UNIVERSITÉ DU QUÉBEC À MONTRÉAL

# TRAITEMENT DU LANGAGE ET SYSTÈMES DE DIALOGUE POUR FACILITER L'ACCÈS À LA JUSTICE

MÉMOIRE
PRÉSENTÉ
COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN INFORMATIQUE

PAR
MARC QUEUDOT

MAI 2020

TABLE DES MATIÈRES

# LISTE DES TABLEAUX

LISTE DES FIGURES

# RÉSUMÉ

En moyenne, un Canadien sur trois sera touché par un problème juridique au cours d'une période de trois ans. Pourtant, tout le monde n'a pas le même accès au droit : qu'on parle de représentation ou de conseil juridique, les coûts très élevés de ces services excluent les personnes défavorisées et les plus vulnérables, leur imposant de se représenter elles-mêmes. Pour ces personnes , l'accès à l'information juridique est donc un outil essentiel. En effet, sans remédier au problème de représentation et de conseil, accéder à l'information permet de limiter les différences de résultats dans les cas les plus simples qui sont aussi les plus nombreux.

Ce travail présente trois réalisations qui visent à faciliter l'accès à l'information juridique. Tout d'abord, les problèmes liées aux données sont examinés. La difficulté d'accéder à des textes de jugements en grande quantité rend complexe la mise en place de systèmes à base d'apprentissage automatique de la jurisprudence. De plus, les données disponibles, peu structurées, sont mal adaptées à un traitement automatisé. Pour pallier à ces problèmes dans le cadre de nos travaux, nous avons collecté, indexé et annoté en partie des jugements rendus par la Cour Fédérale du Canada.

Nous avons ensuite conçu deux robots conversationnels fournissant de l'information juridique à partir de bases de connaissances. Le premier, basé sur les données du gouvernement Canada, traite de sujets d'immigration. Le second, créé pour la Banque Nationale du Canada, renseigne les employés sur des points de droit.

Les expériences menées utilisent plusieurs types de représentation et d'algorithmes de classification pour la conception des systèmes de dialogue, dont des représentations syntaxiques distribuées et un modèle de langue créé par un réseau de neurones profond avec le mécanisme d'attention. Ces deux approches sont également comparées à une technique inspirée du domaine de la recherche d'information qui donne de bons résultats sur le jeu de données de la Banque Nationale. Les annotations du corpus de la Cour Fédérale et le robot conversationnel dédié aux questions d'immigration sont partagées en accès libre.

**MOTS-CLÉS :** Apprentissage automatique, accès à la justice, recherche d'information, robot conversationnel, système de dialogue, traitement automatique du langage naturel.

# INTRODUCTION

La non-représentation d'une forte proportion de justiciables est l'un des facteurs les plus frappants des problèmes d'accès à la justice.

Laniel *et al.* (2018) traite de ce problème et souligne le côté involontaire de cette situation pour beaucoup de justiciables, alors que les juges traitent l'"auto-représentation" comme un choix, voire un privilège. Pour refléter le fait que cette situation n'est souvent pas un choix, Laniel *et al.* préfèrent parler de Justiciables Non-Représentés (JNR).

Schneider (2010) décrit la situation des États-Unis, où la Cour Suprême a statué que les demandes soumises par des JNR (*pro se complaints*) devraient être soumises à un niveau d'exigence moins élevé que des requêtes formulées par des avocats. D'après Schneider, les critères de validité des plaintes qu'utilisent les cours désavantagent encore davantage les JNR. En effet, la responsabilité de montrer que la plainte est crédible revient au plaignant, ce qui est difficile pour les JNR qui manquent typiquement de ressources. Ce décalage est encore plus important lorsqu'une recherche préliminaire est nécessaire, ou que l'opposition détient l'intégralité des éléments clés du procès. D'après Schneider, les contraintes de constitution de dossier et l'étude des plaidoiries mettent en péril le droit à se représenter soi-même devant les tribunaux, et même le droit constitutionnel des citoyens à une opportunité d'être entendu.

Au-delà des coûts de consultation et de représentation élevés qui freinent la représentation des justiciables, d'autres obstacles rendent difficile l'accès à la justice tel l'accès à l'information juridique, mais aussi les contraintes géographiques, etc. La

contribution de cette maîtrise sous la forme d'un robot conversationnel d'immigration pourrait faciliter l'accès à l'information des justiciables dans ce domaine. Les travaux présentés dans ce mémoire ont été réalisés dans le cadre du projet LegalIA [1]. Ce projet regroupe plus d'une douzaine de chercheurs de l'Université du Québec à Montréal (UQAM), de Condordia et de l'Université de Montréal (UdeM) autour de la question du développement éthique et responsable de l'Intelligence Artificielle en droit. Ce projet bénéficie depuis 2017 d'un financement *Audace* des Fonds de Recherche du Québec (FRQ), réservé aux projets «haut risque/haut rendement»interdisciplinaires qui ne sont typiquement pas admissibles aux programmes de financements classiques des trois fonds de recherche.

Dans le cadre de la demande de financement, nous avons réalisé une vidéo [2] pour motiver le besoin derrière ce projet. La première partie de la proposition consiste à créer un robot conversationnel pour assister les acteurs du domaine juridique : avocats, juges et justiciables. Cet outil constituera ensuite un cas d'étude pour les analyses sociologiques et éthiques dans la seconde phase du projet. En pratique, la contribution de cette maitrise s'approche de la partie à destination du justiciable, décrite dans la vidéo. Ce mémoire ne décrit pas explicitement le projet de recommandations de ressources juridiques, mais un projet annexe dans lequel nous avons participé aborde cette problématique et sera mentionné brièvement.

Il existe aussi tout un écosystème d'autres initiatives qui traitent des facettes différentes du problème d'accès à la justice. Certaines associations proposent de l'aide juridique gratuite aux plus démunis. À Montréal, on trouve par exemple la clinique Droits Devant [3] qui accompagne les itinérants et les aide à protéger leurs

---

1. `https://legalia.uqam.ca/`

2. `https://legalia.uqam.ca/video/LegalIA.mp4`

3. `http://www.cliniquedroitsdevant.org/wordpress`

droits, notamment lors de procédures pénales. Dans plusieurs universités, les étudiants ont accès à des conseils juridiques gratuits (c'est le cas à l'Université de Montréal et à l'UQAM notamment), fournis par des étudiants des facultés de droit supervisés par des professionnels. Depuis Mars 2018, le site web *La boussole juridique*[4] recense les organismes qui proposent des services juridiques gratuitement ou à moindre coût.

Au Québec, *Éducaloi*[5] rassemble des experts de nombreux domaines juridiques et rédige des guides concis pour informer les citoyens sur leurs droits. Ils facilitent donc l'accès à l'information juridique par la vulgarisation et l'organisation de l'information en éléments plus faciles à consommer et qui correspondent mieux à ce que les gens recherchent que ne le ferait la seule vulgarisation de chaque texte de loi. C'est dans ce contexte, et en particulier considérant les problèmes d'accès à l'information juridique que notre travail s'insère.

***Question de recherche***

La question de recherche à l'origine de ce travail est *comment faciliter l'accès à la justice au Canada en utilisant des technologies à base d'Intelligence Artificielle (IA)*. Nous raffinerons cet objectif très large par en cherchant à faciliter la recherche en justice prédictive et en fournissant des outils d'accès à l'information juridique. Nous avons donc fabriqué un jeu de données de décisions de justice pour la conception de modèles prédictifs et créé deux robots conversationnels informationnels, l'un dédié à des sujets liés à l'immigration et l'autre au droit dans le domaine bancaire.

---

4. `http://boussolejuridique.ca/`

5. `https://www.educaloi.qc.ca/`

Dans un premier temps, nous présenterons dans le chapitre 1 les travaux récents dans les deux domaines qui nous intéressent : la justice prédictive et les robots conversationnels. Ensuite, nous expliquerons dans le chapitre 3 la méthodologie que nous avons suivie dans ce travail. Nous décrirons dans le chapitre 4 les données utilisées, ainsi que les problématiques que nous avons rencontrées, en particulier pour leur collecte. Le chapitre 5 contiendra les expériences réalisées ainsi que leur analyse. Les contributions présentées ici sont duales : nous décrivons dans un premier temps la construction d'un corpus dédié à la classification de décisions juridiques, puis la conception de deux robots conversationnels. Enfin, le dernier chapitre permettra de faire un bilan des travaux réalisés et de conclure ce mémoire.

Avant de présenter l'état de l'art, nous allons donner un aperçu des 6 publications et des autres activités réalisées dans le cadre de cette maîtrise pour mieux contextualiser notre travail.

*Publications dans le cadre de la maîtrise*

Dans le cadre de cette maîtrise, j'ai eu l'occasion de contribuer à plusieurs projets de recherche liés directement ou indirectement au sujet de ce mémoire. La majorité des travaux repose sur des algorithmes de classification de texte. Deux autres contributions s'intéressent à la séparation de graphes. Les six articles que j'ai rédigés ou auxquels j'ai contribué sont détaillés ci-après :
Dans

> Queudot, M. et Meurs, M.-J. (2018). Artificial Intelligence and Predictive Justice: Limitations and Perspectives. Dans *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, 889–897. Springer

nous décrivons les principales difficultés qui font obstacle au développement de

systèmes d'Intelligence Artificielle dans le domaine du droit, avec une attention particulière portée au sujet de la justice prédictive.

Les articles suivants concernent des développements récents en Traitement Automatique du Language Naturel (*Natural Language Processing* en anglais) (TALN) qui ont inspiré les travaux ce mémoire mais ne les ont pas impacté directement.

La contribution décrite dans

> Maupomé, D., Queudot, M. et Meurs, M.-J. (2019). Inter and Intra Document Attention for Depression Risk Assessment. Dans *Canadian Conference on Artificial Intelligence*, 333–341. Springer

s'applique à la détection précoce de la dépression dans les productions écrites d'internautes. L'accent est mis sur la quantité de données nécessaire à la détection et nous explorons l'impact de différents mécanismes d'attention sur la qualité de représentation des utilisateurs et sur la performance prédictive.

L'article

> Almeida, H., Queudot, M., Kosseim, L. et Meurs, M.-J. (2017). Supervised Methods to Support Online Scientific Data Triage. Dans *International Conference on E-Technologies*, 213–221. Springer

décrit trois approches pour effectuer une première sélection de documents parmi de grands corpus à l'étude par des experts de différents domaines. Le premier système s'applique à des résumés de textes scientifiques pour la recherche d'enzymes fongiques, le second à des articles pour la réalisation d'études systématiques en médecine et le dernier s'intéresse à la modération de contenus en ligne.

Dans

Sarfati, M., Queudot, M., Mancel, C. et Meurs, M.-J. (2017b). Knowledge Discovery in Graphs Through Vertex Separation. Dans *Canadian Conference on Artificial Intelligence*, 203–214. Springer

on applique le nouvel algorithme proposé par Sarfati *et al.* (2017a) à la découverte de connaissance sur un corpus sous forme de graphes et introduit une interface graphique pour manipuler les données et leur transformation.

L'article

Sarfati, M., Queudot, M., Mancel, C. et Meurs, M.-J. (2017a). Formulation relaxée de la séparation équilibrée d'un graphe. Dans *ROADEF 2017, 18ème congrès de la société française de recherche opérationnelle et d'aide à la décision*

présente une généralisation d'un algorithme de séparation de graphes introduit par Mohamed Didi Biha and Marie-Jean Meurs (2011) suivant les travaux de Balas et de Souza (2005). L'approche proposée permet de contrôler la priorité entre la taille relative des graphes résultants et le nombre de nœuds «perdus»dans la séparation pour créer des composantes connexes.

L'article

Almeida, H., Queudot, M. et Meurs, M.-J. (2016). Automatic Triage of Mental Health Online Forum Posts: CLPsych 2016 System Description. Dans *Third Workshop on Computational Linguistics and Clinical Psychology at NAACL HLT*, 183–187

décrit le système développé dans le cadre de la campagne d'évaluation CLPsych 2016 dont le but était de détecter à partir de leurs posts sur un forum les utilisateurs présentant un risque de dépression.

*Autres activités pertinentes*

Le sujet de la justice prédictive et des obstacles à sa mise en pratique au Canada (voir Chapitre 2) a été abordé dans les présentations que j'ai eu l'occasion de donner pendant la maîtrise : la journée IA [6] du département des Sciences de l'UQAM en mars 2019, deux formations à destination des membres du barreau de Montréal en avril 2018 [7] et le congrès de l'ACFAS [8] en mai 2019.

J'ai également participé à deux projets nés de *hackathons* à l'UQAM. À l'issue du premier, avec une équipe d'étudiants en sciences politiques, en droit et en informatique nous avons participé au symposium étudiant sur l'Intelligence Artificielle et les droits humains, d'Affaires Mondiales Canada [9]. Dans le cadre du second, le *Global Legal Hackathon*, nous avons travaillé sur le projet *Juridico* [10] pendant plusieurs mois. L'idée consistait à proposer une interface conversationnelle aux justiciables pour les orienter vers les ressources juridiques les plus adaptées : associations, bureaux d'aide juridiques ou cabinets d'avocats. Ces derniers auraient pu financer le système en contrepartie d'un afflux de prospects et de la première collecte de renseignements. Ce projet a fait partie des 14 équipes finalistes parmi plus de 600 équipes participantes à travers le monde et après deux phases de sélections préalables. En mars 2018, la Société Québécoise d'Information Juridique (SOQUIJ) a lancé une nouvelle version de son service «*la boussole juridique*» [11]

---

6. https://www.actualites.uqam.ca/2019/intelligence-artificielle-les-expertises-UQAM

7. Programme des formations IA et droit du barreau, avril/mai 2018

8. https://www.acfas.ca/evenements/congres/programme/87/600/610/c

9. Page de l'évement du symposium : https://www.cifar.ca/events/student-symposium-on-artificial-intelligence-human-rights

10. https://fspd.uqam.ca/nouvelle/juridico-invite-finale-mondiale-global-legal-hackathon-a-new-york/

11. https://www.newswire.ca/fr/news-releases/a-new-version-of-pro-bono-quebe

qui couvre maintenant la majeure partie du besoin que nous avions identifié.

---

cs-boussole-juridique---modernized-platform-facilitates-search-for-free-or-low-cost-legal-resources-678314323.html

# CHAPITRE I

# DÉFINITIONS ET ÉTAT DE L'ART

Nous allons présenter dans ce chapitre les travaux importants et les meilleures approches récentes dans les domaines auxquels nous nous intéressons. Pour être capable de comprendre les approches employées par ces systèmes et leurs limites, il est important de connaître les métriques utilisées pour les évaluer. Nous allons donc commencer par les définir. Ensuite, nous introduirons les approches de justice prédictive, et nous terminerons avec les systèmes de dialogue. Pour plus de détails sur les techniques présentées, nous conseillons ces deux manuels de référence : (Manning *et al.*, 1999) pour le TALN et (Manning *et al.*, 2008) pour la RI.

## 1.1 Définition des métriques

En Recherche d'Information (RI) et en classification, on utilise souvent les métriques suivantes définies ici pour la classification binaire. Pour ces métriques, on peut calculer les résultats par classes et les agréger par la suite si on a plus de deux classes.

On appelle Vrai Positifs (VP) les instances de la classe positive, correctement classifiées. On définit de la même manière Vrai Négatifs (VN), Faux Positifs (FP) et Faux Négatifs (FN).

La **précision P** est la proportion d'instances réellement positives parmi celles identifiées comme étant positives par le système : $\mathrm{P} = \frac{VP}{VP+FP}$.

Le **rappel R** est la proportion d'instances réellement positives qui ont été correctement classifiées par le système : $\mathrm{R} = \frac{VP}{VP+FN}$.

L'**accuracy A** est la proportion des instances, positives comme négatives, classées correctement, soit $\mathrm{A} = \frac{VP+VN}{VP+VN+FP+FN}$.

On peut l'utiliser si les classes sont équilibrées, mais on obtiendra des résultats biaisés vers la classe majoritaire si ce n'est pas le cas. Par exemple, avec une répartition 90/10 d'instances entre la classe positive et la classe négative, la règle "assigne toujours à la classe positive" obtiendrait 90% d'accuracy.

La **F-mesure $\mathbf{F_m}$** est la moyenne harmonique de la précision et du rappel : $F_m = 2\frac{\text{précision}\times\text{rappel}}{\text{précision}+\text{rappel}}$.
Cette métrique est le meilleur choix si les classes ne sont pas équilibrées.

Chacune de ces métriques est définie dans le cas d'une classification binaire, et elles doivent être aggrégées pour évaluer un cas multi-classes. On utilise 3 moyennes différentes : les moyennes *micro, macro et pondérée.*

Prenons l'exemple d'une classification à 3 classes : A, B et C où il ne peut y avoir qu'une seule classe pour chaque exemple.

Dans le cas de la **micro-moyenne**, on évalue toutes les classes ensemble. Pour la **micro-précision** par exemple, on compte le total de VP et FP en s'aidant de la matrice de confusion ddu tableau 1.1 : $VP = VP_A + VP_B + VP_C = 3+5+6 = 14$, $FP = FP_A + FP_B + FP_C = 4 + 2 + 7 = 13$, puis on calcule la précision : micro-précision $= \frac{\mathrm{VP}}{\mathrm{VP+FP}} = \frac{14}{14+13} \sim 0,52$.

Pour la **macro-précision**, la précision pour chaque classe est calculée (qu'on

Tableau 1.1 Matrice de confusion pour un problème de classification à 3 classes

| | | prédiction | | |
|---|---|---|---|---|
| | | A | B | C |
| vraie classe | A | 3 | 1 | 1 |
| | B | 4 | 5 | 1 |
| | C | 0 | 1 | 6 |

note $p_A$ à $p_C$), puis on effectue la moyenne. D'après le même exemple, on a $p_A = \frac{\text{VP}_A}{\text{VP}_A+\text{FP}_A} = \frac{3}{7}$, idem pour précision$_B$ et précision$_C$. La macro-précision vaut donc $(p_A + p_B + p_C)/3 == (\frac{3}{7} + \frac{4}{7} + \frac{3}{4})/3 = \frac{7}{12} \sim 0,58$.

La moyenne pondérée assigne un poids à chaque classe égal au nombre d'instances qui y sont associées. Dans notre cas, ce sont respectivement 5, 10 et 7 (la colonne support de la matrice). La précision pondérée vaut donc $(p_A*5+p_B*10+p_C*7)/3$ $= (\frac{3*5}{7} + \frac{4*10}{7} + \frac{3*7}{4}/19 =\sim 0,69$

La figure 1.1 [1] illustre quelle partie du jeu de données la précision et le rappel désignent.

La **précision@k** est une variante de la précision, spécifiques aux systèmes à base de RI qui retournent des résultats ordonnés. Plutôt que de considérer tous les résultats pour l'évaluation, on choisit de ne considérer que les $k$ documents les plus hauts dans le classement, ce qui est plus pertinent quand c'est un utilisateur qui doit parcourir ces résultats manuellement. La precision@1 est le cas particulier où $k = 1$, donc on ne considère que le premier document.

---

1. Contribution d'un auteur de wikipedia `https://en.wikipedia.org/wiki/Precision_and_recall#/media/File:Precisionrecall.svg`, visitée le 2 janvier 2020, traduite et ajustée pour un format paysage.

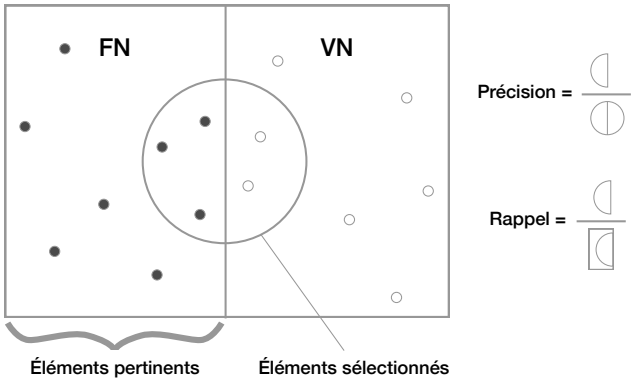



Figure 1.1 Représentation graphique de la précision et du rappel

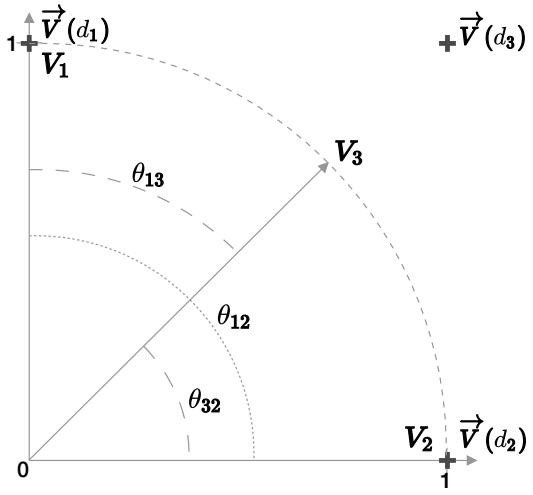



Figure 1.2 Similarité cosinus dans un espace en deux dimensions

Notation

Dans l'équation ci-dessus et le reste de ce document, une variable en gras désigne un vecteur : $\mathbf{v}$ par exemple. La notation $\mathbf{v}_{\langle t \rangle}$ est utilisée pour donner un indice au vecteur $\mathbf{v}$, par opposition à $\mathbf{v}_i$ qui désigne le $i^e$ élément de $\mathbf{v}$.

La **similarité cosinus** entre deux vecteurs permet de comparer deux vecteurs entre eux. Elle est égale à leur produit scalaire divisé par le produit de leurs normes euclidiennes. Avec deux vecteurs $\mathbf{v_1}$ et $\mathbf{v_2}$, on a donc :

$$\text{sim}(\mathbf{v_1}, \mathbf{v_2}) = \cos(\theta) = \frac{\mathbf{v_1} \cdot \mathbf{v_2}}{|\mathbf{v_1}| \times |\mathbf{v_2}|} \tag{1.1}$$

Soit 3 vecteurs à composantes binaires qui représentent des documents $d_1$ à $d_3$ : $\mathbf{d_1} = [1, 0]$, $\mathbf{d_2} = [0, 1]$ et $\mathbf{d_3} = [1, 1]$.

On représente, sur la figure 1.2, $\overrightarrow{V}(d_1)$ à $\overrightarrow{V}(d_3)$ et leur version normalisée

$$\mathbf{v_x} = \frac{\overrightarrow{V}(d_x)}{|\overrightarrow{V}(d_x)|}, \mathbf{x} \in \{1, 2, 3\}$$

On a donc

$$\text{sim}(\mathbf{v_1}, \mathbf{v_3}) = \frac{\mathbf{v_1} \cdot \mathbf{v_3}}{|\mathbf{v_1}||\mathbf{v_3}|} = \frac{[1, 0] \cdot [1, 1]}{|[1, 0]||[1, 1]|} = \frac{1}{2}$$

$$\text{sim}(\mathbf{v_1}, \mathbf{v_2}) = \frac{\mathbf{v_1} \cdot \mathbf{v_2}}{|\mathbf{v_1}||\mathbf{v_2}|} = \frac{[1, 0] \cdot [0, 1]}{|[1, 0]||[0, 1]|} = \frac{0}{1} = 0$$

$$\text{sim}(\mathbf{v_3}, \mathbf{v_2}) = \frac{\mathbf{v_3} \cdot \mathbf{v_2}}{|\mathbf{v_3}||\mathbf{v_2}|} = \frac{[1, 1] \cdot [0, 1]}{|[1, 1]||[0, 1]|} = \frac{1}{2}$$

Sur cette figure en deux dimensions, il aurait été facile de trouver les valeurs de similarité cosinus, par exemple $\text{sim}(\mathbf{v_1}, \mathbf{v_3}) = \cos(\theta_{32}) = \cos(\pi/4) = 1/2$.

## 1.2 Représentation et classification de texte

L'une des sources principales d'amélioration des performances en TALN est liée à la manière dont le texte est représenté pour pouvoir être traité par les différents algorithmes de classification.

Pour comprendre l'intérêt des techniques actuelles, il est important de les replacer dans leur contexte. Cette section décrit les principales méthodes utilisées à travers le temps. Nous verrons plus tard que certaines d'entre elles, bien que développées il y a plusieurs décennies, permettent encore de concevoir dans certaines situations des modèles compétitifs face aux techniques récentes.

### 1.2.1 Sacs de mots et variantes

La manière la plus naïve de représenter un document textuel est probablement de l'identifier à l'ensemble non ordonné de mots qui le composent. On appelle cette technique le «sac de mots»(ou *bag-of-words* en anglais), pour souligner que les mots y sont stockés dans le désordre. La représentation d'un document en particulier est dépendante d'un dictionnaire de termes qu'on construit généralement sur l'ensemble du corpus dont on dispose. On peut ensuite construire, pour chaque document, un vecteur de mots dont les coordonnées seront 0 ou 1 en fonction de la présence ou de l'absence de chacun des mots du dictionnaire. On peut aussi remplacer la variable binaire par le nombre d'occurrences de chaque mot. Dans les deux cas, cela conduit à la production de vecteurs qu'on appelle «clairsemés», parce qu'ils sont principalement constitués de zéros.

Supposons qu'on ait les trois documents suivants :
$A$ : «`Pour demander un visa étudiant, il vous faut un CAQ.`»,
$B$ : «`Comment avoir un visa étudiant ?`»,

Tableau 1.2 Exemple de représentation par sac de mots avec un petit dictionnaire

| | vous | il | un | travailler | alors | CAQ | demander | est | visa | avoir | étudiant | ce | comment | que | je | suis | peux | faut | pour |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| B | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| C | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 |

$C$ : «`Est-ce que je peux travailler alors que je suis étudiant ?`».
Le tableau 1.2 est la représentation en sac de mots de ces phrases, lorsque ces documents sont les trois seuls utilisés pour constituer le dictionnaire. Avec un nombre raisonnable de documents ne traitant pas exactement des mêmes sujets, le nombre de termes du dictionnaire grandit rapidement, alors que le nombre de «1»dans chaque vecteur ne change presque pas. Si le dictionnaire compte 10 000 mots par exemple, on compte seulement 15-20 valeurs non nulles parmi chaque vecteur de taille 10 000, pour des documents de la taille d'une phrase moyenne.

La grande taille de ces représentations implique un coût de calcul élevé pour de nombreux algorithmes, mais le fait que les vecteurs soient clairsemés va permettre une amélioration de ce coût. Ces avancées seront présentés à la section 1.2.2.

On peut ensuite utiliser ces représentations pour n'importe quelle tâche de TALN, en entrée d'algorithmes de classification comme on le décrira plus tard, ou pour comparer directement les documents entre eux, en utilisant par exemple la similarité cosinus que nous avons décrite à la section 1.1.

Un inconvénient majeur de ces représentations est qu'elles donnent un poids identique à tous les mots. Elles ne prennent en compte ni la longueur du document ni le fait que certains mots sont présents dans l'immense majorité des documents.

En effet, un document très long a une grande probabilité de contenir un mot quelconque, mais celui-ci a un impact moins fort sur le sujet du document que dans un document plus court. De la même manière, les mots très fréquents (les pronoms par exemple) n'apportent que peu d'information sur le sujet d'un document qui les utilisent.

La technique appelée *Term Frequency-Inverse Document Frequency (TF-IDF)* calcule un score pour chaque mot qui prend en compte les deux points que nous venons de décrire (longueur du document et fréquence des mots). La fréquence du terme dans le document (TF) pondère le nombre d'occurrences de ce terme par la longueur du document. Cette composante est décrite par l'équation 1.2, dans laquelle on note $f_{t,d}$ le nombre d'occurrences du terme $t$ dans le document $d$.

$$TF(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{1.2}$$

La composante IDF est une heuristique introduite par Karen Spärck Jones (1972) sur l'intuition qu'un terme qui apparaît dans de nombreux documents ne permet pas de bien discriminer ces documents entre eux. Ce score pour un terme $t_i$, qui apparait dans $n_i$ documents parmi $N$ documents, est calculé par l'équation 1.3.

$$IDF(t_i) = \log \frac{N}{n_i} \tag{1.3}$$

Chacune des trois phrases $I$ : «`J'ai obtenu mon visa de travail.`», $J$ : «`Je veux un visa pour étudier.`» et $K$ : «`As-tu mangé un repas de Noël ?`»aurait une représentation par sac de mots aussi différente des deux autres. En effet, le mot `visa` est commun à $I$ et $J$, le mot `un` à $J$ et $K$ et le mot `de` à $I$ et $K$. Puisque `de` et `un` sont des mots très communs comparativement à `visa`, le score associé à ce dernier serait plus élevé que pour les deux autre mots avec la représentation TF-IDF, et donc les documents $I$ et $J$ serait la paire la plus proches parmi les trois possibles.

De nombreux auteurs ont cherché à donner une justification théorique solide à l'intuition de Spärck Jones (1972), en se basant souvent sur la théorie de l'information de Shannon et Weaver (1998). D'après Robertson (2004), on peut trouver de nombreuses limites à ces interprétations a posteriori. Malgré tout, Robertson reconnait l'impact que cette technique a eu sur le domaine du TALN, en s'implantant sous une forme ou une autre dans la majorité des moteurs de recherche du monde, entre autres applications.

### 1.2.2 Représentations lexicales distribuées

Les représentations vectorielles produites par sac de mots ont l'avantage d'être simples à comprendre, et elles sont assez performantes, comme nous l'avons vu. Pourtant, elles ont aussi leurs limites.

Si on reprend les phrases d'exemples $A$, $B$ et $C$ de la section 1.2.1 en remplaçant le mot `visa` (de travail) dans le document $B$ par son synonyme `permis`, les documents $A$ et $B'$ : «`Je veux un permis pour étudier.`» n'ont plus aucun mot en commun.

En utilisant la similarité cosinus sur leur représentations par sac de mots ou tf-idf, ces documents serait donc considérés comme totalement différents (similarité nulle), alors que la similarité entre $A$ et $C$, ainsi qu'entre $B$ et $C$ est non-nulle. Pour la plupart des tâches de classification, obtenir des représentations similaires pour des mots sémantiquement proches serait une qualité intéressante.

Les représentations lexicales distribuées (*embeddings* en anglais) font partie des approches où les représentations vectorielles des mots sont apprises à partir de leur contexte, plutôt que définies explicitement comme nous l'avons décrit. L'hypothèse distributionnelle introduite par Harris (1954) suggère que les mots qui sont proches sémantiquement tendent à être utilisés dans le même contexte. Alors que

l'analyse sémantique latente représente le texte en l'identifiant aux «sujets»auquel il se rapporte (Landauer et Dumais, 1997), les représentations sémantiques distribuées utilisent les mots comme contexte. Mikolov *et al.* (2013a) introduit deux techniques pour apprendre leurs représentations à partir de leur entourage.

Même si ces deux techniques ne sont pas les premières à permettre l'apprentissage de représentations distribuées des mots, la structure allégée du modèle utilisé permet pour la première fois leur entraînement sur des jeux de données de très grande taille avec peu de ressources. D'après Mikolov *et al.* (2013b), des représentations pour 100 milliards de mots peuvent être apprises en une journée de calcul sur une seule machine. Mikolov *et al.* montrent aussi que ces représentations contiennent non seulement de l'information sémantique sur les mots, mais aussi des caractéristiques syntaxiques. Ce qui est surprenant est qu'il est possible, dans cet espace de représentation, d'effectuer des translations linéaires (des déplacements) qui correspondent à ces caractéristiques. En notant $\mathbf{v_m}$, $\mathbf{v_s}$, $\mathbf{v_f}$ et $\mathbf{v_p}$ les vecteurs représentant respectivement les mots «Madrid», «Spain», «France»et «Paris», Mikolov *et al.* montrent que la relation suivante est vraie (équation 1.4).

$$\mathbf{v_m} - \mathbf{v_s} + \mathbf{v_f} \sim \mathbf{v_p} \tag{1.4}$$

On peut voir sur la figure 1.3 que les représentations lexicales distribuées des villes sont proches entre elles, et idem pour les pays, mais surtout la position relative de «France »par rapport à «Paris »est similaire à celle de «Espagne »par rapport à «Madrid ».

Pour reprendre notre exemple, on obtient

$$\mathrm{sim}_{emb}(A, B') \sim \mathrm{sim}_{emb}(A, B) > \mathrm{sim}_{emb}(B, C) \tag{1.5}$$

si on note $\mathrm{sim}_{emb}(X, Y)$ la similarité cosinus entre deux représentations lexicales

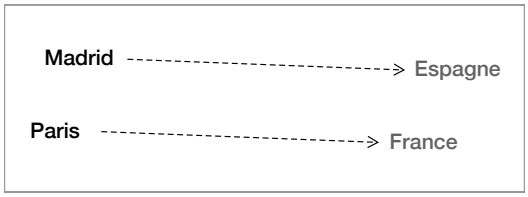



Figure 1.3 Représentation des représentations lexicales distribuées de villes et de pays

distribuées de documents $X$ et $Y$, puisque les synonymes `visa` et `permis` apparaissent souvent dans le même contexte.

### 1.2.3 Modèles de langue et réseaux récurrents

Alors que les représentations lexicales distribuées sont entraînées avec l'objectif explicite de conserver des informations sémantiques dans leurs représentations, les *modèles de langue* l'apprennent implicitement. En effet, l'objectif du modèle de langue est de capturer les distributions de probabilité de séquences de mots. La nature de la langue n'étant pas bien comprise et l'univers des séquences de mots possibles étant extrêmement grand, les modèles de langue utiliseront certaines simplifications pour approximer la probabilité d'une phrase. Par exemple, les modèles $n$-grammes maintiendront un compte des occurrences des séquences de mots de longueur $n$ ou moindre. Pour calculer la probabilité d'une séquence de longueur $m > n$, il suffira de multiplier la probabilité des séquences de $n$ mots qu'elle contient. Une fois entraînés, ces modèles peuvent être utilisés dans des applications plus concrètes, pour représenter les mots comme on le ferait en utilisant des représentations lexicales distribuées.

Les premiers modèles de langue étaient appris à l'aide de réseaux de neurones sans récurrence : Xu et Rudnicky (2000) et Bengio *et al.* (2003) par exemple.

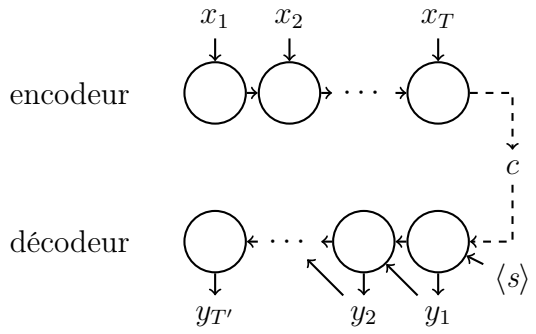



Figure 1.4 Encodeur-décodeur récurrent

Pour pouvoir gérer un contexte de taille variable, on utilise maintenant plutôt l'architecture encodeur-décodeur récurrent (Cho *et al.*, 2014) Le premier des deux réseaux, l'encodeur, est un Réseau de Neurones Récurrents (*Recurrent Neural Network* en anglais) (RNN) qui lit un par un les mots de la phrase à traduire et prédit le mot suivant à chaque étape. Une fois la phrase terminée, l'état caché contient un résumé de la phrase.

On utilise souvent le terme «prédiction»en apprentissage automatique, on parle aussi parfois de «décision». Ces termes désignent la réponse (la sortie) que produit un algorithme après avoir traité l'entrée qui lui a été fournie.

Plus formellement, on note $\mathbf{x} = (x_1, \ldots, x_T)$ une séquence de symboles (qui peuvent être des mots, des caractères), $h_t$ l'état caché à l'étape de temps $t$ et $f$, une fonction d'activation non linéaire. La valeur de l'état caché à l'instant $t$ est donnée par l'équation 1.6.

Une fonction d'activation linéaire peut prendre les valeurs 1 ou 0 (pour VRAI et FAUX par exemple) en fonction de la valeur de ses paramètres. On utilise plutôt des fonctions continues comme fonctions d'activations dans les réseaux

de neurones pour pouvoir trouver leur pente par dérivation (la technique qu'on appelle «descente de gradient»). Puisque ces fonctions ne peuvent pas être décrites par un polynôme de degré 1, on dit qu'elles sont non-linéaires.

$$h_t = f(h_{t-1}, x_t) \tag{1.6}$$

Ensuite, on utilise l'état caché du réseau et les valeurs des symboles précédents dans la séquence pour calculer la probabilité d'apparition d'un symbole $j$, représenté par le vecteur *one-hot* $\mathbf{w}_j$.

Un vecteur *one-hot* est un vecteur contenant une seule valeur différente de zéro. Pour $\mathbf{w}_j$, on parle aussi de vecteur «1 parmi $K$», $K$ étant la taille du vocabulaire.

Ce calcul de probabilité d'un symbole, étant donné les précédents est donné par la formule suivante.

$$p(x_{t,j}|x_{t-1}, \ldots, x_1) = \frac{\exp(\mathbf{w}_j \mathbf{h}_{\langle t \rangle})}{\sum_{j'=1}^{K} \exp(\mathbf{w}_{j'} \mathbf{h}_{\langle t \rangle})} \tag{1.7}$$

Le deuxième réseau, le décodeur, génère une séquence de symboles à partir de l'état caché $c$ du premier RNN et des caractères générés précédemment $y$. La mise à jour de l'état caché de ce deuxième réseau est donnée par le calcul suivant :

$$\mathbf{h}_t = f(\mathbf{h}_{\langle t-1 \rangle}, y_{t-1}, \mathbf{c}) \tag{1.8}$$

Ces deux étapes de l'encodeur-décodeur récurrent sont illustrées par la figure 1.4.

On peut voir qu'à chaque symbole lu, l'encodeur met à jour son état caché à partir de l'ancien et du symbole actuel. Une fois le dernier symbole lu, le réseau a en quelque sorte le contexte en mémoire. Chaque symbole est prédit dans le décodeur en utilisant le contexte, le symbole prédit précédemment et l'état caché actuel. Ce dernier est lui-même issu de l'état caché et du symbole de l'étape de temps précédente.

### 1.2.4 Transformeurs

La technique que nous venons de présenter est basée sur une variante de RNN qui utilise le mécanisme de récurrence pour modéliser l'ordre des mots/symboles dans le texte. Les *transformeurs* sont un nouveau type de réseaux de neurones qui utilisent le mécanisme d'attention à la place de la récurrence. Ce mécanisme introduit par Bahdanau *et al.* (2014), initialement pour la traduction de texte automatique, est devenu la source d'amélioration de performances pour de nombreuses tâches de référence en TALN.

L'idée générale consiste à prédire, à chaque étape de prédiction d'une séquence de texte, l'importance qu'a chacun des éléments passés en entrée du réseau. On dit que c'est l'*attention* que porte le réseau à des éléments particuliers de la séquence en entrée. Dans le cas de la traduction, une partie du réseau apprend la partie de traduction en soi, et l'autre apprend à «aligner»les mots ou groupes de mots.

Les transformeurs (Devlin *et al.*, 2018) utilisent des Réseaux Neuronaux Convolutifs (RNC), combinés avec le mécanisme d'attention. Cette technique permet d'éliminer totalement la dépendance séquentielle des réseaux récurrents, et donc de paralléliser entièrement leur entraînement. Bien que leur entraînement ait un coût de calcul global bien plus élevé que leur équivalent en réseaux récurrents par exemple, cet avantage les rend plus rapides à l'exécution, sous condition d'avoir

accès à de larges ressources de calcul parallèle.

Comme avec les représentations lexicales distribuées (voir section 1.2.2), ce qui a permis la popularisation des transformeurs est la mise à disposition publique de réseaux pré-entraînés. Ceux-ci peuvent être utilisés tels quels pour représenter des séquences de texte, ou ajustés à des tâches spécifiques en un temps raisonnable pour obtenir d'encore meilleures performances. Devlin *et al.* (2018) a introduit le premier transformeur, appelé Bidirectionnal Encoder Representations from Transformers (BERT), dont les modèles pré-entraînés sont très populaires actuellement.

## 1.3 Recherche d'information pour la conception de robots conversationnels

Dans cette partie, nous présentons des techniques de Recherche d'Information (RI) qui nous seront utiles pour la conception des robots conversationnels pour l'immigration et pour l'information juridique, dans les parties 5.1 et 5.2.

Le domaine de recherche de la RI peut être exprimé ainsi d'après Manning *et al.* (2008, p. 1), :

> Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

Les auteurs décrivent ensuite brièvement un problème de classification de la manière suivante (Manning *et al.*, 2008, p. 253) :

> Given a set of classes, we seek to determine which class(es) a given object belongs to.

Dans le cas de la classification de texte, les objets sont des documents, tandis que les classes peuvent être de nature très variées. Un exemple qu'utilisent Manning *et al.* est celui du filtre des pourriels dans une boite de réception : c'est

donc une classification binaire avec des classes pourriel et non-pourriel. On peut aussi considérer les tâches de détection de sentiments pour extraire les critiques négatives d'un produit, ou encore la catégorisation des résultats d'un moteur de recherche en sujets préétablis. Ce sont autant d'exemples dans lesquels des tâches de classification sont utilisées à des fins de recherche d'information.

Certaines techniques développées en RI, comme l'utilisation de la *similarité cosinus* entre deux vecteurs, peuvent aussi être utilisées pour des tâches de classification. Puisque cette mesure de similarité produit un score, il suffit d'y appliquer un seuil pour produire une sortie binaire. Lorsque ces vecteurs représentent des documents, cette mesure permet de mesurer à quel point ils sont différents au regard des mots qui les composent, en faisant abstraction de leur longueur.

On peut reprendre l'exemple de la section 1.1 avec des documents textuels cette fois-ci. On prend $d_1$=«`visa`», $d_2$=«`étudiant`», $d_3$=«`visa étudiant` ». Les représentations en sac de mots des documents $d_1$ à $d_3$ avec le vocabulaire `[«visa», «étudiant»]` valent $\mathbf{v_1} = [1, 0]$, $\mathbf{v_2} = [0, 1]$ et $\mathbf{v_3} = [1, 1]$. La figure 1.5 illustre l'angle entre ces vecteurs. C'est une extension de la figure 1.2 appliqué au contexte de TALN. Le score de similarité $sim_{cosine}(\mathbf{v_1}, \mathbf{v_3}) = \frac{1}{2}$ est supérieur à la similarité $sim_{cosine}(\mathbf{v_1}, \mathbf{v_2}) = 0$, puisque $d_1$ et $d_3$ ont un mot en commun, à l'inverse du couple $d_1$ $d_2$.

On peut remarquer une différence importante entre une recherche par similarité cosinus et la plupart des algorithmes d'apprentissage qui sont utilisés pour la classification de texte. Lors d'une recherche par similarité cosinus, on va comparer le document de la requête avec chacun des documents candidats individuellement. Les algorithmes de classification cherchent plutôt à apprendre une séparation entre des régions de l'espace d'entrée. Ces séparations peuvent être simples comme une régression linéaire, mais aussi très complexe comme avec les

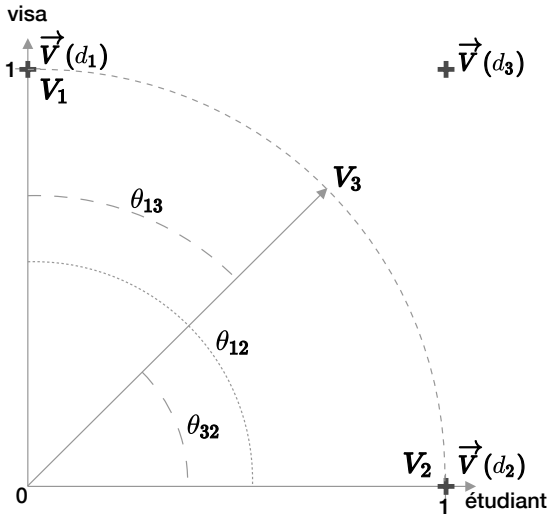



Figure 1.5 Similarité cosinus dans un espace en deux dimensions (taille du vocabulaire=2)

Séparateur à Vaste Marge (*Support Vector Machine* en anglais) (SVM) (Cortes et Vapnik, 1995) ou les Perceptron Multi-Couches (*Multi-Layer Perceptron* en anglais) (MLP) par exemple.

## 1.4 Agents conversationnels

Les robots conversationnels sont une autre catégorie d'outils que nous mettons à profit pour faciliter l'accès à la justice. Nous décrivons ici les contributions importantes du domaine sur lesquelles nous nous appuierons.

Gao *et al.* (2019) classifient les systèmes conversationnels en trois grands groupes. Les SQR constituent la première catégorie. Leur principale caractéristique est l'utilisation de données de types variés, telles que des pages web ou des graphes de connaissances, pour proposer des réponses directes à des questions d'utilisateurs. Le second groupe désigne les systèmes de dialogue dédiés à une tâche spécifique.

Enfin, la troisième catégorie regroupe les robots conversationnels “sociaux”, qui n’ont pas de tâche spécifique à réaliser et sont plutôt des systèmes généralistes.

### 1.4.1 Systèmes de Questions-Réponses

Les moteurs de recherche ou autres système de RI renvoient des documents aux requêtes des utilisateurs. Ces requêtes peuvent être formulées par combinaison de mots-clés, mais aussi parfois en langage naturel (sans contraintes particulières, notamment de vocabulaire, sur le langage permis). À titre d’exemple, l’utilisateur qui cherche à s’informer sur les autorisations de travail dans sa situation pourrait formuler les requêtes suivantes :

— `visa étudiant autorisation travail`

— `J’ai un visa étudiant, est-ce que je peux travailler ?`

Alors qu’un moteur de recherche retournerait sensiblement les mêmes documents à ces deux requêtes, les SQR pourraient utiliser les informations de la requête en langage naturel pour y répondre précisément. Plutôt que de devoir lire les documents complets des règles du visa étudiant, l’utilisateur du SQR pourrait recevoir une réponse comme “En règle générale, les étudiants peuvent travailler 20h/semaine pendant la session, [...]”. D’après Jurafsky et Martin (2014), ces systèmes se présentent sous deux formes. Le premier type de systèmes utilise des bases de connaissances structurées et associe les requêtes à des représentations logiques pour les interroger. Dans la second, on cherche d’abord les documents pertinents avec des techniques de RI, puis on utilise des algorithmes de compréhension du texte pour extraire les passages pertinents.

Dans le cas d’un SQR ayant pour rôle de donner des informations à des immigrants, une table de la base de connaissance pourrait être une table associant nationalité d’origine et type de visa à une liste de pré-requis. On a les ensembles

(incomplets mais pour les besoins de l'exemple) $type = \{travail, études, PVT\}$ et $nationalité = \{américaine, française\}$. La table de base de données correspondante est illustrée dans le tableau 1.3. La figure 1.6 présente un organigramme des interactions avec ce SQR pour répondre à la question «Quels sont les pré-requis pour avoir un visa ?» : les interactions avec l'utilisateur doivent renseigner les deux inconnus (nationalité et type de visa) avant qu'on puisse leur fournir une réponse en utilisant la table de base de données table 1.3. Puisque le système suit des règles très simples, il est entièrement prévisible et, à supposer que les informations sont correctes et complètes, on fournira la bonne réponse à l'utilisateur dans tous les cas.

Il serait possible d'appliquer ce paradigme dans un second temps à des systèmes tels que ceux que nous développons, pour faciliter encore l'accès à l'information. Ce paradigme n'est pas sans faille pour autant, notamment parce qu'il force les interactions à rester dans un cadre très rigide. L'effort de construction de la base de connaissance est aussi bien plus grand que celui nécessaire au regroupement d'un corpus de données non-structurée. Des travaux existent pour extraire de l'information structurée à partir de données textuelles relativement homogènes, notamment (Auer *et al.*, 2007) une ontologie créée à partir de l'ensemble des pages de Wikipedia et des relations entre elles (les liens hyptertexte). Ces approches permettent l'accès à une large base de connaissance relationnelle, au prix d'une baisse dans la qualité des données, par comparaison à un travail manuel. Malgré cette limitation, ce type d'outil constitue une source d'amélioration intéressante pour la suite de nos travaux. Nous avons notamment commencé son intégration dans les outils de TALN de la BNC pour augmenter leur couverture grâce à l'extraction automatique de synonymes depuis cette base de données.

Un SQR basée sur la RI pourrait être conçu avec un corpus contenant notamment les deux documents ci-dessous :

Tableau 1.3 Table de base de données du SQR construit sur une base de connaissance

| nationalité | type | pré-requis |
|---|---|---|
| américaine | travail | Il faut justifier d'une offre d'emploi. |
| américaine | études | Il vous suffit de demander une équivalence. |
| américaine | PVT | Ce permis n'existe pas pour les americains. |
| française | travail | Il faut justifier d'une offre d'emploi et d'une enquête de compétitivité. |
| française | études | Vous devez montrer une preuve d'acceptation et de fonds supérieurs à $10 000. |
| française | PVT | Vous devez vous inscrire au tirage au sort. |

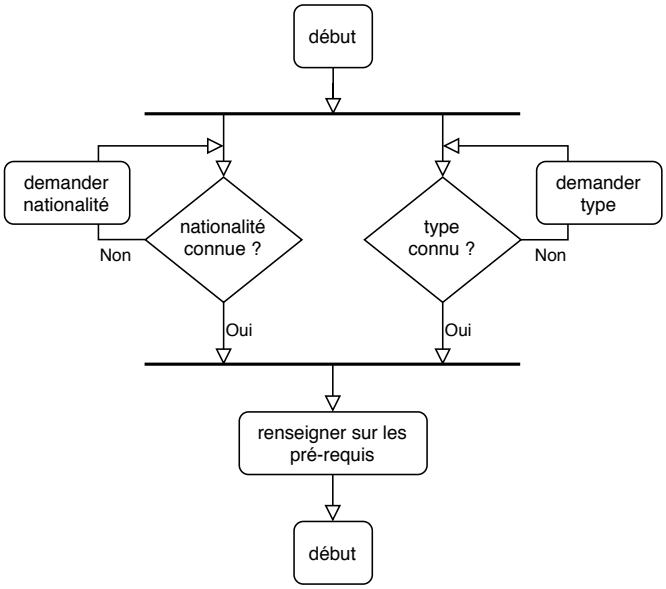



Figure 1.6 Organigramme de la partie "connaître les pré-requis pour un visa" d'un robot conversationnel informationnel d'immigration supporté par une base de connaissance structurée

— $d_a$=«Si vous avez la nationalité américaine, il vous faut justifier d'une offre d'emploi pour avoir un permis de travail. Pour obtenir un permis d'études, il suffit d'appliquer une équivalence de votre status universitaire américain. Le permis vacances-travail, ou PVT, n'existe pas pour les américains.»

— $d_f$=«Si vous avez la nationalité française, pour obtenir un permis de travail, il faut justifier [...]»

Pour répondre à deux requêtes $r_1$ et $r_2$ présentées ci-dessous, la première étape est d'identifier le bon document.

$r_1$=«Quels sont les pré-requis pour avoir un permis de travail pour les détenteurs de la nationalité américaine ? »

$r_2$=«Quels sont les pré-requis pour obtenir un permis de travail pour les américains ? »

Cette étape peut être réalisée à l'aide d'une des méthodes présentées à la section 1.1. Une simple similarité cosinus sur des vecteurs de sac de mots fonctionnerait avec la requête $r_1$ qui a du vocabulaire en commun avec le document $d_a$, mais pas avec la requête $r_2$ qui utilise des mots légèrement différents. Des représentations lexicales distribuées entrainés permettraient de faire correspondre ces deux requêtes avec le document $d_a$.

On peut ensuite utiliser diverses techniques pour extraire la partie pertinente du document pour répondre à la question. On commence généralement par appliquer une analyse grammaticale de la requête pour déterminer le type de question («Qui », «Quoi / Qu'est-ce qui », «Quand », etc.). Il est ensuite possible d'utiliser des outils d'étiquetage morpho-syntaxiques pour filtrer les résultats et de faire une comparaison par similarité cosinus par exemple pour sélectionner la ou les phrases les plus pertinentes.

### 1.4.2 Agents conversationnels généralistes

La conception de robots conversationnels qui ne soient pas restreints à une tâche spécifique est difficile, notamment parce que l'évaluation de tels systèmes est complexe. La caractérisation de ce qui constitue une bonne conversation a motivé de nombreux auteurs mais reste encore aujourd'hui une question de recherche ouverte (Deriu *et al.*, 2019a).

Le premier robot conversationnel généraliste ELIZA (Weizenbaum *et al.*, 1966) a été créé pour imiter une conversation avec un psychothérapeute. L'algorithme consistait simplement à associer certains mots-clés à des phrases prédéfinies. Grâce au choix de "persona" choisi par Weizenbaum *et al.*, le système recevait des réactions d'utilisateurs très positives, sans prendre de grands risques puisqu'il répondait principalement par de nouvelles questions.

Le système ALICE (Wallace *et al.*, 2003) était le premier à traiter les entrées en langue naturelle. Le langage Artificial Intelligence Markup Language (AIML) a été développé dans ce cadre, pour concevoir un système de règles de conversation. De nombreux robots conversationnels créés grâce à ce langage sont recensés par Satu *et al.* (2015).

Comme pour d'autres domaines du TALN, les systèmes de dialogue à base d'apprentissage sont actuellement les plus répandus. Vinyals et Le (2015) utilisent une architecture de réseau de neurones de type séquence vers séquence (*seq2seq*, introduite par Sutskever *et al.* (2014)) pour entraîner un robot conversationnel de bout en bout et sans connaissances à priori. Le modèle apprend simplement à prédire la réponse la plus probable (caractère par caractère) à une question tirée du corpus d'entraînement. La simplicité de ce modèle présente plusieurs avantages. D'abord, les seules données nécessaires sont des échanges de conversations : aucune annotation particulière n'est nécessaire, donc les corpus de données poten-

tiellement utilisables sont nombreux. Ensuite, puisque cette méthode ne nécessite pas de connaissances spécifiques au domaine, elle est très facilement applicable à des contextes variés.

Cette approche a toutefois ses limites que les auteurs reconnaissent. Premièrement, l'objectif de prédire la prochaine étape de conversation la plus probable n'est pas un très bon indicateur de l'objectif d'une conversation. En effet, le but d'une conversation est souvent un objectif à plus long terme comme le partage d'une certaine information ou l'accomplissement d'une tâche précise. Ensuite, le modèle a tendance à favoriser des réponses peu risquées, ce qui revient souvent à des réponses très courtes et peu intéressantes. Enfin, le modèle ne contient pas de mécanisme pour s'assurer que les réponses sont cohérentes entre elles. Ce dernier point est, d'après les auteurs, l'un des points qui empêchent leur modèle de passer le test de Turing.

Qiu *et al.* (2017) expliquent que les méthodes à base de RI échouent souvent à traiter des questions longues et précises. Les modèles de génération de texte risquent de générer des réponses incohérentes ou qui n'ont pas de sens. Les limitations de ces approches motivent la mise au point d'un robot conversationnel hybride, qui combine RI et génération de texte. Qiu *et al.* entraînent trois modèles : un modèle de RI, un de génération de texte, et un troisième qui choisit la réponse à donner à l'utilisateur en fonction du seuil de confiance du premier algorithme. Lors de l'évaluation effectuée manuellement, le modèle hybride dépasse le modèle simple de RI avec 60% de *precision@1* contre 40%.

*Évaluation*

Le *test de Turing*, introduit par Turing (1950) sous le nom d'*Imitation Game* consiste, pour une machine, à chercher à se faire passer pour un humain aux yeux d'un examinateur par l'intermédiaire d'une conversation écrite. D'après (Radziwill

et Benton, 2017), cet objectif a guidé la recherche du développement de robots conversationnels depuis ELIZA (Weizenbaum *et al.*, 1966).
Ramos (2017) ainsi que Radziwill et Benton (2017) suggèrent que cette aptitude à imiter un comportement humain n'est pas nécessairement une qualité désirable, et argumentent que même l'empathie des humains à l'égard de ces systèmes n'en souffrirait pas.

La compétition *ConvAI*[2] de la conférence Neural Information Processing Systems (NeurIPS) évalue les systèmes sur leur capacité à converser en maintenant l'apparence d'une personnalité consistante. Chaque conversation est précédée par quelques phrases décrivant la personnalité à adopter et les modèles ont ensuite des conversations avec des utilisateurs humains. Chacun d'eux (modèle et humain) a la même tâche : poser des questions à son interlocuteur pour apprendre à le connaître et répondre aux questions en respectant la personnalité qui a été attribuée. L'interlocuteur humain indique ensuite le persona avec lequel il pense avoir discuté, entre le vrai et un autre, choisi au hasard, et à quel point il a apprécié la conversation.
Cette compétition n'est pas la seule dédiée aux robots conversationnels, Deriu *et al.* (2019b) listent d'autres compétitions et métriques sans qu'un standard soit identifié. Une de ces techniques consiste à évaluer à quel point la discussion avec le robot conversationnel est pertinente. Les métriques BLEU (Papineni *et al.*, 2002) et ROUGE (Lin, 2004) notamment, mesurent le chevauchement entre le dialogue du robot conversationnel et des phrases pré-établies. Certaines approches, comme (Lowe *et al.*, 2017) utilisent un RNN pour essayer de prédire les notes que donneraient des juges à des phrases ou la conversation au complet. Cette approche nécessite une annotation manuelle de grande ampleur mais ses résultats sont corrélés assez étroitement avec des jugements d'utilisateurs. Malgré de nombreuses pistes

2. `http://convai.io/`

prometteuses, ce problème reste encore ouvert puisqu'il semble difficile d'identifier les facteurs de qualité d'un robot conversationnel en général.

### 1.4.3 Robots conversationnels spécialistes

Pour les robots conversationnels spécialistes, qu'ils soient à but informationnel ou plutôt transactionnel, le domaine est connu à l'avance. Dans le deuxième cas, le rôle du robot conversationnel est d'automatiser les échanges dans le but de réaliser une transaction : la résiliation d'un abonnement, ou l'envoi d'un virement bancaire par exemple. Cela a permis le développement de systèmes à états finis (Bobrow *et al.*, 1977) et à base de règles pour régir les transitions entre états. Ces techniques ont eu un grand succès et continuent à être utilisées encore aujourd'hui. Le système développé par Bobrow *et al.* est encore à la base de nombreux systèmes de réservations de vol aujourd'hui (Jurafsky et Martin, 2017).

Ces techniques ont pourtant l'inconvénient d'être peu flexibles. En effet, les transitions entre états sont normalement des questions (ou des patrons de questions, comme avec ALICE). Si plusieurs questions devraient mener à la même action, il faut dupliquer la règle. D'autres informations peuvent aussi être prises en compte dans le choix d'une transition entre états (par exemple la position géographique de l'utilisateur ou le nombre de jours avant la prochaine lotterie de visa, lorsque c'est applicable). De manière générale, si on a $n$ variables binaires, la taille de l'arbre des états possibles sera de $2^n$. La figure 1.7 illustre cela avec 3 variables binaires $A$, $B$ et $C$ dont dépend l'état $E$. Pour ces 3 variables, on a $2^3 = 8$ états par exemple. Comme nous le verrons dans les sections 5.1 et 5.2, les systèmes que nous développons ne nécessitent pas nécessairement un mécanisme de ce genre dans leur première version, mais pourraient en bénéficier dans des versions ultérieures.

Pour cette raison, il est souvent difficile d'énumérer chacuns d'entre eux et de

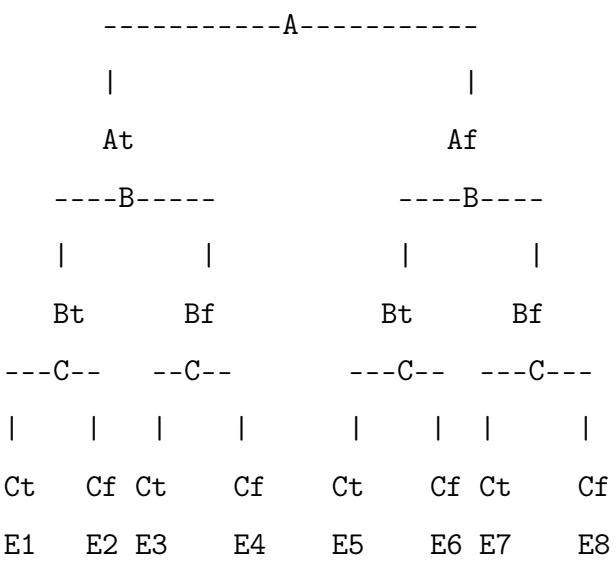



Figure 1.7 Croissance exponentielle du nombre d'états en fonction du nombre de variables

maintenir un tel système. Le framework Rasa [3] et d'autres outils récents de conception de robots conversationnels utilisent l'abstraction d'«intention»(Bocklisch *et al.*, 2017). Une intention est définie en fonction du résultat attendu d'un tour de conversation. Les deux phrases `Qui a le droit de travailler au Canada ?` et `Quels visas me permettent d'avoir une job au Canada ?` pourraient être considérées comme relevant de la même intention, par exemple. Un modèle de classification d'intention est chargé d'assigner la bonne intention à chaque interaction de l'utilisateur, puis, un second modèle utilise le contexte de la conversation et cette intention pour identifier la meilleure réaction à cette interaction. Si on ignore l'historique par exemple, on répondrait sûrement avec une action comme `retourner_infos_visa_travail`. Si on connait des informations sup-

3. `https://rasa.com`

plémentaires, l'action pourrait être personalisée, par exemple en retournant les informations sur le travail au Québec pour un immigrant français.

StarSpace (Wu *et al.*, 2018) est un algorithme pour apprendre des représentations lexicales distribuées d'entités de différents types dans un même espace, de manière supervisée. Pour la classification d'intentions, les entités sont donc les documents (le texte produit par les usagers et les questions de la FAQ), ainsi que les étiquettes des intentions. L'algorithme consiste à rapprocher les paires document-intention positives (celles où l'intention correspond au document) et à limiter la proximité des paires document-intention négatives. En pratique, les entités sont représentées par des caractéristiques, et c'est la représentation de ces caractéristiques qui est mise à jour en minimisant la fonction de perte (formule 1.9) par descente de gradient stochastique.

$$\text{Fonction de perte : } \sum_{\substack{a,b \\ b^- \in E^-}} L^{batch}(sim(a,b), sim(a,b_1^-), \cdots, sim(a,b_k^-)) \qquad (1.9)$$

Où $a$ est un document, $b$ l'intention associée à ce document et $E^-$ un sous-ensemble de $k$ intentions non associées à $a$, qu'on note $b_i^-$. La fonction de similarité `sim` est une similarité cosinus.

L'échantillonnage des instances négatives est une technique introduite par Mikolov *et al.* (2013a) qui permet un entraînement bien plus rapide qu'en calculant la perte sur toutes les instances du jeu de données. Avec cette technique, on met seulement à jour la représentation de k instances négatives, sélectionnées aléatoirement, plutôt que du nombre total d'exemples du jeu de données $n$, avec $k \ll n$.

Comme mentionné précédemment, on cherche à minimiser la distance entre $a$ et $b$, tout en cherchant à éloigner $a$ et les instances négatives ($b^- \in E^-$). Pour limiter

le temps d'entraînement, on arrête de se focaliser sur les instances qui sont déjà au-delà d'une certaine marge $\mu$. Cette technique s'appelle *margin ranking loss*, et se traduit par l'équation 1.10.

$$L = \begin{cases} -sim(a, x) & \text{si } x = b \\ \max(0, \mu - sim(a, x)) & \text{si } x \in E^- \end{cases} \tag{1.10}$$

On voit que la perte diminue lorsque la similarité entre $a$ et $b$ augmente. Pour les labels négatifs, $b^- \in E^-$, si la similarité entre $a$ et $b^-$ est supérieure à la marge $\mu$, la perte augmente.

Nous avons présenté des approches dans les domaines de la justice prédictive, puis des systèmes de robots conversationnels. Voyons maintenant la méthodologie que nous allons adopter pour répondre à notre question de recherche en mettant à contribution ces techniques.

# CHAPITRE II

# JUSTICE PRÉDICTIVE

Nous présentons dans ce chapitre l'état de l'art des travaux qui appliquent des techniques de TALN au domaine de la justice, qui informent notre contribution décrite à la section 4.3.1. Nous donnerons ensuite un aperçu de l'organisation de la justice au Canada et au Québec, afin de mieux comprendre le contexte dans lequel s'insère la suite de nos travaux.

## 2.1 État des travaux en justice prédictive

Loevinger (1948) débat longuement du besoin d'une science dédiée à l'étude des questions de droit. Il oppose cette science qu'il nomme Jurimétrique (*Jurimetrics* en anglais) à l'étude théorique du droit réalisée principalement par les philosophes et théoriciens du droit (la *jurisprudence* en anglais, à ne pas confondre avec le mot jurisprudence en français, qui se traduit par *case law*). En 1963, il précise la portée de ce domaine de recherche émergent :

> Jurimetrics is concerned with such matters as the quantitative analysis of judicial behavior, the application of communication and information theory to legal expression, the use of mathematical logic in law, the retrieval of legal data by electronic and mechanical means, and the formulation of a calculus of legal predictability. (Loevinger, 1963)

Loevinger défend la possibilité de prédire des décisions de justice contre l'un de ses contemporains, Wiener, qui estime que ces tentatives sont absurdes. Wiener (1962) critique les hypothèses sur lesquelles ces techniques s'appuient, et l'inévitable perte de compétence des praticiens qui utiliseraient des machines, plutôt que leur propre raisonnenemt, dans leur travail. Premièrement, la supposition que les décisions suivent la règle du précédent ne tient pas dans le cas de la Cour Suprême d'après lui. Puisqu'elle est au sommet de la hiérarchie des cours des États-Unis, la Cour Suprême a en effet le pouvoir de rejeter un précédent établi par n'importe quelle autre cour. Wiener rejette aussi l'idée selon laquelle les juges sont consistants dans leurs décisions et que connaître leurs décisions passées suffit à prédire les futures. Pour Zambrano (2015), il est évident que des prédictions peuvent être faites, puisque c'est ce que font les avocats sur la base de leur expérience, afin de conseiller leurs clients. La question devient donc «comment faire effectuer de telles prédictions par une machine sur une base quantitative et non intuitive» (Zambrano, 2015, p. 1). On notera que ce type particulier de problèmes est le terrain de jeu privilégié de l'apprentissage automatique. Les tâches de reconnaissances d'image et de TALN par exemple, principalement intuitives chez les humains, sont particulièrement difficiles à réaliser de manière automatique. Actuellement, les algorithmes à base d'apprentissage sont seulement capables de réaliser partiellement certaines de ces tâches, dans des cas souvent simples.

Zambrano évoque deux approches concurrentes à ce problème. L'approche par les règles d'abord, qui consiste à créer un formalisme permettant d'encoder les lois et les descriptions de faits. Un algorithme procède ensuite de manière déductive à partir de la représentation d'un cas pour prescrire la manière dont le Droit devrait être appliqué. L'utilisation de cette méthode porte ses fruits dans certains cas bien spécifiques, c'est la base des logiciels TAXMAN (McCarty, 1976) et TURBOTAX (Intuit Inc., 2019), voir (Goolsbee, 2004) pour le droit fiscal par exemple. L'approche par les cas ensuite, désigne l'entraînement d'un modèle sur

de nombreux exemples de décisions de justice. Ici, les règles de droit sont apprises implicitement et peuvent ensuite être appliquées à de nouveaux cas.

Katz (2012) va dans le même sens que Zambrano : il remarque que la «prédiction»d'éléments liés à un conflit juridique potentiel constitue une grosse partie de la valeur ajoutée par un avocat et relève plusieurs raisons pour lesquelles les algorithmes prédictifs peuvent avoir un avantage sur les experts humains. Tout d'abord, la quantité de données que les algorithmes sont capables de traiter leur permet de rencontrer plus de cas rares, qui pourraient autrement être totalement inconnus d'un avocat. Ce même argument est avancé, et supporté expérimentalement, pour des modèles entraînés sur des millions d'images de radiologie, dont la performance de détection d'un tissu cancéreux est excellente (Cruz-Roa *et al.*, 2013). Aussi, d'après Katz, la démarche même de conception d'un tel système nécessite une certaine transparence qui permettrait de tenter de limiter certains biais cognitifs humains qui passeraient autrement inaperçus.

Katz *et al.* (2017) analysent les jugements de la Cour Suprême des États-Unis depuis sa création grâce aux données de la *Supreme Court Database* (Spaeth *et al.*, 2016). Cette base de données structurée recense plus de 19 000 jugements de 1791 à 1945 dans sa version *Legacy*, et environ 13 000 jugements depuis 1946 dans la version *Modern*.

Cette base de données contient de nombreuses variables qui caractérisent un cas en particulier. Certaines variables sont d'ordre chronologique, d'autres renseignent sur le contexte (de quelle cour est issue la décision à étudier, quelle était cette décision, mais aussi la question juridique, etc.) et d'autres encore détaillent les votes de chacun des juges, l'orientation politique de la Cour à cette époque, etc.

Cette approche ne relève donc pas spécifiquement du TALN mais de l'apprentissage automatique classique. En utilisant une variante des Forêts Aléatoires (Breiman,

2001), Katz *et al.* parviennent à prédire les issues de jugements de la cour, ainsi que les votes individuels des juges qui la composent, avec une F-mesure de 0,69.

Aletras *et al.* (2016) s'est intéressé aux jugements de la Cour Européenne des Droits de l'Homme (CEDH) qui doit déterminer de la constitutionnalité d'une décision de juridiction inférieure. La décision est donc prise sur un nombre limité de critères, et Aletras *et al.* ont réduit la portée de leur étude à 3 articles en particulier. De plus, il est important de noter que toutes les applications reçues par la cour sont passées par un premier filtre sous la forme d'une liste d'admissibilité. Les demandes rejetées de cette manière ne sont pas conservées, et le système décrit ici n'en a pas connaissance. En gardant en tête ces limites, les résultats (79% d'accuracy sur un jeu de données équilibré) sont assez encourageants. Il est aussi intéressant de noter qu'en utilisant seulement la section «Circonstances», le modèle obtient une exactitude de 76%, donc une perte assez faible par rapport à la décision au complet.

Sulea *et al.* (2017) ont travaillé avec les décisions de la Cour de Cassation, l'équivalent français de la Cour Suprême du Canada. Les auteurs ont entraîné un modèle à prédire les jugements de la Cour, mais aussi le domaine du droit et l'époque à laquelle la description du cas et le jugement ont été rédigés. Contrairement à d'autres auteurs du domaine, Sulea *et al.* s'intéressent à limiter le biais introduit par la nature des données d'entrée de leur algorithme. En effet, les approches existantes utilisent le texte des décisions pour en «prédire »certaines caractéristiques comme l'issue du jugement, par exemple.

Plutôt que de rejeter l'approche entière comme étant non pertinente, Sulea *et al.* adoptent la méthode suivante : ils retirent les références directes aux éléments qu'ils cherchent à prédire et s'approchent ainsi d'une description de cas que pourrait faire un avocat. Le système développé obtient 98.6% de F1-score pour la

classification de l'issue d'un cas, parmi 6 classes possibles. Ces résultats sont très impressionnants et démontrent qu'avec un jeu de données adapté et dans un contexte juridique limité, il est possible d'obtenir un système très performant. Il faut toutefois noter que les données utilisées contiennent toujours des indicateurs très forts de l'issue du jugement, qui ne seraient pas disponibles avant que celui-ci soit prononcé. Des efforts supplémentaires à l'approche de masque qu'ont utilisée les auteurs seraient nécessaires pour rendre ce genre d'approches utilisables dans un contexte plus réaliste et nous pensons que des jeux de données de qualité permettront d'y contribuer.

Nous avons présenté dans cette section une évolution des approches à base de règles vers des modèles d'apprentissage automatique dans le domaine de la jurimétrique. Bien que ces dernières techniques rencontrent du succès dans de nombreuses applications du TALN, il est possible qu'elles ne suffisent pas à devenir applicables dans des situations pratiques dans le domaine juridique. Nous avons mentionné certaines limitations que les travaux que nous présenté comportent, et de nombreux auteurs travaillent plutôt à approcher un processus de raisonnement juridique qu'à traiter le texte de manière brute comme nous l'avons vu. On trouve notamment dans la conférence ICAIL [1] beaucoup de tâches présentées comme intermédiaires, ou pré-requises à la conception de justice prédictive. Parmi ces tâches, on trouve l'identification d'arguments, le résumé automatique ou l'extraction de règles à partir d'un corpus de lois.

## 2.2 Aperçu de la situation du droit au Canada et au Québec

Le droit au Canada est divisé en deux branches. Le *droit public* gouverne tout ce qui est d'intérêt général, c'est a dire tout ce regroupe tout ce qui a trait aux

1. `https://icail2019-cyberjustice.com/`

institutions, à leurs relations entre elles et avec les individus, et les relations entre individus qui concernent directement la société (Law, 2018). Les relations entre les individus, tant qu'elles ne concernent pas directement la société, ainsi qu'entre les individus et les compagnies sont de l'ordre du *droit privé*. C'est notamment ce qui est considéré par le droit de la famille, des contrats, de la propriété et du commerce.

La «compétence»désigne «l'aptitude légale d'une autorité publique à accomplir un acte dans un domaine donné» (Reid, 2001). Au Canada, certains domaines du droit relèvent très clairement de la compétence du fédéral ou de la province. Les questions liées aux armées par exemple sont toujours jugées dans des cours fédérales, alors que les institutions municipales sont uniquement gouvernées par leur province (Brun *et al.*, 2014). D'autres situations sont plus délicates à départager et Brun *et al.* passent plus de 150 pages à décrire la répartition des compétences en détail. Pour ce qui nous intéresse, rappelons-nous qu'en général, la compétence provinciale porte sur le droit privé, ainsi que les affaires sociales et économiques sur son territoire. Le Parlement fédéral a compétence sur tout le reste, principalement du droit public.

La common law est le système juridique issu du droit anglais, on le retrouve ainsi dans de nombreuses anciennes colonies britanniques. Selon ce système, la jurisprudence (l'ensemble des décisions juridiques passées) constitue le droit. Plus précisément, chaque décision rendue en cour devient potentiellement un *précédent*, c'est-à-dire qu'elle pourra être citée pour appliquer le même jugement dans un cas similaire. Si le jugement vient de la même cour ou d'une cour supérieure, le même raisonnement *devra* être appliqué, alors qu'un jugement d'une cour inférieure ou d'une *juridiction* différente sera considéré, mais pourra ne pas être suivi.

Dans le système dit «de tradition civiliste», aussi appelé droit romano-germanique

ou simplement *droit civil* au Québec, le droit est un ensemble de règles qu'on trouve, non pas dans les décisions passées, mais dans des textes de loi. Ces textes sont généralement regroupés par sujets dans des «codes».

> Note : le terme *droit civil* désigne aussi une branche du droit privé, dans laquelle on trouve aussi le droit commercial ainsi que le droit international privé. Les droits relatifs aux personnes mais aussi le droit de la famille, des successions et de la propriété sont englobés par le droit civil (Émond et Lucie, 2004).

Le Canada au niveau fédéral ainsi que toutes les provinces et les territoires, excepté le Québec, utilisent un droit par common law. Le Québec a hérité du droit de tradition civiliste. Le *Code Civil du Québec* recense les lois de droit privé en application au Québec. Même si ces deux systèmes sont très différents, ils s'influencent l'un l'autre très régulièrement. L'utilisation de la common law ne veut pas non plus dire qu'il n'y a aucun texte de loi en tant que tel. Les exemples les plus connus sont probablement la Charte canadienne des droits et libertés et la Constitution du Canada.

Nous avons déjà présenté (voir section 2.1) des systèmes à base d'apprentissage pour la Cour Européenne des Droits de l'Homme, basée sur la constitution européenne, et pour la Cour Suprême des États-Unis par exemple, qui utilise la common law. De la même manière, des approches à base de règles existent dans les deux types de systèmes juridiques. Malgré cela, le droit de tradition civiliste, se prête mieux à un système à base de règles que la common law puisqu'il repose sur un ensemble de règles, sous forme de textes juridiques dans les différents codes. Il est moins facile de concevoir un système de règles qui couvre l'ensemble de la jurisprudence puisque cet ensemble est toujours changeant (même si les lois

changent aussi) et de très grande taille. De plus, chaque décision en remplace potentiellement une autre plus ancienne, en totalité ou partie. Dans ce contexte, il serait intéressant de disposer d'un jeu de données avec suffisamment de structure pour pouvoir choisir d'inclure ou d'ignorer certaines parties, selon l'objectif de classification. Idéalement, la prédiction de l'issue d'un jugement ne serait pas basée sur des informations disponibles seulement après le jugement, ce qui n'est pas le cas dans les approches que nous avons décrites. Nous présenterons dans la section 4.3.1 la création d'un jeu de données pour remplir ce besoin.

# CHAPITRE III

# MÉTHODOLOGIE

Avant de présenter la méthodologie, il est important de comprendre dans quel contexte ce projet s'insère. Nous allons donc présenter la situation du droit au Québec et au Canada, qui influence directement notre approche de recherche.

## 3.1 Travail multidisciplinaire

Notre travail s'est effectué dans un contexte largement multidisciplinaire. Le projet LegalIA [1], dans lequel ce travail est intégré, regroupe des chercheurs de plusieurs disciplines différentes : principalement en informatique et en droit, mais aussi en éthique, anthropologie et linguistique. Travailler avec des personnes aux formations, sensibilités et intérêts si différents apporte des perspectives et des idées qui n'émergeraient pas nécessairement en isolation. L'effort de vulgarisation nécessaire à une communication avec des personnes en dehors de son champ d'expertise est souvent intéressant en soi et l'interaction réciproque nous permet d'apprendre des autres domaines. Enfin, puisque nul n'est expert dans toutes les disciplines impliquées dans le projet, nous sommes naturellement poussés à collaborer, ce qui est très enrichissant.

---

1. https://legalia.uqam.ca/

La multidisciplinarité vient aussi avec son lot de difficultés. Le principal défi est sûrement celui de la communication : les habitudes, les attentes et parfois même le vocabulaire peuvent être si éloignés entre des personnes venant de deux domaines différents que des incompréhensions sont inévitables. Bien sûr, aborder un projet s'étendant sur plus d'un domaine augmente aussi le risque, par le nombre de variables à contrôler et aussi parce que l'on dépend fortement des experts des autres domaines.

## 3.2 Orientation du projet en fonction de la disponibilité des données

La première tâche que nous avons abordée s'inscrit dans le cadre de la justice prédictive. L'objectif initial était de fournir à une personne une estimation de ses chances de succès en fonction de sa situation, si elle décidait d'entreprendre une démarche juridique. Répondre à cette question présente de nombreux défis, dont celui de la disponibilité de données nécessaires à l'entraînement d'un système à base d'apprentissage automatique.

La description que ferait une personne donnée de sa propre situation risque d'être très différente de celle que ferait une autre personne à sa place. Nous avons fait l'hypothèse qu'une personne formée au droit formulerait généralement des descriptions plus consistantes entre elles. Ce type de descriptions existe sous forme des notes que prennent assistant(e)s juridique ou avocat(e)s lors de leur première rencontre avec un client, mais ce sont des données couvertes par le secret professionnel entre un client et son avocat. D'après Dodek qui analyse l'état du privilège du secret professionnel entre l'avocat et son client dans la jurisprudence, «le Privilège»est devenu un «droit quasi constitutionnel» (Dodek, 2011, p. 3). Il est donc impossible, au moins dans le cadre de notre projet d'avoir accès à ce genre de données.

Une autre manière de capter ce qui constitue la situation du justiciable serait d'utiliser tous les documents qu'il/elle a soumis à la cour. Pour la prédiction, on lui demanderait de soumettre les documents à sa disposition. Concevoir une représentation capable de combiner ces informations de différents types aurait été sans nul doute un énorme défi. Encore une fois pourtant, nous sommes limités par l'accès aux données avant même de nous heurter au quelconque défi technique. En effet, ces documents sont accessibles au cas par cas, en se rendant sur place (dans la cour qui a rendu le jugement), mais ils ne sont pas numérisés et nous sont donc en pratique inaccessibles.

Nous avons donc décidé d'utiliser à la place les documents produits par une cour lorsqu'elle prend une décision, c'est-à-dire le texte du jugement. Ces données présentent l'avantage d'être plus facilement accessibles que les notes de consultations d'avocat(e)s, même si leur collecte présentera aussi des difficultés. Ces documents contiennent, dans le cas des cours de première instance, une description de la situation qui pourrait se substituer à celle que nous décrivions plus tôt.

## 3.3 Suivi d'expériences et reproductibilité

Ces dernières années, le phénomène de "crise de la reproductibilité" en sciences a été beaucoup discuté. Baker (2016) reporte les résultats d'un sondage où plus de 1500 chercheurs sont interrogés sur leur expérience à reproduire des travaux de recherche. Plus de 70% des participants ont déjà échoué à reproduire les expériences d'autres chercheurs, et plus de 50% pour leurs propres expériences. Cette étude porte sur des domaines variés des sciences mais l'informatique n'est pas épargnée.

Collberg et Proebsting (2016) ont pu reproduire seulement 32% parmi 402 expériences choisies au hasard et Gundersen et Kjensmo (2018) font un constat

similaire en passant en revue 400 articles de conférences de rang A [2] en informatique, et constatent que la plupart de ces expériences sont non reproductibles par manque de documentation.

Gundersen et Kjensmo suggèrent que le format des articles scientifiques ne se prête pas à la transmission de toutes les informations précises nécessaires à la reproduction des expériences, par opposition au code. Le partage du code, et des paramètres des modèles permettrait donc un niveau de reproductibilité supplémentaire par rapport à la simple description de l'expérience dans un article scientifique. La gestion des versions de code est permise par des outils comme subversion ou git.

Ces outils interviennent dans plusieurs étapes avant la publication de modèles et d'articles pour décrire les expériences. La conception d'un système à base d'apprentissage contient beaucoup d'éléments en interaction. Plus tôt, nous avons traité de ce sujet pour appuyer la recommandation de partager le code source des expériences mais avant ça, il faut souvent des dizaines de tentatives d'entraînement de modèle. La gestion de ces expériences est une tâche complexe mais nécessaire, pour laquelle il n'y a pas de solution largement acceptée comme les systèmes de versionnement de code.

Nous utilisons le framework MLflow (Zaharia *et al.*, 2018) pour deux éléments du processus d'élaboration de modèles. Pour le suivi d'expérience d'abord, nous gardons une trace de chaque exécution des scripts de collecte, de préparation de données, mais aussi de l'entraînement et l'évaluation de modèles.

Cet outil nous permet de revenir plus tard sur un résultat d'expérience, retrouver exactement son contexte pour l'analyser et la décrire, et la reproduire, ou faire de

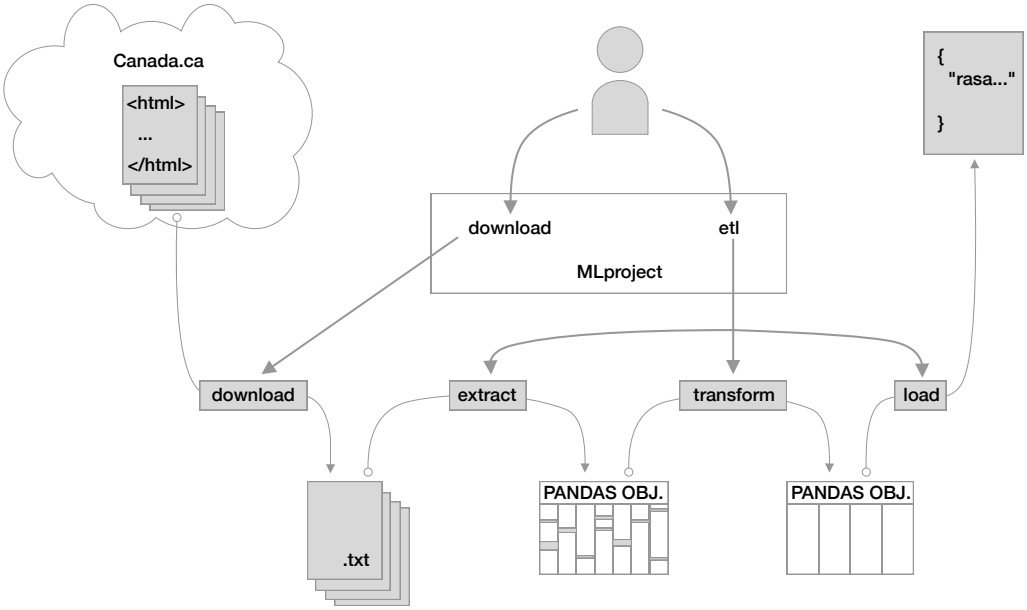
2. IJCAI et AAAI, notations ERA à travers `http://www.conferenceranks.com/`.

nouvelles expériences à partir du même montage expérimental. Nous avons aussi ajouté un niveau d'abstraction au dessus de chacun des scripts que nous utilisons (un *projet* MLflow) pour pouvoir les appeler simplement avec des valeurs par défaut. Ces scripts permettent une meilleure répétabilité de nos expériences puisqu'il suffit de quelques commandes pour entraîner les modèles que nous décrivons.

Pour le robot conversationnel d'immigration par exemple, la figure 3.1 décrit l'interaction d'un utilisateur avec le projet de collecte et préparation des données. Comme on peut le voir, l'utilisateur utilise uniquement les deux points d'entrées exposés dans le fichier *MLproject*. Le premier point d'entrée déclenche le téléchargement de la collection de documents, et le deuxième démarre leur traitement. Ce traitement de données est décomposé selon le patron de conception «extraire, transformer et charger ». Les trois étapes sont respectivement la récupération des données depuis la source, le nettoyage et le formattage des données, puis l'export vers le format final utilisable par le cadriciel RASA.

Après avoir décrit la méthodologie que nous avons employée, nous allons discuter des problématiques de disponibilité et de qualité des données qui ont été structurantes dans notre travail.

Figure 3.1 Processus de collecte et préparation de données pour le robot conversationnel d'immigration et interaction de l'utilisateur avec le système



# CHAPITRE IV

# DISPONIBILITÉ, COLLECTE ET UTILISATION DES DONNÉES

L'accès aux données a été un défi majeur qui a façonné le sujet de recherche et la manière dont nous l'avons abordé. Nous allons présenter ces enjeux dans la première sous partie, puis nous détaillerons les corpus que nous avons utilisés dans nos travaux.

## 4.1 Disponibilité des données

Comme il est mentionné dans (Queudot et Meurs, 2018), l'accès à large échelle aux décisions de justice à des fins d'analyses est souvent compliqué, voir impossible selon les juridictions. Cela s'explique principalement par des raisons économiques mais aussi de protection des données personnelles et de la vie privée des individus.

Les cours de justice du Canada gardent les documents liés aux procédures juridiques qu'elles traitent au format papier et il est toujours possible d'y accéder au cas par cas en faisant la demande sur place. La gestion et la diffusion de l'information juridique est ensuite déléguée à des organismes partenaires.

Dans le cas du Québec, c'est la Société québecoise d'information juridique (SOQUIJ) qui a pour mandat de permettre l'accès aux citoyens à l'information juridique (incluant les décisions des tribunaux judiciaires et administratifs). Puisque

l'organisme ne dispose pas de budget à cette fin, il doit donc trouver des façons de générer des revenus pour couvrir ses coûts. Ces derniers concernent non seulement la numérisation, le stockage et la distribution des documents, mais aussi l'anonymisation de certains jugements, et un choix éditorial sur les décisions à publier et la rédaction de résumés et d'annotations d'une sélection de jugements importants. La SOQUIJ met donc à disposition gratuitement tous les jugements sur une base individuelle par l'intermédiaire de son site web : la collecte de masse de ces données est interdite. Pour les cabinets d'avocats, mais aussi pour les organismes dont les besoins dépassent ce qui est disponible publiquement, des forfaits d'accès ou des partenariats sont négociés sur une base individuelle.

D'autres organismes privés effectuent leur propre indexation des décisions de nombreuses cours et proposent le même type de forfaits d'accès, mais aussi l'accès à des documents avec un prix à la pièce, de l'ordre de 2-3$ chacun. Ces offres, plutôt orientées vers des cabinets d'avocats, mais aussi adaptées à quiconque a besoin de quelques décisions particulières sont déraisonnables pour nos fins d'analyses dans le cadre académique.

D'après la *Loi sur l'accès aux documents des organismes publics et sur la protection des renseignements personnels*[1], il est aussi possible de faire une "demande d'accès à l'information" comme pour les données de n'importe quel organisme public, moyennant les frais de traitement. Dans notre cas, cette option n'était pas envisageable à cause des délais de traitement de ces demandes.

À part ces organismes qui possèdent et traitent les données juridiques à travers plusieurs juridictions et les rendent faciles d'accès et d'exploitation, il est aussi possible de visionner la plupart des décisions sur les sites web des cours où elles sont rendues. Sur chacun des sites de cour auxquels nous avons accédé, il n'est

1. `http://legisquebec.gouv.qc.ca/en/showdoc/cs/A-2.1?langCont=fr` D

pas permis de récupérer des documents à large échelle.

Cet état de fait n'est pas un hasard, la position du Barreau du Québec notamment, est que les données de justice doivent être "facile d'accès, mais pas trop"[2]. En théorie, cela voudrait dire faciliter l'accès à l'information juridique pour les citoyens et les praticiens du droit, mais en complexifier l'accès en masse. Le gouvernement veut absolument éviter que les jugements soient accessibles par une simple recherche web par exemple. Un abus de ce genre a eu lieu aux États-Unis sur le site web `mugshots.com`. Les auteurs de ce site récupéraient les photos prises au moment de l'arrestation des suspects dans de nombreux comtés des États-Unis pour les rendre disponibles et extorquer de l'argent des suspects pour retirer leur photos. Dans ce cas, les photos étaient déjà partagées publiquement sur les sites web des agences individuelles de police mais les rassembler et leur donner une audience a conduit à de nombreuses pertes d'emploi et autres discriminations[3].

En juillet 2017 toutefois, la Cour Fédérale du Canada faisait exception : aucune mention limitant l'accès aux décisions n'était présente sur son site. Après avis favorable des juristes du projet, nous avons donc téléchargé l'ensemble des pages web contenant des décisions (détails à la section 4.3.1). Pour éviter la fuite de ces données, nous les avons stockées sur un serveur interne à accès restreint.

En juillet 2017, nous avons aussi négocié un accès à des jugements avec la SOQUIJ pour l'entrainement de systèmes de classification et de dialogue. Malgré une entente signée à cette époque, ce projet n'a pas pu aller de l'avant puisque la SOQUIJ n'a finalement jamais donné suite à cet engagement.

---

2. Déclaration du Batonnier lors d'une présentation à destination du barreau de Montréal le 27 avril 2018.

3. `https://www.chicagotribune.com/business/ct-biz-mugshot-website-owners-extortion-20180518-story.html`

## 4.2 Qualité des données

À la problématique de l'accès au données s'ajoute celle de leur qualité. Dans le cas du jeu de données de la Cour Fédérale que nous décrivons à la section 4.3.1, les données ne sont pas organisées de manière à pouvoir être collectées ou analysées facilement. Toutes les décisions ne sont pas non plus présentes, comme nous l'avons décrit dans (Queudot et Meurs, 2018). Enfin, même une fois collectés, les documents sont sous la forme de texte brut sans annotation. La structure HTML des pages contenant les décisions ne permet pas de séparer simplement les sections.

Même si nous avons présenté des travaux qui réussissent à utiliser la jurisprudence pour prédire l'issue d'un cas, ces approches ont toutes de fortes limitations. Par exemple, tous les systèmes développés à partir de la jurisprudence dont nous avons connaissance utilisent directement le texte d'une décision existante pour en prédire l'issue. Certains auteurs masquent une plus ou moins grande partie du texte mais l'information utilisée n'est pas disponible en tant que tel dans une situation réaliste.

Pour concevoir de tels systèmes, nous manquons de données qui soient disponibles avant même d'entamer une démarche en justice. Un exemple de données qui pourrait être adaptées à ce contexte serait une description de la situation du requérant par un avocat, prise lors d'un premier rendez-vous. Pour des raisons évidentes de confidentialité et de secret professionnel, ce genre de données n'est pas disponible au public.

Tableau 4.1 Statistiques du jeu de données de la Cour Fédérale

| | Occurrences | % du total |
|---|---|---|
| Nombre de décisions uniques | 46 369 | 100% |
| Décisions uniquement en anglais | 2 329 | 5% |
| Décisions uniquement en français | 602 | 1.2% |
| Nombre de juges différents [4] | 41 | - |

Tableau 4.2 Nombre de documents des différentes composantes de la Cour Fédérale

| Base de données | Nombre de documents |
|---|---|
| FC | 9 729 |
| CF | 29 393 |
| CAF | 7 260 |
| Total | 46 382 |

## 4.3 Corpus utilisés

### 4.3.1 Cour Fédérale

*Détails sur le corpus*

Les décisions de la Cour Fédérale peuvent être rendues en français ou en anglais mais sont souvent traduites dans les deux langues. Si l'un des partis le demande, la décision doit aussi être traduite. Le détail de la répartition des langues est présent dans le tableau 4.1.

Ce jeu de données, bien qu'extrait du site web de la Cour Fédérale du Canada,

4. Statistique issue du site de la Cour Fédérale.

contient des documents issus de trois bases de données distinctes : la Cour Fédérale elle-même (CF), la Cour d'Appel Fédérale (CAF) et la Cour Canadienne sur l'Impôt (CCI).

La Cour Fédérale traite des cas de droit public qui incluent l'immigration, le droit des réfugiés, le droit maritime, la propriété intellectuelle, la défense et sécurité nationale et les relations internationales. La CCI traite uniquement du droit des impôts et du revenu, et la CAF reçoit les demandes d'appel de la CF et la CCI, ainsi que de certains tribunaux provinciaux en particulier. Le tableau 4.2 recense le nombre de documents dans chacune de ces trois cours.

*Indexation*

Nous avons collecté tous les documents disponibles sur le site de la Cour Fédérale. Comme nous l'avons décrit à la section 4.2, la structure des pages ne permet pas d'identifier les sections des décisions. Nous avons donc seulement extrait le texte brut des pages à l'aide de la librairie BeautifulSoup [5]. Pour explorer le corpus facilement, nous l'avons indexé à l'aide du moteur d'indexation Apache Lucene [6] après un pré-traitement simple pour retirer les doublons et rassembler les versions anglaise et française des mêmes cas.

*Annotation du corpus*

Comme nous l'avons décrit dans la section 3.1, nous avons collaboré avec une étudiante à la maîtrise en droit et avocate au Barreau de Montréal qui a annoté manuellement les documents tel que nous allons le décrire dans cette section. Pour cette tâche d'annotation, nous avons installé et utilisé une instance locale de l'outil

5. `https://www.crummy.com/software/BeautifulSoup/`

6. `https://lucene.apache.org/`

open source d'annotation en ligne *brat*[7].

La conception du schéma d'annotation et la mise en place de l'environnement nécessaire est notre contribution.

Comme nous le décrivons dans la section 4.1, les applications possibles de l'Intelligence Artificielle pour la justice sont limitées par la disponibilité de ressources seulement textuelles (les jugements par exemple). Nous avons donc entrepris d'annoter une partie des documents collectés afin d'y ajouter des données structurées.

Ces annotations sont composées d'*entités* qui peuvent être reliées par des *relations*. Par exemple, une demande de contrôle judiciaire peut être motivée par une loi en particulier. On affecterait alors l'entité *judicial_review* (demande de contrôle judiciaire) à la portion de texte qui en parle, l'entité *law* à la mention d'un ou plusieurs règlements, et la relation *motivated_by* (justifié(e) par) pour les relier.

Présentées dans l'annexe A, le tableau A.1 liste les entités utilisées pour l'annotation des décisions (Federal Courts Act, 1985) et le tableau A.2 contient les descriptions des relations entre ces entités.

### 4.3.2 FAQ d'immigration Canada

Nous avons collecté 1088 pages en anglais sur le centre d'aide du gouvernement du Canada sur les thèmes d'immigration et citoyenneté[8].

L'ensemble de ces pages constitue une Foire Aux Questions (FAQ) d'immigration. Chaque page commence par une question associée à une catégorie (par exemple «Étudier »/ «Studying») et est suivie d'une réponse constituée d'un ou plusieurs

---

7. `https://brat.nlplab.org`

8. `https://www.cic.gc.ca/francais/centre-aide/index-en-vedette-can.asp`

paragraphes de texte. La page contient aussi une liste de questions que d'autres utilisateurs ont trouvées utiles. Enfin, deux listes de mots-clés sont aussi présentes dans la page mais invisibles à l'utilisateur. Après vérification, nous avons choisi de ne pas extraire ces informations puisqu'elles ne changent pas d'une page à l'autre.

Une fois les documents récupérés, nous utilisons la structure HTML pour extraire les différentes informations de chaque page. Le code de collecte et de préparation des données est rendu disponible [9] pour faciliter la reproduction des résultats.

Notre jeu de données consiste en 1088 classes avec seulement un (1) exemple par classe. Pour les modèles à base d'apprentissage, ces exemples constitueront notre jeu de données d'entraînement. Afin de pouvoir guider la phase d'optimisation de notre système et en faire une évaluation préliminaire, nous avons créé des paraphrases de 88 exemples du jeu d'entraînement. Avec ce jeu de données de validation, il nous est possible de valider nos hypothèses et d'itérer rapidement sur la conception du système. Ce jeu de données bénéficierait d'un volume plus important en augmentant le nombre d'exemples annotés pour évaluer la performance du système de manière plus rigoureuse. Pour les besoins de la preuve de concept du travail de mémoire cependant, ce corpus de petite taille est suffisant.

### 4.3.3 FAQ juridique interne

Ce deuxième robot conversationnel est développé au sein de la Banque Nationale du Canada (BNC) pour répondre aux questions d'ordre légal des employés. J'ai travaillé sur ce projet lors d'un stage, puis en temps qu'employé à temps partiel en parallèle de la fin de ma maîtrise. Comme pour le cas du robot conversationnel d'immigration, ce robot est basé sur une FAQ et les tours de conversations

9. `https://gitlab.ikb.info.uqam.ca/marc/immigration_faq_scrapper`

sont assez indépendants les uns des autres. La principale difficulté réside donc à nouveau en la classification de l'intention de l'utilisateur en fonction des intentions auxquelles notre modèle a été exposé lors de sa conception. Dans ce cas aussi, peu de données sont disponibles initialement pour la conception du robot conversationnel.

Ce jeu de données contenait initialement 2 formulations pour chacune des 275 questions. Un premier robot conversationnel a été entrainé sur ce corpus, il est décrit comme l'expérience de référence avec StarSpace classique à la section 5.2.2. Le robot conversationnel a ensuite été exposé à une trentaine de ses utilisateurs - les membres des services juridiques - à la BNC, dans cette version préliminaire, pour collecter de vraies interactions. Les nouvelles formulations pour des intentions existantes ont permis de générer un jeu de test représentatif des questions réelles des utilisateurs. Après avoir filtré les questions en dehors du cadre de la FAQ et retiré les duplicats, nous avons annoté 292 interactions couvrant 126 intentions pour constituer le jeu de données de test.

Ce robot conversationnel interne à la BNC est actuellement prêt à être mis en production.

Dans ce chapitre, nous avons présenté trois jeux de données liés à l'information juridique. Le premier est constitué de décisions de justice, les deux autres sont des FAQ. Dans le chapitre suivant, nous allons décrire la conception de robots conversationnels à partir de ces jeux de données de petite taille. Le premier a pour vocation de donner des renseignements sur des sujets liés à l'immigration au Canada. Le second s'intègre dans un cadre corporatif pour informer ses utilisateurs sur les règles de droit liées à leur emploi.

# CHAPITRE V

# EXPÉRIENCES ET ANALYSE DE RÉSULTATS

Dans cette section, nous présentons les expériences réalisées sur les jeux de données décrits précédemment. Le premier robot conversationnel répond à des questions d'immigration grâce à des données publiques. Le second utilise un jeu de donné privé pour traiter des questions juridiques au sein de la BNC.

## 5.1 Robot conversationnel pour des questions d'immigration

### 5.1.1 Expérience de référence avec StarSpace

Nous avons entrainé un classifieur sur le jeu de données de questions-réponses d'immigration, décrit à la section 4.3.2, en utilisant l'algorithme StarSpace décrit à la section 1.4.3. Nous créons donc une représentation commune aux phrases des utilisateurs et aux intentions (une par exemple du jeu d'entraînement). L'implémentation de RASA [1] retourne aussi un classement des instances ayant la plus haute valeur de similarité avec la valeur en entrée.

Pour éviter le suraprentissage de notre modèle, nous utilisons une couche de décrochage (*dropout*) pour supprimer 20% des neurones à chaque phase d'entrainement.

1. `https://github.com/RasaHQ/rasa`

Nous entrainons deux modèles pendant 40 époques, c'est à dire que l'algorithme verra l'ensemble du jeu de données 40 fois au total. Le premier utilise la perte standard décrite dans (Wu *et al.*, 2018), avec le paramètre *mu* (qui contrôle la marge) à 0,8. Pour le deuxième modèle, la perte est contrôlée par une fonction *softmax*, il n'y a donc pas de marge à partir de laquelle l'entrainement s'arrête pour un couple d'instances en particulier. La fonction *softmax* prend en entrée un vecteur de $K$ nombres réels et produit un vecteur de $K$ nombres réels strictement positifs dont la somme vaut 1. Du fait de cette caractéristique, on peut interpréter cette sortie comme une distribution de probabilités. Dans le cas d'une classification multi-étiquettes, ces probabilités sont celles que l'exemple appartienne à chacune des classes possibles.

### 5.1.2 Expériences à base de RI

Nous utilisons aussi une version modifiée de l'algorithme StarSpace qui abandonne le concept d'intention pour directement associer une phrase de l'utilisateur à une réponse issue de la base de connaissance. Plutôt que d'apprendre à associer un document texte à une classe (qui ne porte pas d'information), on rapproche ici deux documents texte. Nous avions décrit à la section 1.4.3 que l'apprentissage de la position des entités était indirect, c'est en effet la position des caractéristiques (des sacs de mots) qui est apprise. La représentation des mots déjà connus informe donc la représentation lexicale distribuée des futurs documents qui les emploient, comme dans l'algorithme original de StarSpace. Cette approche apporte toutefois deux différences : premièrement l'ajout des réponses augmente le volume de données disponible pour l'apprentissage des représentations ; ensuite, au moment de prédire la bonne réponse à une interaction de l'utilisateur, on peut mettre à profit les informations contenues dans la réponse.

Tableau 5.1 Résultats de la classification d'intention du robot conversationnel d'immigration

| | microP | microR | microF | macroP | macroR | macroF | wP | wR | wF |
|---|---|---|---|---|---|---|---|---|---|
| softmax | 0,92 | 0,52 | 0,67 | 0,51 | 0,52 | 0,52 | 0,51 | 0,52 | 0,52 |
| marge | 1 | 0,60 | 0,75 | 0,60 | 0,60 | 0,60 | 0,60 | 0,60 | 0,60 |
| RI | 1 | 0,60 | 0,75 | 0,60 | 0,60 | 0,60 | 0,60 | 0,60 | 0,60 |

La première représentation des documents utilise des $n$-grammes avec $n \in \{1, 4\}$, puis on entraîne les représentations lexicales distribuées pendant 40 epochs avec les mêmes paramètres que le modèle StarSpace classique décrit à la section 5.1.1.

### 5.1.3 Résultats

Les résultats des expériences de la partie compréhension du langage naturel (classification d'intention ou sélection de la meilleure réponse) sont reportés dans le tableau 5.1.2. Entre les deux versions de StarSpace classique, le modèle qui utilise la perte avec une marge dépasse les performances de la version avec softmax. On remarque que dans ce cas-ci, l'utilisation de la variante RI de StarSpace n'apporte absolument aucune amélioration sur la version qui utilise les intentions. Plusieurs raisons à cela peuvent être envisagées. Par exemple, les instances faciles du corpus de test ont déjà été classées correctement et les instances difficiles ne peuvent pas l'être par simple similarité de vocabulaire. Aussi, le test ne couvrant que 88 classes parmi les 1088, il n'est peut être pas suffisamment représentatif de l'ensemble du domaine pour que le modèle s'applique avec succès. Des variantes utilisant des modèles préentrainés (Devlin *et al.*, 2018) se comporteraient peut-être mieux mais auraient l'inconvénient de présenter plus de classes pour moins d'instances. La taille relative des questions-réponses pourrait également avoir une influence sur

les résultats. Ces pistes pourront être investiguées dans la suite de ces travaux.

## 5.2 Robot conversationnel interne pour des questions de droit

### 5.2.1 Référence avec StarSpace classique

Nous utilisons également le framework open-source Rasa pour entraîner nos robots conversationnels.

Ce modèle de référence utilise une représentation initiale des documents par $n$-grammes avec $n \in \{1, 2\}$ Le modèle est entraîné pendant 20 époques avec la perte à marge de StarSpace. La marge pour les instances positives (quand arrêter de rapprocher les couples positifs) est $\text{mu}_{\text{pos}} = 0,8$ et la marge négative (quand arrêter d'éloigner les couples négatifs) est $\text{mu}_{\text{neg}} = -0,4$.

Pour comprendre pourquoi la valeur négative de $\text{mu}_{\text{neg}}$ est négative, il faut revenir sur la définition de la similarité cosinus que nous avons donné. La figure 1.5 représente visuellement la similarité cosinus dans un cas où chaque dimension signifie la présence ou l'absence d'un terme puisque l'exemple utilise une représentation par sac de mots. Dans ce cas, la valeur minimum que peut prendre $\text{sim}_{\text{cosine}}$ est 0. Pourtant, pour certaines caractéristiques comme le sentiment associé à un terme, des valeurs négatives peuvent avoir du sens. Les représentations lexicales distribuées, et celles apprises par StarSpace notamment peuvent apprendre ce genre de caractéristiques de manière implicite. Dans ces cas-là, la similarité la plus basse qu'il est possible d'obtenir est de -1, ce qui correspond à un angle cos de $\pi$, ou 180 degrés. C'est ce qui explique la valeur négative de $\text{mu}_{\text{neg}}$, puisque la plage des valeur possibles est en fait de [-1 ;1].

### 5.2.2 Référence améliorée

Une autre expérience de référence à laquelle nous allons comparer notre approche utilise l'algorithme BERT (Devlin *et al.*, 2018) pour représenter les conversations. BERT est un *transformeur* (comme nous l'avons décrit à la section 1.2.4) pour lequel un modèle préentrainé est disponible. Il existe une version anglaise, mais pas de version française, nous utilisons donc un modèle entraîné sur 104 langues (dont le français).

Le modèle pré-entrainé est la version M-BERT$_{BASE}$ [2], c'est-à-dire la version de base de BERT avec 110M de paramètres, par opposition aux 345M de paramètres de $BERT_{LARGE}$ (non disponible en version multilingue). Le modèle a ensuite été finement réglé sur le jeu de données d'entrainement pendant 100 époques, sur une machine avec une carte graphique Nvidia GTX1070.

### 5.2.3 Expérience à base de RI

De manière similaire au modèle décrit à la section 5.1.2, nous utilisons une variante de StarSpace classique qui met à profit les réponses (aux questions de FAQ) pour répondre aux questions.

La figure 5.1 illustre le déroulement des étapes du pipeline de l'algorithme StarSpace classique dans le contexte de la BNC. Dans cet exemple, on voit l'identifiant de l'intention du premier document être rapproché de chacun des mots qui composent ce dernier. La figure 5.2 représente la version où les réponses (en noir) sont utilisées comme entités. Les mots des questions sont en bleu, ceux des réponses en noir et le mot "compte" souligné en noir est présent à la fois dans les questions

2. `https://storage.googleapis.com/bert_models/2018_11_23/multi_cased_L-12_H-768_A-12.zip`

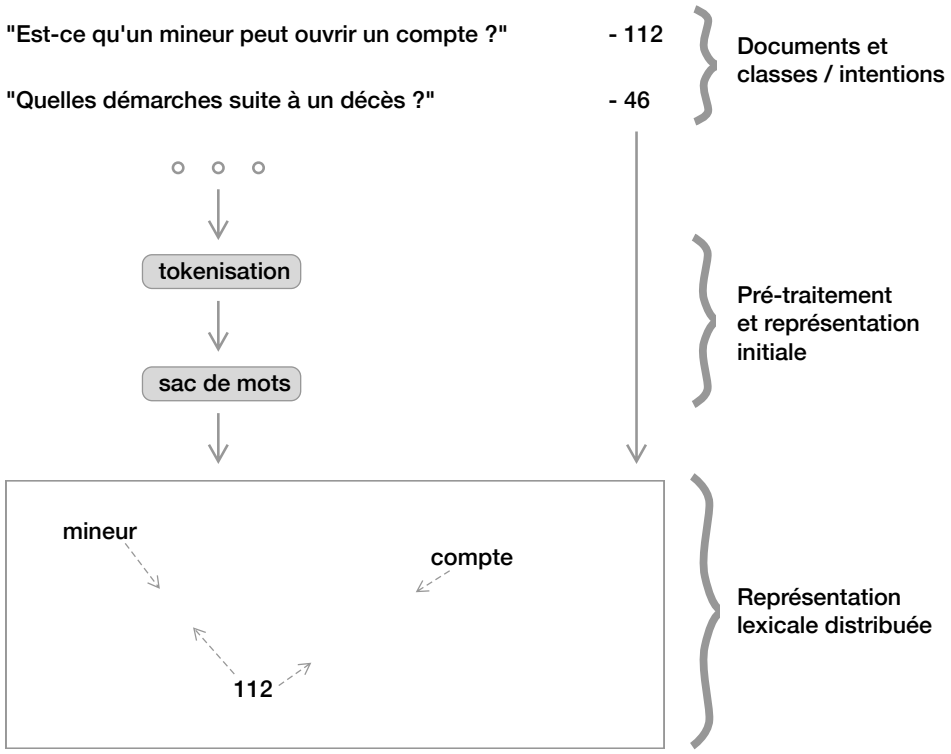



Figure 5.1 Les étapes d'apprentissage de représentation sémantique distribuée avec des intentions

et dans les réponses.

### 5.2.4 Résultats

Les résultats des expériences sur les données légales sont reportés dans le tableau 5.2.4. On note précision, rappel et F-mesure par les lettres P, R et F. Micro, macro et w (pour *weigthed*, c'est à dire pondéré ; la lettre P étant déjà utilisée) désignent le mode d'agrégation. Ces métriques sont décrites à la section 1.1.

La version standard de StarSpace est dépassée par les deux autres approches. Les performances en utilisant une moyenne des prédictions au niveau des instances

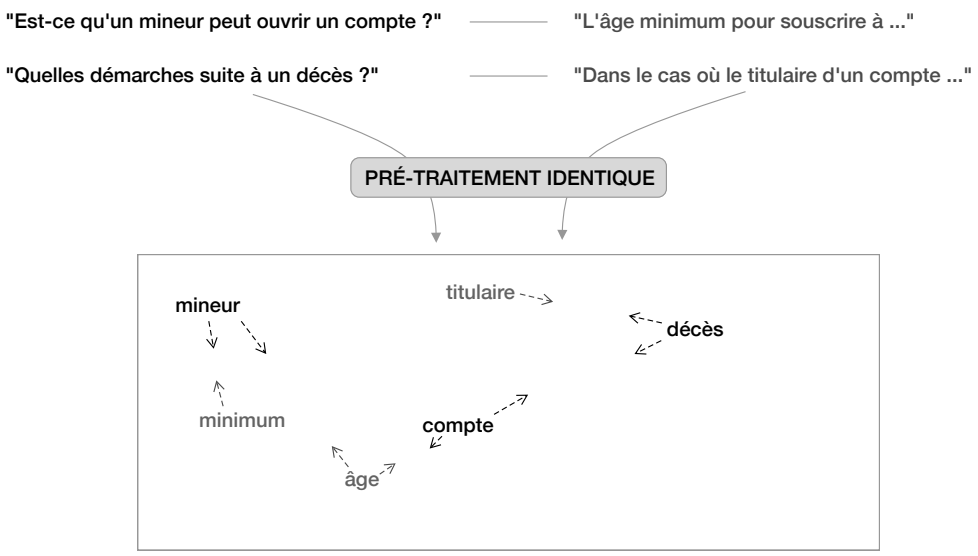



Figure 5.2 Étapes d'apprentissage de représentation utilisant les réponses

(micro-précision, rappel et F-mesure) donne l'avantage à notre modèle, alors que pour les deux autres types de moyenne (macro et pondérée), le modèle BERT est meilleur.

Si nous considérons les différences entre le modèle StarSpace original et la modification RI, on observe des différences de plus de 10% de F-mesure au profit du second. Il semble que le modèle RI utilise les informations supplémentaires

Tableau 5.2 Performances des 3 systèmes sur la prédiction d'intention du jeu de données légales de la BNC

| | microP | microR | microF | macroP | macroR | macroF | wP | wR | wF |
|---|---|---|---|---|---|---|---|---|---|
| StarSpace | 0,61 | 0,61 | 0,61 | 0,52 | 0,48 | 0,48 | 0,73 | 0,61 | 0,63 |
| BERT | 0,70 | 0,67 | 0,66 | 0,75 | 0,75 | **0,75** | 0,85 | 0,75 | **0,76** |
| RI | 0,78 | 0,64 | **0,70** | 0,64 | 0,62 | 0,60 | 0,72 | 0,64 | 0,65 |

contenues dans les réponses pour mieux classer les interactions utilisateurs. Par contre, cette différence n'est pas aussi marquée si on utilise une moyenne pondérée par le nombre d'instances dans chaque classe, ce qui pourrait rejoindre les questionnements présentés en 5.1.2.

Le niveau de performance de RI n'est pas non plus suffisant pour dépasser celui que permet un modèle de langue comme BERT. Ce dernier met à profit des informations apprises sur le contenu entier de wikipédia. Malgré le coût rédhibitoire (plus de 10 000$) pour entraîner un modèle au complet, il semble que pour un jeu de données de petite taille comme le notre, l'utilisation de modèles préentraînés est bénéfique. Si on considère le coût d'entraînement, le poids du modèle et surtout le temps d'inférence, un modèle simple comme RI offre déjà une amélioration significative sur un modèle StarSpace simple.

# CONCLUSION

L'accès à l'information juridique est un frein majeur à l'accès à la justice. Dans ce mémoire, nous avons présenté trois travaux liés à l'intelligence artificielle qui peuvent aider à améliorer cette situation.

Dans un premier temps, nous avons collecté et indexé toutes les décisions rendues par la Cour Fédérale du Canada, la Cour Canadienne de l'Impôt et la Cour d'Appel du Canada. Ce faisant, nous avons soulevé des problématiques importantes d'accès et de qualité des données. Nous avons ensuite conçu un schéma d'annotation en s'appuyant sur la loi sur les cours fédérales (Federal Courts Act, 1985) qui les régit, puis nous en avons annoté une centaine avec l'aide d'une avocate et étudiante à l'UQAM. Ce corpus pourra être utilisée pour diverses tâches de justice prédictive comme la prédiction de l'issue d'un cas mais aussi du domaine du droit concerné, de la / des lois pertinentes, etc. Chacune de ces tâches a un impact sur l'information donnée aux citoyens qui envisagent une démarche juridique.

Nous avons aussi conçu deux robots conversationnels dans le but d'informer leurs utilisateurs sur des questions de droit. Le premier répond à des questions liées à l'immigration, et l'autre, basée sur une base de connaissance de la Banque Nationale du Canada, répond aux questions de droit de ses employés. Tous deux se basent sur une Foire Aux Questions dont le nombre de questions ne dépasse pas, ou de peu, le nombre de réponses. La tâche de classification sous-jacente a donc un nombre d'exemples par classes très bas (moins de 5) pour un nombre de classes très élevé (275 et 1088 respectivement). Nous avons expérimenté avec un algorithme permettant d'apprendre des représentations lexicales distribuées de manière supervisée.

Nous avons remarqué que, dans le cas le moins extrême du jeu de données de la BNC, l'utilisation d'une variante qui représente aussi les réponses dans le même espace et les utilise pour faire les prédictions apporte une amélioration très significative. Sur ce même jeu de données, l'utilisation d'un modèle de langue préentraîné sur un corpus de très grande taille et finement réglé sur le nôtre augmente encore les performances. Le surcoût d'entraînement est irréaliste pour tous, si ce ne sont les plus gros acteurs du domaine, mais l'existence de réseaux préentraînés sur certaines langues limite l'impact de ce défaut. Sur le jeu de données d'immigration, la variante n'apporte aucune amélioration de performance. Ces résultats sont à mettre en perspective avec la faible taille du jeu de données de tests par rapport au jeu de données d'entraînement (seulement 10% des classes sont couvertes par le jeu de test).

Nous avons décrit les deux robots conversationnels comme des systèmes de classification. Le robot conversationnel de la BNC emploie une technologie de reformulation des questions posées basée sur des mots-clés pour augmenter son score global. Hors du champ de ce travail et faisant l'objet d'un brevet, cette technique ne pourra pas être utilisée dans le système de dialogue pour l'immigration. En revanche, puisque plus de 17 000 documents de la Cour Fédérale traitent d'immigration, nous envisageons de compléter les réponses tirées de la FAQ par des jugements pertinents. Pour cela, nous utiliserons l'index de documents de la Cour déjà existant et le document de la question provenant de l'utilisateur.

Ce type de documents ne sera pas nécessaire à la plupart des utilisateur d'un système comme le nôtre mais il pourrait constituer un premier pas vers un outil plus large comme nous l'avions décrit dans la vidéo de présentation de LegalIA [3], et donc vers un meilleur accès à la justice pour tous.

---

3. `https://legalia.uqam.ca/video/LegalIA.mp4`

# APPENDICE A

# ENTITÉS UTILISÉES POUR ANNOTER LE CORPUS DE LA COUR FÉDÉRALE

Tableau A.1: Description des entités utilisées pour annoter les décisions

| *Entity* | *Category* | *Description* |
|---|---|---|
| Judge | Decision_maker | |
| Prothonotary | Decision_maker | |
| Applicant | Party | |
| True_applicant | Party | |
| Respondant | Party | |
| Judicial_review | Action | |
| JR_jurisdiction | Action<br>→ Judicial_review | Request for judicial review when the decision maker "acted without jurisdiction, acted beyond his or her jurisdiction or refused to exercise his or her jurisdiction" (Federal Courts Act, 1985) |

| | | |
|---|---|---|
| JR_PNJ | Action<br>→ Judicial_review | Request for judicial review when the decision maker "failed to observe a principle of natural justice, procedural fairness or other procedure that it was required by law to observe" (Federal Courts Act, 1985). |
| JR_legal_error | Action<br>→ Judicial_review | Request for judicial review when the decision maker "erred in law in making a decision or an order, whether or not the error appears on the face of the record" (Federal Courts Act, 1985). |
| JR_facts_error | Action<br>→ Judicial_review | Request for judicial review when the decision maker "based its decision or order on an erroneous finding of fact that it made in a perverse or capricious manner or without regard for the -material before it" (Federal Courts Act, 1985). |
| JR_fraud | Action<br>→ Judicial_review | Request for judicial review when the decision maker "acted, or failed to act, by reason of fraud or perjured evidence" (Federal Courts Act, 1985). |

| | | |
|---|---|---|
| JR_illegal | Action<br>→ Judicial_review | Request for judicial review when the decision maker "acted in any other way that was contrary to law" (Federal Courts Act, 1985). |
| Appeal | Action | |
| Constitutional_challenge | Action | |
| Mandamus | Action | |
| Objection | Action | Motion to strike |
| Intervention | Action | Motion for intervention |
| Other | Action | |
| Law | | |
| Ans_Granted | Answer | |
| Ans_Dismissed | Answer | |
| Issue_Granted | Issue | |
| Issue_Dismissed | Issue | |
| Correct_decision | Standard_of_review | Standard of review : correct decision |
| Reasonable_decision | Standard_of_review | Standard of review : reasonable decision |
| Facts | | |

**Note** : Une demande de contrôle judiciaire (en anglais, *a request for judicial review*) consiste à évaluer si une décision prise par une instance inférieure devrait être ré-examinée. Dans le cas de l'immigration par l'exemple, une décison de refuser un visa peut être prise par un agent d'immigration, arriver devant la Cour Fédérale à la demande de la personne concernée, et peut être renvoyée devant un nouvel agent pour une nouvelle évaluation du cas si la cour décide que la loi n'a pas bien été respectée.

Tableau A.2: Description des relations entre les entités d'annotation

| Arg1 | Relation | Arg2 |
|---|---|---|
| Action | Motivated_by | Law |
| Standard_of_review | Analysis_of | Action |
| Answer | Answers | Action |
| Judge | Rules | Answer |

# GLOSSAIRE

**apprentissage automatique** (*Machine Learning* en anglais). L'Apprentissage Automatique désigne le domaine de recherche en informatique d'étude des algorithmes qui apprennent leur comportement grâce aux données. Ce fonctionnement est décrit par opposition à la manière traditionnelle de concevoir des algorithmes pour résoudre des problèmes avec une machine. La programmation consiste à rédiger une suite d'instructions fixes et précises que la machine exécute. L'Apprentissage Automatique permet d'apprendre une partie de ce comportement. On dit aussi que ces algorithmes «s'améliorent avec l'expérience», ce qui signifie que leur performances augmentent en relation avec la quantité de données dont on dispose pour les entraîner. 20, 38, 39, 46

**caractéristique** En Apprentissage Automatique, les caractéristiques (*features* en anglais) sont des propriétés individuellement mesurables d'un phénomène observé. Ce sont ces représentations simplifiées qui sont fournis entrée des algorithmes d'apprentissage, plutôt que la donnée elle-même. Pour une tâche d'estimation du prix de vente d'une maison, on pourrait par exemple utiliser les caractéristiques «surface», «nombre de chambres»et «distance au centre-ville»pour représenter chaque maison. Un autre choix, peut-être moins judicieux pourrait être la couleur des volets et la hauteur de la porte d'entrée. On voit bien que la question du choix des caractéristiques utilisés et de leur nombre est loin d'être triviale, et elle influencera directement la qualité des résultats de nos algorithmes. 35, 62

**décrochage** Le décrochage (ou *dropout* en anglais), permet d'éviter le surapprentissage dans les réseaux de neurones en supprimant aléatoirement une partie des neurones. 61

**Forêts Aléatoires** (*Random Forest* en anglais). Cet algorithme d'Apprentissage Automatique combine les prédictions d'un ensemble d'arbres de décisions (*Decision Trees*), chacun construits sur un sous-ensemble aléatoire de traits (*features*) des données pour limiter la sur-généralisation du modèle. 39

**Intelligence Artificielle** L'intelligence artificielle est un domaine de recherche de l'informatique qui est défini par Russell et Norvig (2016) comme la conception et la construction d'agents intelligents qui perçoivent un environnement et prennent des mesures qui influent sur cet environnement. 2, 3, 5, 57

**jurisprudence** La jurisprudence désigne l'ensemble des décisions de justice sur un sujet donné. Ces décisions combinées sont supposées combler les failles et les ambiguïtés les unes des autres et les juges sont lié à elles par la règle du précédent. Ce principe dicte que le même raisonnement juridique doit être appliqué aux cas similaires, et c'est en étudiant la jurisprudence que cela se traduit. La jurisprudence est un faux-ami en anglais, où il désigne l'étude théorique du droit. Le mot anglais pour «jurisprudence»est «case law». 42

**représentation lexicale distribuée** Les représentations lexicales distribuées, ou embeddings en anglais, sont des représentations vectorielles qui mettent à profit le contexte des éléments représentés pour conserver certaines informations sémantiques. Voir la description en détails à la section 1.2.2. vii, 17, 18, 19, 29, 35, 62, 63, 64, 69

**robot conversationnel** (*chatbot* en anglais). Un robot conversationnel est un programme informatique conçu pour simuler une conversation avec des utilisateurs humains, en particulier sur internet. iii, iv, vii, 2, 4, 23, 25, 28, 30, 31, 32, 33, 34, 36, 49, 58, 59, 64, 69, 70

# ACRONYMES

**BERT** Bidirectionnal Encoder Representations from Transformers 23

**BNC** Banque Nationale du Canada v, 27, 58, 61, 67, 69, 70

**CEDH** Cour Européenne des Droits de l'Homme 40, 43

**FAQ** Foire Aux Questions 57, 58, 59, 65, 69, 70

**JNR** Justiciables Non-Représentés 1

**MLP** Perceptron Multi-Couches (*Multi-Layer Perceptron* en anglais) 25

**NeurIPS** Neural Information Processing Systems 32

**RI** Recherche d'Information iv, 9, 11, 23, 24, 26, 27, 31, 65

**RNC** Réseau Neuronal Convolutif (*Convolutional Neural Network* en anglais) 22

**RNN** Réseau de Neurones Récurrents (*Recurrent Neural Network* en anglais) 20, 21, 22, 32

**SOQUIJ** Société Québécoise d'Information Juridique 7, 51, 52, 53

**SQR** Systèmes de Question-Réponses v, 25, 26, 27, 28

**SVM** Séparateur à Vaste Marge (*Support Vector Machine* en anglais) 25

**TALN** Traitement Automatique du Language Naturel (*Natural Language Processing* en anglais) 5, 9, 14, 15, 17, 22, 24, 27, 30, 37, 38, 39, 41

**TF-IDF** Term Frequency-Inverse Document Frequency 16

# RÉFÉRENCES


Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D. et Lampos, V. (2016). Predicting Judicial Decisions of the European Court of Human Rights : A Natural Language Processing Perspective. *PeerJ Computer Science*, *2*, e93.

Almeida, H., Queudot, M., Kosseim, L. et Meurs, M.-J. (2017). Supervised Methods to Support Online Scientific Data Triage. Dans *International Conference on E-Technologies*, 213–221. Springer.

Almeida, H., Queudot, M. et Meurs, M.-J. (2016). Automatic Triage of Mental Health Online Forum Posts : CLPsych 2016 System Description. Dans *Third Workshop on Computational Linguistics and Clinical Psychology at NAACL HLT*, 183–187.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R. et Ives, Z. (2007). Dbpedia : A nucleus for a web of open data. Dans *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, p. 722–735., Berlin, Heidelberg. Springer-Verlag.

Bahdanau, D., Cho, K. et Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv :1409.0473*.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News*, *533*(7604), 452.

Balas, E. et de Souza, C. C. (2005). The vertex separator problem : a polyhedral investigation. *Mathematical Programming*, *103*(3), 583–608.

Bengio, Y., Ducharme, R., Vincent, P. et Jauvin, C. (2003). A Neural Probabilistic Language Model. *Journal of machine learning research*, *3*(Feb), 1137–1155.

Bobrow, D. G., Kaplan, R. M., Kay, M., Norman, D. A., Thompson, H. et Winograd, T. (1977). GUS, a Frame-Driven Dialog System. *Artificial intelligence*, *8*(2), 155–173.

Bocklisch, T., Faulkner, J., Pawlowski, N. et Nichol, A. (2017). Rasa : Open source language understanding and dialogue management. *arXiv preprint arXiv :1712.05181*.

Breiman, L. (2001). Random Forests. *Machine learning*, *45*(1), 5–32.

Brun, H., Tremblay, G. et Brouillet, E. (2014). Droit constitutionnel. *Éditions Yvon Blais*, p. 1055.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. et Bengio, Y. (2014). Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv preprint arXiv :1406.1078*.

Collberg, C. et Proebsting, T. A. (2016). Repeatability in Computer Systems Research. *Communications of the ACM*, *59*(3), 62–69.

Cortes, C. et Vapnik, V. (1995). Support-Vector Networks. *Machine learning*, *20*(3), 273–297.

Cruz-Roa, A. A., Ovalle, J. E. A., Madabhushi, A. et Osorio, F. A. G. (2013). A Deep Learning Architecture for Image Representation, Visual Interpretability and Automated Basal-Cell Carcinoma Cancer Detection. Dans *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 403–410. Springer.

Deriu, J., Rodrigo, A., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E. et Cieliebak, M. (2019a). Survey on Evaluation Methods for Dialogue Systems. *arXiv preprint arXiv :1905.04071*.

Deriu, J., Rodrigo, A. M., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E. et Cieliebak, M. (2019b). Survey on Evaluation Methods for Dialogue Systems. *ArXiv*, *abs/1905.04071*.

Devlin, J., Chang, M.-W., Lee, K. et Toutanova, K. (2018). BERT : Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv :1810.04805*.

Dodek, A. M. (2011). Solicitor-Client Privilege in Canada : Challenges for the 21st Century. *Canadian Bar Association*.

Émond, A. et Lucie, L. (2004). Introduction à l'étude du droit. *Les Cahiers de droit*, *45*(3), 611–613.

Federal Courts Act (1985). . R.S., 1985, c. F-7, 18.1(4). Récupéré de `https://www.canlii.org/fr/ca/legis/lois/lrc-1985-c-f-7/derniere/lrc-1985-c-f-7.html#art18.1par4`

Gao, J., Galley, M., Li, L. *et al.* (2019). Neural Approaches to Conversational AI. *Foundations and Trends® in Information Retrieval*, *13*(2-3), 127–298.

Goolsbee, A. (2004). The TurboTax Revolution : Can Technology Solve Tax Complexity ? *The Crisis in Tax Administration*, 124–38.

Gundersen, O. E. et Kjensmo, S. (2018). State of the art : Reproducibility in artificial intelligence. Dans *Thirty-Second AAAI Conference on Artificial Intelligence*.

Harris, Z. S. (1954). Distributional Structure. *WORD*, *10*(2-3), 146–162. http://dx.doi.org/10.1080/00437956.1954.11659520

Intuit Inc. (2019). *TurboTax*. Récupéré le 30 Janvier 2019 de https://turbotax.intuit.ca/

Jurafsky, D. et Martin, J. H. (2014). Question Answering. In *Speech and Language Processing*, volume 3. Pearson.

Jurafsky, D. et Martin, J. H. (2017). Dialog Systems and Chatbots. *Speech and language processing*, *3*.

Katz, D. M. (2012). Quantitative Legal Prediction – or – How I Learned to Stop Worrying and Start Preparing for the Data Driven Future of the Legal Services Industry. *Emory LJ*, *62*, 909.

Katz, D. M., Bommarito II, M. J. et Blackman, J. (2017). A General Approach for Predicting the Behavior of the Supreme Court of the United States. *PlOS ONE*, *12*(4), e0174698.

Landauer, T. K. et Dumais, S. T. (1997). A Solution to Plato's Problem : The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological review*, *104*(2), 211.

Laniel, R.-A., Bahary-Dionne, A. et Bernheim, E. (2018). Agir seul en justice : du droit au choix—état de la jurisprudence sur les droits des justiciables non représentés. *Les Cahiers de droit*, *59*(3), 495–532.

Law, J. (2018). *A dictionary of law* (9 éd.). Oxford University Press. http://dx.doi.org/10.1093/acref/9780198802525.001.0001

Lin, C.-Y. (2004). ROUGE : A package for automatic evaluation of summaries. Dans *Text Summarization Branches Out*, 74–81., Barcelona, Spain. Association for Computational Linguistics. Récupéré de https://www.aclweb.org/anthology/W04-1013

Loevinger, L. (1948). Jurimetrics–The Next Step Forward. *Minnesota Law Review*, *33*, 455.

Loevinger, L. (1963). Jurimetrics : The Methodology of Legal Inquiry. *Law and Contemporary Problems*, *28*(1), 5–35.

Lowe, R., Noseworthy, M., Serban, I. V., Angelard-Gontier, N., Bengio, Y. et Pineau, J. (2017). Towards an Automatic Turing Test : Learning to Evaluate Dialogue Responses. *arXiv preprint arXiv :1708.07149*.

Manning, C. D., Manning, C. D. et Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT press.

Manning, C. D., Raghavan, P. et Schütze, H. (2008). *Introduction to Information Retrieval*. New York, NY, USA : Cambridge University Press.

Maupomé, D., Queudot, M. et Meurs, M.-J. (2019). Inter and Intra Document Attention for Depression Risk Assessment. Dans *Canadian Conference on Artificial Intelligence*, 333–341. Springer.

McCarty, L. T. (1976). Reflections on TAXMAN : An experiment in artificial intelligence and legal reasoning. *Harvard Law Review*, *90*, 837.

Mikolov, T., Chen, K., Corrado, G. et Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv :1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. et Dean, J. (2013b). Distributed Representations of Words and Phrases and Their Compositionality. Dans *Advances in Neural Information Processing Systems*, 3111–3119.

Mohamed Didi Biha and Marie-Jean Meurs (2011). An exact algorithm for solving the vertex separator problem. *J. Global Optimization*, *49*(3), 425–434.

Papineni, K., Roukos, S., Ward, T. et Zhu, W.-J. (2002). BLEU : a Method for Automatic Evaluation of Machine Translation. Dans *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics.

Qiu, M., Li, F.-L., Wang, S., Gao, X., Chen, Y., Zhao, W., Chen, H., Huang, J. et Chu, W. (2017). AliMe Chat : A Sequence to Sequence and Rerank Based Chatbot Engine. Dans *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, volume 2, 498–503.

Queudot, M. et Meurs, M.-J. (2018). Artificial Intelligence and Predictive Justice : Limitations and Perspectives. Dans *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, 889–897. Springer.

Radziwill, N. M. et Benton, M. C. (2017). Evaluating Quality of Chatbots and Intelligent Conversational Agents. *arXiv preprint arXiv :1704.04579*.

Ramos, R. (2017). Screw the Turing Test-Chatbots don't need to act human. *VentureBeat. Retrieved on March*, *13*, 2017.

Reid, H. (2001). *Dictionnaire de droit québécois et canadien*. Montréal : Wilson & Lafleur.

Robertson, S. (2004). Understanding inverse document frequency : on theoretical arguments for IDF. *Journal of documentation*, *60*(5), 503–520.

Russell, S. J. et Norvig, P. (2016). *Artificial Intelligence : a Modern Approach*. Malaysia ; Pearson Education Limited,.

Sarfati, M., Queudot, M., Mancel, C. et Meurs, M.-J. (2017a). Formulation relaxée de la séparation équilibrée d'un graphe. Dans *ROADEF 2017, 18ème congrès de la société française de recherche opérationnelle et d'aide à la décision*.

Sarfati, M., Queudot, M., Mancel, C. et Meurs, M.-J. (2017b). Knowledge Discovery in Graphs Through Vertex Separation. Dans *Canadian Conference on Artificial Intelligence*, 203–214. Springer.

Satu, M. S., Parvez, M. H. *et al.* (2015). Review of Integrated Applications With AIML Based Chatbot. Dans *2015 International Conference on Computer and Information Engineering (ICCIE)*, 87–90. IEEE.

Schneider, R. K. (2010). Illiberal Construction of Pro Se Pleadings. *University of Pennsylvania Law Review*, *159*, 585.

Shannon, C. E. et Weaver, W. (1998). *A Mathematical Theory of Communication*. University of Illinois press.

Spaeth, H. J., Epstein, L., Martin, A. D., Segal, J. A., Ruger, T. J. et Benesh, S. C. (2016). 2016 Supreme Court Database, Version 2016 Legacy Release v01. (SCDB_Legacy_01). `http://supremecourtdatabase.org/`.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, *28*(1), 11–21.

Sulea, O.-M., Zampieri, M., Malmasi, S., Vela, M., Dinu, L. P. et van Genabith, J. (2017). Exploring the Use of Text Classification in the Legal Domain. *arXiv preprint arXiv :1710.09306*.

Sutskever, I., Vinyals, O. et Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. Dans *Advances in Neural Information Processing Systems*, 3104–3112.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, *59*(236), 433.

Vinyals, O. et Le, Q. (2015). A Neural Conversational Model. *arXiv preprint arXiv :1506.05869*.

Wallace, R., Tomabechi, H. et Aimless, D. (2003). Chatterbots Go Native : Considerations for an Eco-System Fostering the Development of Artificial Life Forms in a Human World. *Published online : http ://www. pandorabots. com/pandora/pics/chatterbotsgonative. doc*.

Weizenbaum, J. *et al.* (1966). ELIZA—a Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM*, *9*(1), 36–45.

Wiener, F. B. (1962). Decision Prediction by Computers : Nonsense Cubed — and Worse. *American Bar Association Journal*, 1023–1028.

Wu, L. Y., Fisch, A., Chopra, S., Adams, K., Bordes, A. et Weston, J. (2018). Starspace : Embed all the things ! Dans *Thirty-Second AAAI Conference on Artificial Intelligence*.

Xu, W. et Rudnicky, A. (2000). Can artificial neural networks learn language models ? Dans *Sixth International Conference on Spoken Language Processing*.

Zaharia, M., Chen, A., Davidson, A., Ghodsi, A., Hong, S. A., Konwinski, A., Murching, S., Nykodym, T., Ogilvie, P., Parkhe, M. *et al.* (2018). Accelerating the Machine Learning Lifecycle with MLflow. *IEEE Data Eng. Bull.*, *41*(4), 39–45.

Zambrano, G. (2015). Précédents et prédictions jurisprudentielles à l'ère des Big Data : parier sur le résultat (probable) d'un procès. Récupéré le 28 Janvier 2018 de `https://hal.archives-ouvertes.fr/hal-01496098`