sont pertinents	Les relations sont pertinentes en général, mais il y en a quelques-uns qui ne le sont pas comme budget will show canadian

Tableau F.8. Évaluations qualitatives du graphe résumé relatif au sujet National Defense du 7 décembre

	Le graphe résume-t-il le sujet abordé ? Pourquoi ?	Les concepts du graphe sont-ils pertinents ? Pourquoi ?	Les relations du graphe sont-elles pertinentes ? Pourquoi ?
Réponse de l'évaluateur 1	Mieux que les deux précédents : beaucoup d'information est présente, mais les sous-graphes ne sont pas connexes (il n'y a pas de lien visible entre "air strike" et le graphe "principal")	Oui, malgré quelques erreurs (comme "last night president obama")	Certaines sont inutiles (liberal government told canadian : il manque quelque chose), mais globalement oui
Réponse de l'évaluateur 2	a mon avis le graphe résume bien le sujet abordé	il y a quelques concepts qui sont inutiles comme way , et il y aussi un manque, comme obama a mentionné germany mais il y aussi france et le UK je ne sais pas si c'est important à extraire ou pas, mais si germany est pertinent alors les autres le sont aussi	n générale elles sont pertinentes, mais comme way is appropriate for canada ce n'est pas pertinent je crois

Tableau F.9. Évaluations qualitatives du graphe résumé relatif au sujet Taxation du 8 mai

	Le graphe résume-t-il le sujet abordé ? Pourquoi ?	Les concepts du graphe sont-ils pertinents ? Pourquoi ?	Les relations du graphe sont-elles pertinentes ? Pourquoi ?
Réponse de l'évaluateur 1	Non : beaucoup trop de sous-graphes non reliés	la plupart sont pertinents, mais mal reliés	Beaucoup de relations qui ne sont pas fausses, mais n'apportent rien (liberal announced with great fanfare, commitment to create fairness)
Réponse de l'évaluateur 2	le graphe résume une partie du texte comme les autres graphe mais un concepts comme TFSA qui a été mentionné 4 fois dans le texte je crois qu'il est important d'extraire une information qui prouve que ce concept a été bien mentionné et discuté dans la discussion	les concepts et les relations qui ont été détectés sont bien pertinents je pense	

Tableau F.10. Évaluations qualitatives du graphe résumé relatif au sujet Infrastructure du 8 juin

	Le graphe résume-t-il le sujet abordé ? Pourquoi ?	Les concepts du graphe sont-ils pertinents ? Pourquoi ?	Les relations du graphe sont-elles pertinentes ? Pourquoi ?
Réponse de l'évaluateur 1	Plus ou moins : beaucoup d'informations pertinentes, mais pas suffisamment pour parler de "résumé". Un bon début de résumé	Oui, malgré quelques coquilles ("touch", venant de l'expression "out of touch")	Oui, mais certaines sont trop longues : "will invest an additional 1c a litre from", ou incohérentes (exemple précédent : "is completely out of")
Réponse de l'évaluateur 2	e pense que le graphe ne résume pas trop bien le sujet, dans le texte ils font mention des rôles des maires, ainsi que plusieurs villes qui ont été mentionnées dans le texte n'existent pas dans le graphe	oui les concepts qui ont été extraites sont pertinents en générale, mais quelques uns sont inutile comme touch,	oui les relations sont pertinentes

Tableau F.11. Évaluations qualitatives du graphe résumé relatif au sujet Citizen Voting Act du 1 mai

	Le graphe résume-t-il le sujet abordé ? Pourquoi ?	Les concepts du graphe sont-ils pertinents ? Pourquoi ?	Les relations du graphe sont-elles pertinentes ? Pourquoi ?
Réponse de l'évaluateur 1	Oui, la plupart des concepts pertinents sont là, cependant certains triplets pourraient être omis, ex: Canada -> is -> lower voter turnout Le graph est un peu mélangeant.	La plupart des concepts pertinents sont là	Certaines relations sont pertinentes, d'autres ne le sont pas, ex: right -> to vote in -> quebec; le concept right est lié à canadian (canadian -> have> right). C'est faux que les canadiens ont le droit de vote au québec, bien qu'il est question du vote pour le gouv fédéral De manière générale, les relations sont pertinentes.
Réponse de l'évaluateur 2	Moyennement. Je m'attendais à voir au moins l'ID du bill dans les concepts. On peu tout de même avoir une idée du sujet (droit au vote, canadiens non-résidant, élections, ...)	La majorité oui. Certain sont très généraux (new way, thing), et j'ai l'impression que certains sont manquants.	En majorité. D'autres comme "Canada->is->low voter turnout" ou "Canada->are->citizen of other country" n'ont pas de sens à mes yeux. Et je me serais attendu à voir des relation par rapport au "bill" en question.

Tableau F.12. Évaluations qualitatives du graphe résumé relatif au sujet Respecting Families of Murdered and Brutalized Persons Act du 8 2 juin

	Le graphe résume-t-il le sujet abordé ? Pourquoi ?	Les concepts du graphe sont-ils pertinents ? Pourquoi ?	Les relations du graphe sont-elles pertinentes ? Pourquoi ?
Réponse de l'évaluateur 1	Oui, les idée principales du sujet sont quasiment toutes illustrées par le graph	Oui, pour la plupart. certains ne le sont pas, ex symbol	Oui, pour la plupart, elles visent le bon objets
Réponse de l'évaluateur 2	Oui. Les concepts et relations sont pertinentes. On peut savoir qu'on parle d'au moins un projet de loi. Mais il serait dure de savoir vraiment ce que ce projet de loi contient.	Oui. Les concepts sont tous relativement importants au sujet.	Oui. Elles sont importantes dans par rapport au sujet, mais plus de relations auraient d'avantage aidé.

Tableau F.13. Évaluations qualitatives du graphe résumé relatif au sujet Aboriginal Affairs du 3 juin

	Le graphe résume-t-il le sujet abordé ? Pourquoi ?	Les concepts du graphe sont-ils pertinents ? Pourquoi ?	Les relations du graphe sont-elles pertinentes ? Pourquoi ?
Réponse de l'évaluateur 1	Non, le graphes concerne plusieurs affirmations isolée parfois impertinentes, il ne renvoie pas une image globale juste du texte	Certain oui, d'autres non, eX: disgrace, average, close, sad chapter	idem, ex: non, is one of , recieve an average of...
Réponse de l'évaluateur 2	Moyennement. On peu avoir une idée du sujet avec des conecepts comme "first nation student" ou "reserve" mais sans plus	En général oui, mais certains concept principaux répétés semblent manquant dans le texte semblent manquant (Truth and Reconciliation comittee report)	And grande partie oui, car elles résumant certain éléments du texte, mais il semble en manquer d'autres au moin autant importantes.

Tableau F.14. Évaluations qualitatives du graphe résumé relatif au sujet The economy du 3 juin

	Le graphe résume-t-il le sujet abordé ? Pourquoi ?	Les concepts du graphe sont-ils pertinents ? Pourquoi ?	Les relations du graphe sont-elles pertinentes ? Pourquoi ?
Réponse de l'évaluateur 1	Non, il manque de l'info	Non, ex: island	Non, ex : according to (the speaker -> according to > international labour certaines relations sont trop longue et contiennent leur objet ex: downgraded its forecasts for (downgraded what? its forecasts)
Réponse de l'évaluateur 2	ssez bien. On a une bonne idée du sujet. Des relations peuvent aussi indiquer certaines positions face au débat (les relation partant de "OECD") se contredisent.	Oui. À l'exception de "the speaker", les autres concepts sont des éléments important du débat.	La quasi-totalité des relations sont pertinentes et ont un lien direct avec le débat, touchant des points importants. Les relations contradictoires donnent des bonne idées de différentes positions.

Tableau F.15. Évaluations qualitatives du graphe résumé relatif au sujet National Defence du 4 mai

	Le graphe résume-t-il le sujet abordé ? Pourquoi ?	Les concepts du graphe sont-ils pertinents ? Pourquoi ?	Les relations du graphe sont-elles pertinentes ? Pourquoi ?
Réponse de l'évaluateur 1	L'idée générale est résumée, toutefois, certains concepts sont imprécis, exemple: la relation canadian -> is very disrespectful of -> sergeant doiron; le sujet devrait être canadian government	La plupart des concepts pertinents sont là, i.e minister of national defence, sergeant doiron, allegation of inappropriate conduct, etc.	Oui, les relations retenues sont correctement extraite et désignent précisément les relations entre les concepts
Réponse de l'évaluateur 2	Moyennement. Le seul graphe peut donner une idée du sujet très général traité, mais en dit peu sur le/les débats précis. Peut-être car les concepts donne un idée, mais les relations ne représentent pas beaucoup le débat sur le traitement des prisonniers ou l'enquête sur la mort du sergeant Doiron.	Oui, il en manque, mais ceux présents sont pertinent au débat.	Peu. La plus intéressante touche le "lack of transparency" du gouvernement. Les phrases de politiciens répétées qui n'apportent rien au débat sont malheureusement représentées dans la sortie du système. (Veteran->served->afghanistan / government-> takes->allegation of inappropriate conduct)