Open Law

Restitution du programme IA & Droit

OPEN LAW*
* Le droit ouvert

Données d'apprentissage

Livre Blanc

Restitution du programme

IA & Droit
-
Données d'apprentissage

« Ce programme a offert, dans une approche collaborative, une première appréhension concrète des utilisations possibles de l'intelligence artificielle pour l'analyse des décisions de justice. Cela a permis de prendre la mesure des perspectives ouvertes mais également de l'ampleur des obstacles à surmonter. »

Edouard Rottier, responsable du bureau des diffusions numériques au service de documentation, des études et du rapport (SDER) de la Cour de cassation

L'association Open Law*
Le droit ouvert

L'association Open Law*, Le droit ouvert est un espace de travail et d'expérimentation pour l'ensemble des acteurs du monde du droit prêts à innover dans un mode collaboratif et ouvert. Open Law se présente également comme un espace d'exploration de problématiques émergentes, telles que la Blockchain et les Smart Contrat, le Legal Design ou encore l'Open Gov.

Toutes ses actions aboutissent à la production de ressources communes ouvertes (référentiels, ontologies, livres blancs, modèles de statuts, cahiers des charges, bases de données, développements Open Source, etc.).

Aujourd'hui, l'association Open Law réunit des représentants de l'ensemble des acteurs de l'information juridique et, plus largement, du monde du droit. Sa communauté rassemble environ 2 000 professionnels invités à interagir lors de plus de 100 temps forts depuis sa création.

Sommaire

L'association Open Law*, Le droit ouvert 5

Sommaire 7

1 La genèse du programme 9

2 Le cas d'usage : zonage des décisions de justice 15

3 Le choix de l'outil 19

3.1 Le choix de l'annotateur Brat 20
3.2 Facile à déployer, difficile à utiliser 21
3.3 Quels outils pour de futurs projets d'annotation de paragraphes ? 22

4 Une méthodologie d'annotation à parfaire 25

4.1 Le travail de création du plan d'annotation avec la Cour de cassation 26
4.2 Le retour à la réalité lors du travail d'annotation 28
4.3 Ce que l'on en retient pour les prochains projets 30

5 L'évaluation de la qualité des annotations 35

5.1 Les méthodes d'évaluation mises en place 36
5.2 Evaluation du dataset 38
5.3 Pistes d'amélioration 42

6 Bibliographie 45

1

La genèse du programme

Le projet *IA et Droit - données d'apprentissage* a pour objectif la création collaborative de jeux de données d'apprentissage pour des algorithmes d'intelligence artificielle dans le domaine du droit.

L'apprentissage automatique (ou *machine learning*) révolutionne la façon de concevoir les outils d'intelligence artificielle, en prenant le contre-pied des systèmes à base de règles (*rule-based*) utilisés jusqu'ici dans le monde juridique.

Dans son livre blanc sur l'IA[1] publié en 2016, l'INRIA indique que « *les résultats les plus remarquables obtenus dans le domaine de l'apprentissage automatique sont basés sur l'apprentissage supervisé, c'est-à-dire l'apprentissage à partir d'exemples dans lesquels le résultat attendu est fourni avec les données d'entrée. Cela implique d'étiqueter les données avec les résultats attendus correspondants, un processus qui nécessite des données à grande échelle* ». Dans le monde juridique français, ces données d'apprentissage n'existent pas, ce qui induit un coût d'entrée conséquent pour tous les acteurs souhaitant intégrer ce marché émergent. Par ailleurs, ce manque de données d'entraînement de qualité est à l'origine de beaucoup des risques éthiques (manque de transparence, amplification, reproduction voire création de biais...) liés au développement des outils d'intelligence artificielle.

C'est la raison pour laquelle l'association *Open Law**, *Le droit ouvert* a lancé le projet *IA et droit - données d'apprentissage*, en partenariat avec la mission «Éthique et Algorithme» de la CNIL[2]. Répondant à l'objectif de «do-tank» que l'association s'est fixé et grâce à sa communauté diversifiée, qui rassemble les compétences juridiques et techniques nécessaires, ce projet vise à constituer des jeux de données d'entraînement exploitables par tous.

En 2017, la première saison du programme *IA et droit - données d'apprentissage* avait pour objectif de constituer un premier jeu de données utile au plus grand nombre (le zonage des décisions de justice, voir 2.), et ce faisant de :

- démontrer la faisabilité et l'utilité de constituer des jeux de données d'entraînement de qualité ;
- explorer des méthodologies de création de données d'apprentissage, ainsi que les outils et compétences nécessaires ;
- documenter la démarche afin de lui conférer un caractère réplicable pour d'autres jeux de données.

1. Livre blanc IA 2016 – INRIA - https://www.inria.fr/actualite/actualites-inria/livre-blanc-sur-l-intelligence-artificielle
2. https://www.cnil.fr/fr/ethique-et-numerique-les-algorithmes-en-debat-0

Conformément aux principes érigés par l'association, le jeu de données produit est mis à la disposition de tous et exploitable dans un format ouvert.

Au terme de ces 6 mois d'expérimentation, un jeu de données d'environ 400 décisions a été réalisé, soit près de 39 000 paragraphes annotés[3]. Parmi les annotations effectuées, certaines sont exploitables en l'état, d'autres non ou nécessitent d'être retravaillées. Les enseignements méthodologiques en revanche, furent nombreux et précieux. C'est l'objet de ce livre blanc que de les rapporter.

3. Toutes les ressources issues de ce programme (documents de travail, méthodologie, code, jeu de données annoté, évaluation et publications) sont disponibles sur le site d'Open law : https://openlaw.fr/ressources-open-law

Pourquoi les données de jurisprudence ?

Si différents types de données avaient été initialement envisagées (on pense aux conventions collectives, ou encore aux contrats), ce sont finalement les décisions de justice qui ont été retenues dans le cadre de ce programme.

En effet, l'association venait alors de conclure un programme sur l'Open Data jurisprudentiel - Open Case Law -, qui avait permis de réunir l'ensemble des acteurs intéressés par cette matière première. Réfléchir à l'enrichir pour constituer un dataset d'apprentissage était un continuum logique, et permettait de faire vivre ce jeune écosystème autour d'un nouveau projet d'exploration tout en prolongeant la dynamique d'innovation collaborative initiée à cette occasion.

Les pré-requis du programme

Ouverture :

- Programme ouvert à tous
- Données brutes disponibles en open data et outil open source
- Livrable, code et documentation en format ouvert

Qualité et utilisabilité du livrable :

- Évaluation du jeu de données
- Volume et représentativité
- Possibilité d'utilisation directe

2

Le cas d’usage : zonage des décisions de justice

Lors d'un atelier où de nombreux cas d'usage ont été évoqués (voir encadré), ce fut finalement le zonage des décisions de justice qui a été retenu. Il s'agissait de repérer les différentes parties d'une décision de justice : présentation des parties, avocats ou magistrats, exposé des faits et de la procédure, exposé de la règle de droit applicable, argumentation du juge, solution...

Autres cas d'usage envisagés lors de l'atelier du 7 avril 2018

- Constitution d'un jeu de données juridiques neutre (sans surreprésentation de certaines catégories de la population)
- Identification des dépendances grammaticales entre les mots spécifiques au langage juridique
- Identification des phrases dans les décisions de justice (où la ponctuation n'est pas décisive)
- Anonymisation des décisions de justice
- Identification des rôles dans les décisions de justice : magistrat, avocat, greffier...
- Identification des montants et leurs informations associées dans les décisions de justice : fondement, brut ou net, en capital ou en rente mensuelle/annuelle...
- Classification des décisions d'appel ou de renvoi dans les documents de cours d'appel
- Identification des clauses dans les appels d'offres publics
- Identification des chiffres (métrages), noms (qui est exproprié), lieux dans les arrêtés des conseils municipaux
- Zonage et identification des montants et de leurs fondements dans les conventions collectives

Travailler sur les décisions de justice était une suite naturelle du programme Open Case Law, qui avait pris fin en décembre 2016[4]. Les données étaient disponibles et bien connues, les partenaires identifiés et motivés.
Par ailleurs, ce cas d'usage répondait au besoin exprimé par plusieurs membres de l'association, pour faciliter leurs travaux d'extraction d'informations dans les décisions de justice (montants, chance de réussite, concepts...). En effet, en distinguant les différentes parties d'une décision de justice, on parvient à lever les ambiguïtés liées à la présence de la même information à plusieurs endroits de la décision (par exemple, dans une décision d'appel, un montant sera présent 4 fois au moins : celui

4. https://openlaw.fr/sites/default/files/2018-07/Livret_blanc_interactif21_04_0.pdf

alloué en première instance, ceux demandés par chaque partie en appel, celui alloué par le juge). Ce zonage des décisions servira en outre la création de nombreux outils (voir encadré ci-dessous), et la méthodologie servira à d'autres types de données, pour lesquelles le zonage est également nécessaire : textes, conventions collectives, contrats, appels d'offres...

Exploitation du zonage - brainstorming issu des ateliers des 7 novembre et 7 décembre 2018

- Affichage de la décision : pouvoir afficher les arrêts avec un confort de lecture plus grand, en identifiant visuellement les zones et en améliorant la lisibilité des décisions de justice
- Sommaire de la décision : permettre une navigation dans la décision
- Moteur de faits : permettre la recherche d'une décision par ses faits
- Identification de la règle de droit : pouvoir rechercher une décision par la règle de droit appliquée / rapprocher des décisions sur la base de la règle de droit appliquée
- Chaînage des décisions d'une même procédure : identifier les différentes décisions de la procédure et naviguer de l'une à l'autre
- Extraction de montants : extraire les montants avec plus de précisions en utilisant les différentes zones identifiées (montants demandés par les parties, montants alloués en première instance...)
- Moteur de parties : pouvoir rechercher les parties personnes morales afin de connaître leur passé judiciaire

La structure des décisions étant différente d'une juridiction à une autre, il a été décidé de limiter le périmètre du projet à un seul niveau de juridiction. Les décisions de cassation étant bien normées, le choix a été fait de se concentrer sur les juridictions du fond, et plus précisément les cours d'appel, les jugements de première instance n'étant quasiment pas présente dans les fonds disponibles en open data.

3

Le choix de l’outil

1. Le choix de l'annotateur Brat

Le choix de l'outil d'annotation des textes est structurant pour le projet. Certains des outils disponibles permettent la pré-annotation par règle, d'autres des interfaces pour gérer des utilisateurs multiples. Les plus aboutis sont souvent des outils propriétaires. Outre les contraintes budgétaires, le choix d'outil sous licence open source a été fait pour favoriser la reproductibilité de la méthodologie.

Nos besoins étaient les suivants :

- pouvoir qualifier des paragraphes ;
- pouvoir travailler à plusieurs ;
- disposer d'annotations simples à exploiter ;
- pouvoir déployer facilement l'outil.

Après recherche, il est rapidement apparu que la plupart des outils n'étaient plus maintenus, et que d'autres impliquaient un déploiement lourd (serveur Tomcat, etc.) ou offraient des fonctionnalités très (voire trop) basiques.

La contrainte la plus rarement satisfaite était la possibilité de travailler à plusieurs sur le même jeu de données. Sans cela, nous aurions eu des allers retours entre les annotateurs et les responsables du projet en charge du contrôle qualité des annotations. Cela aurait sensiblement ralenti tout le processus.

Notre attention s'est finalement portée sur Brat (http://brat.nlplab.org/), très certainement le logiciel d'annotation le plus connu dans l'univers open source et le seul à notre connaissance à remplir notre cahier des charges.

2. Facile à déployer, difficile à utiliser

Notre première (bonne) surprise fut la simplicité de déploiement. La documentation est complète et la configuration est très simple.
Par ailleurs, comme nombre d'outils collaboratifs, Brat est en fait un site auquel on peut accéder via un classique navigateur. Pour ces outils, il faut souvent disposer d'un serveur logiciel de pages Web, qui requiert une configuration plus ou moins complexe. Brat contient entre autres un serveur basique préconfiguré.
Grâce à cela, le déploiement sur un serveur AWS (Amazon) n'a pris que quelques heures à une seule personne, découverte de l'outil incluse.
Après avoir chargé quelques décisions, nous nous sommes rapidement rendu compte que pour certaines d'entres elles, l'affichage avait de graves ratés sur les phrases les plus longues. Nous avons tenté de corriger le bug, mais la façon dont est écrit le logiciel a rendu cette tâche trop coûteuse en temps. Par pragmatisme, nous avons choisi de transformer nos données (en découpant les phrases longues) de sorte à ne plus être exposés à ce bug.

Après cela, les tests nous avaient paru concluants si ce n'est un détail… une petite difficulté à sélectionner le texte. La sélection sur cet outil dédié à l'annotation de mots et non de paragraphes, a la fâcheuse tendance à s'étendre au tout début document, même si ce n'est pas la volonté de l'utilisateur. Ce «détail» c'est révélé être extrêmement pénalisant, et a sensiblement réduit la productivité des annotateurs. Faute de pouvoir corriger le bug, nous avons mis en place des mesures de contournement qui ont rendu l'annotation moins pénible, mais aussi moins intuitive.

3. Quels outils pour de futurs projets d'annotation de paragraphes ?

Le choix d'un outil inadapté, car pénible à l'usage, s'est traduit par un ralentissement des annotations, et donc, à budget constant, une réduction de la taille du jeu de données constitué.

Le choix de l'outil est une question essentielle à aborder dès le départ du projet. Dans notre cas, nous avons peut être écarté certaines pistes trop vites.

La première était de développer nous même un outil d'annotation. Nous aurions ainsi pu parfaitement couvrir nos besoins (l'annotation de paragraphes entiers n'est pas une fonction que l'on trouve facilement) en consommant une partie des 6 mois alloués au projet. En revanche, nous aurions évité quelques longues séances de debuggages et surtout facilité la vie des annotateurs. Enfin, nous aurions pu ainsi développer un outil ad hoc, contenant des fonctionnalités spécifiques, comme par exemple un système de droits en fonction du «niveau» de l'annotateur.

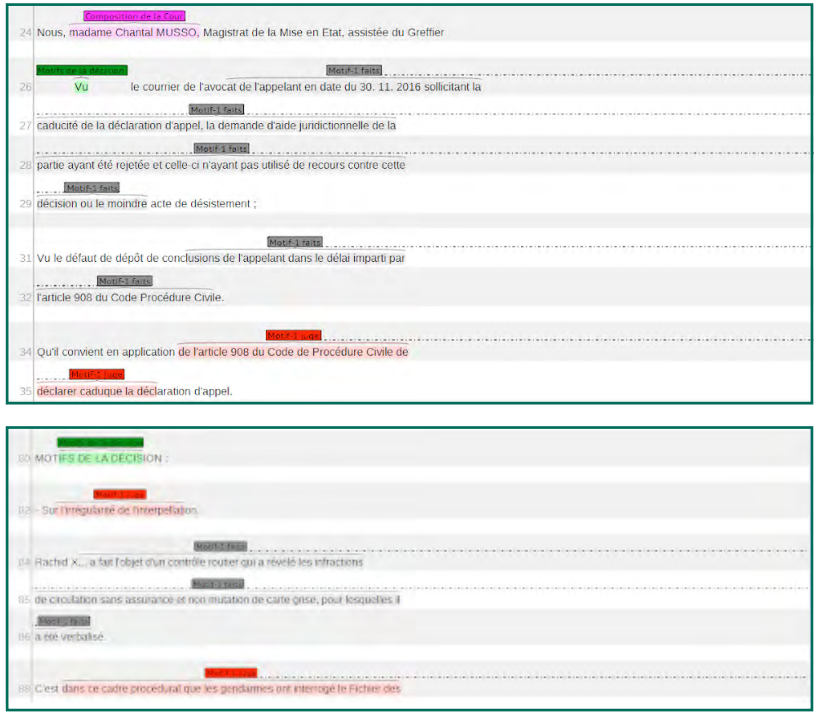
La seconde piste était d'utiliser un outil non dédié à l'annotation. Excel (ou son alter ego open source LibreOffice Calc) permet, en tant que tableur, d'afficher le texte de chaque paragraphe de chaque décision dans un tableau, et de mettre en face une colonne annotation. Nous avons depuis la fin du projet testé cette solution et elle nous a semblé plus agréable à l'usage. Elle n'aurait cependant pas permis une annotation collective. Pour cela, une solution cloud, du type Excel 360 ou la Google Spreadsheet aurait convenu (attention cependant aux fonctionnalités de tri et de recherche, nécessaires pour ce projet et très lentes dans les solutions cloud).
Enfin, la troisième piste était celle des outils propriétaires payants. Depuis le début du projet, de nombreux outils ont été mis sur le marché, certains incluant même la partie apprentissage (active learning). Sans être exhaustifs, on pourra citer par exemple Dataturk ou Prodi.gy.

Témoignages

«Au niveau de l'outil, le point négatif principal est qu'il fallait aller à deux endroits différents, l'un pour choisir son lot et le second pour faire les annotations. De plus, la compréhension de l'outil était assez complexe au départ.»

«En ce qui concerne le plan d'annotation, il était parfois difficile de distinguer les différentes catégories de motif (juge, règle de droit…), ce qui rallongeait le temps d'annotation d'une décision.»

Marie-Laure Bellamy-Brown et Alexia Dary, Rédaction technique, Editions Législatives



Captures d'écran de l'outil utilisé pour le programme IA & Droit - Brat - 2017

4

Une méthodologie d’annotation à parfaire

1. Le travail de création du plan d'annotation avec la Cour de cassation

Le plan d'annotation permet de normaliser la structure du document, en l'espèce la décision de justice. Sa qualité est essentielle, puisqu'un plan mal construit conduit à des annotations inutiles et finalement rend difficile l'apprentissage d'un modèle de machine learning.

Pour identifier les différentes catégories de notre plan d'annotation (les zones de la décision que l'on souhaite identifier), l'expertise du service de documentation, des études et du rapport (SDER) de la Cour de cassation a été décisive. Nous sommes partis des cours dispensés à l'ENM et l'avons comparé avec la réalité des arrêts pour constituer la trame des décisions et identifier les types de «zones» que les annotateurs devaient identifier.

Nous avons tout d'abord listé quatre zones principales, appelées macro_zones dans notre plan d'annotation :

- Entête : le début d'une décision de justice, qui contient en principe la présentation de la décision (date, juridiction, numéro identifiant, références de la décision attaquée...), des parties, des avocats, des membres de la Cour, des caractéristiques de la décision (ordonnance ou au fond, publique ou non, contradictoire ou non...) ;
- Exposé du litige : la présentation des éléments utiles du litige, c'est-à-dire l'exposé des faits, de la procédure, et des moyens et prétentions des parties ;
- Motifs de la décision : présentation des arguments retenus et discutés par la Cour et solution ;
- Dispositif : la dernière partie de la décision, qui décrit la solution du litige.

En parallèle de ce typage des paragraphes, nous avons souhaité annoter le lien entre chaque moyen soulevé et la «réponse» du magistrat, qu'elle soit dans les motifs ou le dispositif, en leur donnant un numéro commun (Motif-1 correspond à Dispositif-1, Motif-2 à Dispositif-2, etc).

Par ailleurs, dans la partie Entête, nous souhaitions identifier les paragraphes mentionnant les différents protagonistes (les parties, avocats, magistrats, etc.) et ceux contenant la référence de la décision attaquée. Dans la partie Exposé du litige, nous souhaitions distinguer les faits, la procédure et les prétentions de chaque partie. Dans la partie Motifs, nous avons tenté de distinguer les différentes prémisses du syllogisme judiciaire : la règle de droit, les faits, la conclusion. L'idée était de pouvoir, par exemple, identifier l'énoncé d'une règle de droit même lorsqu'aucun texte n'était expressément cité.

Nous avons également distingué d'autres informations, telles les «demandes accessoires» (dépens et frais d'avocat) dans les zones «Motifs» ou «Dispositif» et quelques catégories «hors zone», comme le caractère contradictoire ou non, le caractère public ou non ou les dates de l'audience (date de clôture, prorogations...).

Après quelques tests d'annotation, nous avons procédé à la rédaction des guidelines à destination des annotateurs. Ces guidelines devaient être le plus simple d'accès possible, tout en permettant de répondre au maximum de questions. Nous avons donc écrit deux documents :

- un document de prise en main rapide contenant les informations sur le fonctionnement de l'outil et des explications dans les grandes lignes sur chaque catégorie du plan d'annotation ;
- une FAQ contenant les réponses à toutes les questions que peuvent se poser les annotateurs.

C'est à ce stade que nous avons confié le travail aux annotateurs, en espérant les rendre le plus autonome possible tout en restant à leur disposition pour les précisions nécessaires.

2. Le retour à la réalité lors du travail d'annotation

Dès les premiers tests du groupe de travail en charge de la création du plan d'annotation, il est apparu que notre plan était trop théorique et complexe par rapport à la réalité de la rédaction des décisions de justice. Nous avons à ce stade supprimé certaines catégories lorsqu'à l'usage on s'apercevait que même des utilisateurs très experts (les concepteurs du plan d'annotation) ne parvenaient pas à les identifier dans nombre de décisions ou lorsqu'elle n'étaient pas ou peu présentes dans les données. Il en fut ainsi, par exemple, de la zone «publicité de la décision» : malgré son intérêt indéniable (pour, par exemple, filtrer les décisions à verser en open data de celles qui devaient rester non publiques), cette information était présente sous des formes et à des emplacements trop variés pour être repérée, sauf par une catégorie très précise d'experts (les magistrats et les greffiers).

Les difficultés lors de la phase d'annotation proprement dite furent d'un autre ordre. Notre objectif initial était de permettre au plus grand nombre de personnes de participer, de façon la plus autonome possible. En réalité, le coût d'entrée sur l'outil et le plan d'annotation était trop grand pour permettre un réel crowdsourcing. En effet, si les annotations prenaient en moyenne 5 minutes par décision, la prise en main prenait de l'ordre de 1h en moyenne (pour 1 ou 2 arrêts), en présence d'un formateur (en plus d'une introduction générale au machine learning). De plus, on s'est rapidement rendu compte que certains profils d'annotateurs étaient plus efficaces que d'autres, car habitués à passer en revue des arrêts sans les lire dans le détail (greffiers, services sources des éditeurs juridiques, magistrats, veilleurs dans les services juridiques…) et qu'à l'inverse la tâche était très difficile pour des avocats, habitués à analyser le fond des arrêts.

Par ailleurs, ces profils efficaces étaient aussi «meilleurs» que d'autres, car habitués à ce type de tâches nécessitant beaucoup de rigueur et donc qu'ils lisaient et respectaient scrupuleusement les guidelines. Le gros du travail a donc été effectué par un groupe d'annotatrices au sein des Editions Lefebvre Sarrut (ELS). De nombreux échanges ont ainsi pu avoir lieu, et le plan d'annotation et les guidelines furent constamment mis à jour au fur et à mesure de l'avancée du projet et complétés avec les nouveaux éléments soulevés.

Néanmoins, malgré ce travail de grande qualité, nous arrivons à la conclusion que nous avons été trop ambitieux. Le plan de classement était trop complexe, l'outil trop fastidieux à utiliser, ce qui, pour la majorité des catégories, n'a pas permis de constituer un jeu de données d'une quantité et d'une qualité suffisantes pour l'entraînement d'un modèle. Une reprise manuelle de l'ensemble du dataset a été nécessaire pour gommer certaines erreurs trop fréquentes, comme par exemple le fait d'inclure l'aide juridictionnelle dans la zone «Entete_avocat». Certaines catégories, sources de trop nombreuses approximations, ont lorsque c'était possible été fusionnées. Ainsi, les catégories «Faits» et «Procédure» furent fusionnées dans une catégorie unique «Faits et procédure».

Enfin, il s'avère - mais c'était aussi l'un des objectifs du projet que d'identifier cela - que certains de nos objectifs initiaux ne sont tout simplement pas atteignables par la classification. Ainsi en est-il de l'association entre un argument du juge dans la zone «Motifs» et sa reprise dans la zone «Dispositif» : l'association n'est pas du tout systématique dans un arrêt d'appel. En effet, elle est souvent remplacée par des phrases englobantes comme «infirme en toutes ses disposition» ou «confirme pour le surplus».

3. Ce que l'on en retient pour les prochains projets

On retiendra de cette expérience plusieurs choses.

Une tâche à la fois tu feras

Il vaut mieux faire plusieurs séances d'annotation, chacune concernant une seule catégorie, avec une réponse de type oui/non, que d'essayer de demander dans le même temps à l'annotateur de faire plusieurs tâches, surtout si elles sont de complexité différente. Dans le projet *IA et droit* saison 1, même si nous avions envisagé cet écueil et réduit considérablement nos ambitions à la baisse (il fut un temps envisagé de demander aux annotateurs, en plus du zonage, d'identifier les phrases au sein des décisions de justice !), nous avons réalisé un plan d'annotation trop complexe et trop long. En conséquence, nous ne sommes pas parvenus à un volume satisfaisant de décisions pour la majorité des catégories.

Pour le crowdsourcing simplissime tu resteras

Lors des ateliers réellement ouverts (*crowdsourcing*) du projet, on constate que seules certaines catégories (les macro-catégories Entête, Exposé du litige, Motifs et Discussion) ont été vraiment comprises et parfaitement annotées.
Pour de futurs projets, nous réserverons donc aux ateliers de *crowdsourcing* des tâches plus simples, comme l'identification des parties. Cette simplicité permettra de baisser la barrière à l'entrée en simplifiant les *guidelines*, le nombre d'opérations à retenir pour annoter, etc.
On pourra également faciliter la formation et les échanges entre les utilisateurs en instaurant un système de badges certifiants - dits *Open Badges* - donnant à des «niveaux» différents d'annotation, par exemple : un premier niveau de formation, où l'on annote des données déjà connues pour évaluer la qualité de l'annotation et fournir les guidelines ; un second niveau d'annotation ; un troisième niveau de modération et d'évaluation.
Pour cela, le développement d'un outil *ad hoc* sera nécessaire.

Pour des projets complexes, itératif tu seras

Pour des tâches plus complexes (comme par exemple l'identification de la règle de droit applicable ou celle des faits dans les motifs), une intégration des annotateurs dès la rédaction des *guidelines* aurait été bénéfique. Ce travail collaboratif aurait permis d'identifier plus rapidement les tâches «infaisables», les adaptations à réaliser, les difficultés récurrentes… Dans les faits, c'est ce qui s'est passé avec un petit groupe d'annotatrices, mais un peu tard dans la vie du projet.
A noter qu'il nous semble indispensable pour cela que les annotateurs comprennent le fonctionnement d'un algorithme de machine learning, afin d'éviter des annotations approximatives ou des plans d'annotations trop complexes (avec des rôles ou des relations, par exemple).

Pour des projets simples et rapides, solo tu travailleras

Certaines tâches ne nécessitant ni outil ni compétences particulières ni volume particulièrement grand de données auraient pu être effectuées par une seule personne et sans autre outil qu'un tableur. Il en sera ainsi, par exemple de l'identification de la composition de la Cour ou même des *macro_zones* (ce qui a été fait par la suite, pour compléter le *dataset*).

Dans tous les cas, un outil adapté tu fourniras

Quelle que soit la tâche choisie, il convient de bien s'assurer que l'annotateur dispose des outils (et, au travers des outils, des fonctionnalités) qui lui sont nécessaires. Ainsi, si pour identifier le paragraphe contenant les références de la décision attaquée, il doit pouvoir visualiser l'ensemble de la décision (et notamment les paragraphes précédents et suivants), alors l'outil devra lui permettre de visualiser du contexte. De même, si pour identifier les paragraphes concernant la publicité du débat, l'annotateur doit disposer d'un moteur de recherche pour trouver certains termes récurrent, il faudra le prévoir directement dans l'outil.

Et évidemment, une tâche utile tu demanderas

Enfin, cela va sans dire mais va encore mieux en le disant, il convient bien sûr de s'assurer que le travail demandé est réellement utile, sous peine d'"user" inutilement son vivier d'annotateurs. Ainsi, on s'attachera à ne pas constituer par annotation manuelle un *dataset* qui pourrait parfaitement et facilement l'être avec des règles.

Témoignages

Corine Dauchez Maître de conférences à l'Université Paris Nanterre, Diplômée notaire Membre du CEDCACE

Retour sur la formation au *Machine learning* : Atelier du 27 novembre 2017
L'algorithme n'est « rien » sans un jeu de données de qualité !

«A l'occasion d'un atelier «Machine learning» que j'avais organisé pour les étudiants du M2 Droit des technologies numériques et société de l'information de l'Université Paris Nanterre, Camille Le Douaron a formé les étudiants afin qu'ils puissent aider à la constitution d'un jeu de données destiné à terme à entraîner des algorithmes de zonage de décisions de justice. Camille a, entre autres, souligné l'importance de la quantité des données en entrée et, ce dont nous avions peut-être moins conscience, de la qualité des données collectées pour nourrir les algorithmes.

Aussi, l'annotation des décisions de justice, destinée à en extraire et classer les données pertinentes, doit reposer sur une « méthodologie de zonage » rigoureuse. Elle implique l'élaboration d'un plan d'annotation de la décision de justice qui doit être précis, sans être trop détaillé, et simple pour que les annotateurs (nos étudiants !) puissent aisément se l'approprier.

Nous avons ainsi pu participer concrètement à la phase primordiale de collecte humaine des données. A cette occasion, nous avons réalisé le degré d'exigence requis par ce travail d'annotateur dont le niveau d'expertise juridique devait notamment permettre de faire face à l'hétérogénéité structurelle des décisions de justice. L'atelier a été riche d'enseignements permettant de se rendre compte que la méthode, l'expertise et la rigueur étaient des qualités humaines indispensables au travail d'apprentissage des algorithmes !»

Témoignage de Brunos Mathis (extrait)

«Dans le cadre du programme IA et Droit, j'ai participé à l'exercice collaboratif de constitution d'un corpus d'apprentissage d'une annotation automatisée des décisions de justice. J'ai ainsi annoté plusieurs dizaines d'arrêts de Cour d'Appel.
Il s'agissait d'effectuer un zonage de chaque texte, c'est-à-dire d'annoter les fragments de texte significatifs, afin de les structurer en différents paragraphes ou sous-paragraphes.

Alors même qu'il était limité à un corpus d'arrêts d'appel, l'exercice montre bien le grand nombre de variantes dans la façon dont les greffes rédigent leurs décisions. Des différences apparaissent selon les thématiques ; par exemple, une décision en droit du travail est plutôt plus longue, et donne lieu à davantage d'annotations, qu'une autre en droit d'asile. Mais même à thématique égale, des différences sont visibles selon les ressorts. Cela tient sans doute au choix de templates que chaque greffe s'est constitué pour se faciliter la tâche, voire de son outillage bureautique.

L'ordre des rubriques, au sein du texte d'une décision, est globalement invariant. L'exposé du litige vient avant les motifs de la décision, qui précèdent eux-mêmes le dispositif, les prétentions de l'appelant sont décrites avant celles de l'intimé, etc… Cependant, les titres sont rares et c'est une grande variété de tournures de phrase qui peut marquer la transition entre deux rubriques. Par exemple, « par ses conclusions » ou « dans ses écritures », identifient pareillement un paragraphe de moyens et prétentions.

On devine ainsi qu'un processus de machine-learning permettra de localiser les transitions de rubriques, par-delà les nuances du langage et les erreurs de syntaxe, avec un meilleur niveau de qualité que ne le ferait un système basé sur des règles de gestion, nécessairement approximatives.»

Plan d'annotation du programme *IA & droit - données d'apprentissage*

- Exposé du litige
- Motifs de la décision
- Dispositif
- Références décision attaquée
- Appelant-demandeur
- Intimé-défendeur
- Avocat
- Composition de la Cour
- Faits et procédure
- Moyens et prétentions appelant
- Moyens et prétentions intimé
- Motif-1 (ou -2 ou -3) juge
- Motif-1 (ou -2 ou -3) règle de droit
- Motif-1 (ou -2 ou -3) faits
- Motif-1 (ou -2 ou -3) prétentions appelant
- Motif-1 (ou -2 ou -3) prétentions intimé
- Motif-demandes accessoires
- Dispositif-1 (ou -2 ou -3)
- Dispositif-demandes accessoires

5

L’évaluation de la qualité des annotations

1. Les méthodes d'évaluation mises en place

L'évaluation est un enjeu majeur de la constitution de *dataset* d'apprentissage. L'algorithme cherchant à définir le modèle sous-tendant les exemples qui lui sont soumis, il est absolument indispensable de fournir des données soit en très grand nombre, soit d'excellente qualité (cohérentes entre elles et représentatives de l'ensemble des données), idéalement les deux.

En matière juridique, nous ne pourrons que difficilement parvenir à un volume suffisant de données d'apprentissage pour nous passer d'un réel travail sur leur qualité. En effet, le nombre total de décisions de justice rendues par les différentes juridictions par an n'est pas suffisant pour nous permettre, en en prenant un échantillon, de créer un jeu de données d'apprentissage dans lequel les erreurs deviendraient quantité négligeable.

Pour étudier la qualité de notre jeu de données, il nous a fallu étudier d'une part la qualité des annotations elles-mêmes, et d'autre part la représentativité du jeu de données d'apprentissage vis-à-vis de l'ensemble des données disponibles.

Taille du dataset

C'est quoi un échantillon suffisant ? Quelle est la masse critique ?
On se pose cette question pour chaque projet impliquant de l'apprentissage supervisé. Mais la réponse dépend du projet : plus les données brutes sont cohérentes et le signal à capter est simple, moins il faudra de données d'apprentissage pour parvenir à une qualité satisfaisante.
Ainsi, pour le zonage des décisions de Cour de cassation, qui sont toujours rédigées de la même manière, une centaine de décisions suffiraient sûrement. Pour notre projet, les 400 décisions disponibles sont suffisantes pour les analyses simples (catégories macro notamment), mais ne le sont pas pour d'autres (identification de la règle de droit par exemple).

La qualité des annotations

Les méthodes d'évaluation de la qualité des annotations sont relativement variées et plutôt bien documentées. Une évaluation manuelle est évidemment toujours possible, mais on peut aussi automatiser les sondages via des techniques de comparaison des annotations, entre les annotateurs (calcul de l'interagreement) ou par rapport à un référentiel pré-établi et validé comme correct (que nous avons appelé « gold dataset »). Nous avons exploré chacune de ces méthodes.

La représentativité du jeu de données

Il est nécessaire d'évaluer la représentativité du jeu de données par rapport à l'ensemble des données disponibles ("données brutes", ou "données réelles"), car si le jeu de données d'apprentissage ne représente pas correctement l'ensemble des cas de figure présents dans les données brutes à enrichir, alors ces cas de figure ne seront pas pris en compte par le modèle.

Pour ce faire, nous avons décidé de prendre un échantillon de décisions de toutes les chambres de toutes les cours d'appel. En revanche, nous avons pris le parti de ne pas nous intéresser aux évolutions dans le temps de la structure des décisions. Dès lors, pour s'assurer de la représentativité de notre échantillon par rapport aux données réelles, nous avons sélectionné toutes les décisions sur une période donnée (en l'occurrence l'année 2016), partant du principe que la répartition des décisions d'une juridiction à une autre et d'une chambre à une autre est stable dans le temps.

Il convient donc de s'assurer de bien connaître le jeu de données global (statistiques de représentation géographique, temporelle, matérielle…) parallèlement au travail de constitution du jeu de données d'apprentissage, afin de pouvoir vérifier la représentativité de ce dernier.

2. Evaluation du dataset

Evaluation des annotations

Après évaluation des données, on constate une qualité générale des annotations plutôt bonne, même si une passe manuelle sur l'ensemble pour les fautes les plus répandues et les modifications de guidelines intervenues en cours de travail d'annotation fut nécessaire. En revanche, les méthodes d'évaluation basées sur la comparaison des annotations (interagreement et gold dataset) ne furent pas concluantes du fait du trop faible nombre de documents annotés (35 lots, soit seulement 70 documents comparés). Quelques lots ont été exclus du fait d'une qualité générale trop faible ou de l'irrégularité des annotations (documents annotés seulement en partie). Ces documents ont été regroupés sous le tag «annotation_difficulty : impossible» pour pouvoir être isolés.

Le plan d'annotation ayant été construit dans une étape précédente et, comme expliqué ci-dessus, sans connaissance suffisante des données, ce n'est que lorsque nous avons appliqué pour la première fois un apprentissage et que nous avons manuellement analysé les résultats que nous avons réalisé qu'il nous fallait raffiner le jeu de données et le plan d'annotation. Ainsi, nous avons procédé de façon itérative et avons corrigé le jeu de données en fonction des endroits où l'algorithme butait le plus. Tout le long de ce processus, nous avons tenté de rester vigilants et de n'apporter des corrections que lorsque nous comprenions la cause de la difficulté de l'algorithme, et que cette cause nous semblait légitime, c'est à dire que l'annotation était ambiguë, et qu'il nous fallait lever cette ambiguïté, par exemple en supprimant une catégorie du plan d'annotation, ou en en regroupant plusieurs, etc.. Le risque majeur de cette façon de faire est de biaiser le jeu de données et d'annoter que pour avoir les meilleurs performances, sans prise en compte du besoin juridique.

Ce qui s'est dégagé lors de cet exercice, c'est que notre plan d'annotation initial était trop complexe et déconnecté de la réalité, et qu'il nous fallait le simplifier. Une fois que le plan a été stabilisé, il a été appliqué sur de nouvelles données qui n'avaient jamais été annotées auparavant (test set), ce qui nous a permis de vérifier l'absence de biais forts dans l'annotation. Nous sommes pleinement conscients que cette méthodologie n'était pas parfaite, mais elle nous semble avoir permis de rendre le jeu de données davantage réutilisable.

Test de classification automatique

Nous avons ensuite effectué à partir de ces données un premier test de classification (*multiclass classification* avec *FastRtext*[5]). Les résultats sont plutôt encourageants : pour les macro-catégories, le bon type est trouvé en moyenne dans plus de 93% des cas, même pour les décisions «atypiques» comme les ordonnances.

paragraph_type_macro	nb_mots_moyen	nb_items	micro_recall	micro_precision	micro_f1
1 Entête	11	2513	96,74	98,94	97,83
2 Motif_de_la_decision	50	1714	96,5	93,92	95,19
3 Dispositif	17	758	89,45	99,56	94,23
4 Expose_litige	38	1564	93,29	88,42	90,79

Showing 1 to 4 of 4 entries Previous 1 Next

En moyenne, le bon type macro est trouvé en **95,01%** des **6549** paragraphes utilisés pour les tests.

Apprentissage des catégories macro
Source : Michaël Benesty pour ELS, https://els-rd.github.io/iaetdroit/

Pour les catégories de second niveau, les résultats sont plus aléatoires. Ainsi, alors que nous pensions que l'identification des parties, avocats, magistrats de la décision serait aisée, en réalité, ces informations sont bien souvent mélangées et les zones difficiles à dissocier (mais ceci est probablement dû au fait que l'ordre des paragraphes entre eux est important - le défendeur venant toujours après le demandeur - et doit donc pouvoir être résolu relativement facilement). A l'inverse, les références de la décision attaquée étant formulées de façon très variées mais toujours isolées du reste de la décision, les résultats sont plutôt bons (88% lors de notre premier test).

5. Disponible en open source via ce lien : https://github.com/pommedeterresautee/fastrtext

	paragraph_type_micro_cleaned	nb_mots_moyen	nb_items	micro_recall	micro_precision	micro_f1
1	Entete_avocat	21	189	91,01	93,99	92,48
2	Entete	6	686	88,78	90,49	89,63
3	Motif	53	1366	94,51	84,66	89,31
4	Moyens_et_pretentions	42	649	89,37	84,8	87,03
5	Entete_composition_de_la_cour	15	434	82,26	90,38	86,13
6	Dispositif_demandes_accessoires	20	151	80,79	91,04	85,61
7	References_decisions_attaquees	23	101	72,28	97,33	82,96
8	n_a	8	1564	84,21	77,2	80,55
9	Faits_et_procedure	38	824	76,7	84,49	80,41
10	Dispositif	20	332	69,58	75,99	72,64

Showing 1 to 10 of 12 entries Previous 1 2 Next

En moyenne, le bon type macro est trouvé en **83,81%** des **6549** paragraphes utilisés pour les tests.

Apprentissage des catégories de second niveau
Source : Michaël Benesty pour ELS, https://els-rd.github.io/iaetdroit/

Analyse de la représentativité du dataset

En choisissant de sélectionner les décisions à annoter de façon chronologique, toutes juridictions confondues, nous nous sommes assurés d'une bonne représentativité géographique et matérielle de notre dataset sans avoir besoin de faire un travail méthodique d'analyse des données brutes. Ce fut une erreur, car cela nous a empêché de remarquer un biais majeur dans ces données brutes, à savoir les données Légifrance. En effet, certaines juridictions versent dans Légifrance proportionnellement plus de décisions que d'autres juridictions. Il s'agit notamment des juridictions de Corse et des juridictions d'outre-mer (environ 20% du nombre total de décisions). Ces décisions en surnombre sont majoritairement des décisions de procédure, ce qui fausse encore notre dataset car elles ont une structuration qui leur est propre et à laquelle ne correspond pas notre plan d'annotation. Ainsi, par exemple, sur 63521 arrêts d'appel disponibles dans Légifrance, 3419 émanent de la cour d'appel de Bastia, soit exactement autant que de la cour d'appel de Rennes, pour un rapport de 1 à 15 en terme de population couverte[6] et de 1 à 6 en terme de nombre de nouvelles affaires par an[7].

Par ailleurs, même si c'est moins problématique car ces décisions n'ont de toute façon pas vocation à être manipulées, certains contentieux ne sont pas présents dans Légifrance (pénal, petits contentieux) et ne sont donc pas du tout représentés dans notre dataset.

6. Le ressort de la cour d'appel de Rennes couvre les 4 départements bretons et le département de la Loire Atlantique, soit environ 5 millions d'habitants. Le ressort de la cour d'appel de Bastia couvre les départements de la Corse, soit environ 330 000 personnes.
7. En 2017, le nombre d'affaires nouvelles était de 1678 pour la cour d'appel de Bastia et de 10 398 pour la cour d'appel de Rennes. Source : activité statistique des juridictions, données statistiques de la Justice, http://www.justice.gouv.fr/statistiques.html

3. Pistes d'amélioration

La première piste d'amélioration découle naturellement de celles évoquées aux précédents chapitres : une meilleure adéquation entre la tâche confiée aux annotateurs, leurs compétences et leurs disponibilités et l'outil qui est mis à leur disposition améliorera, de fait, la qualité des annotations.

Pour de véritables campagnes de *crowdsourcing*, à condition de s'assurer de disposer de suffisamment d'annotateurs, on pourra mettre en place une technique de vérification consistant à faire annoter le même item par 3 personnes différentes (voire 5 en cas de divergences sur les 3 premières). On pourra aussi prévoir un système d'évaluation par les annotateurs eux-mêmes, via un système de certification/modération.

Un autre axe d'amélioration consistera à appliquer le modèle au fur et à mesure de l'annotation, ce qui permet une évaluation en temps réel du *dataset*. Cette solution est rendue d'autant plus possible que des outils dits d'*active learning* sont aujourd'hui disponibles.

En ce qui concerne la représentativité du *dataset*, on préconisera un travail approfondi d'analyse des données brutes (ou «réelles»), afin d'en connaître le contenu et la répartition, mais aussi de ses biais intrinsèques : le jeu de données surreprésente la Corse par rapport aux autres ressorts, comme dans notre cas, ou les hommes blancs par rapport aux autres catégories, ou encore les votants par rapport aux abstentionnistes…

De l'importance de l'ouverture des données de jurisprudence

Aujourd'hui et dans l'attente de l'ouverture en open data des décisions de justice décidée par la loi pour une République numérique du 7 octobre 2016[8], les travaux sur la jurisprudence judiciaire ne portent que sur une petite portion du contentieux. En effet, même en comptant les données accessibles via des licences, seuls les contentieux d'appel et de cassation sont disponibles. Les travaux menés par les acteurs de la legaltech ne peuvent donc pas être représentatifs de la réalité de la jurisprudence française.

Ainsi, le taux d'appel sur les jugements prononcés en 2015 était de 67,8% sur les jugements des conseils de prud'hommes, de 21,4% sur les jugement de TGI, de 13,7% sur les jugements des tribunaux de commerce… et de 5,6% sur les jugement de TI[9]. Par essence, seuls les contentieux «importants» (financièrement ou symboliquement ou impliquant des catégories sociales aisées) font l'objet de recours. Sans oublier le «facteur avocat», la représentation par avocat n'étant pas toujours obligatoire en 1ère instance, mais l'étant en revanche en appel.

Juridiction	2011	2012	2013	2014	2015
Tribunal de grande instance sur jugements en premier ressort	18,7	19,7	20,8	21,4	21,4
Tribunal d'instance	5,1	5,3	5,1	5,9	5,6
Conseil de prud'hommes sur jugements en premier ressort	64	67	67,7	68,3	67,8
Tribunal de commerce sur jugements en premier ressort	12,8	13,2	13,7	14,7	13,7

Taux d'appel des jugements prononcés au fond
Source : Ministère de la Justice/SG/SEM/SDSE/Exploitation statistique du Répertoire Général Civil

8. https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000033202746&categorieLien=id
9. Sources : chiffres-clés de la Justice 2017

6

Bibliographie

Bibliographie

Sur l'apprentissage automatique

BRAUNSCHWEIG Bertrand, 2016, Intelligence Artificielle, les défis actuels et l'action d'Inria, *Livre blanc INRIA* [en ligne]
Disponible sur : https://www.inria.fr/ [consulté le 11 novembre 2018]

CHU Tony et YEE Stéphanie, 2015. L'apprentissage automatique en images [en ligne]
Disponible sur : www.R2D3.us [consulté le 11 novembre 2018]

Sur la constitution des jeux de données d'apprentissage

ALTEXSOFT, 2018, How to Organize Data Labeling for Machine Learning: Approaches and Tools [en ligne]
Disponible sur : www.altexsoft.com [consulté le 11 novembre 2018]

ORAIN Grégoire, 2017, Sur Amazon Turk, les forçats du clic, *Le Monde*, Edition du 22 mai 2017.

RUETTE, Thomas, 2014, What is inter-annotator agreement [en ligne]
Disponible sur : https://corpuslinguisticmethods.wordpress.com/ [consulté le 11 novembre 2018]

Sur les enjeux éthiques de l'apprentissage automatique

VILLANI, Cédric, 2018, Donner un sens à l'intelligence artificielle. Pour une stratégie nationale et européenne. Rapport de la mission Villani [en ligne]
Disponible sur : https://www.aiforhumanity.fr/ [consulté le 11 novembre 2018]

DEMIAUX, Victor et SI ABDALLAH, Yacine, 2017, Comment permettre à l'Homme de garder la main ? Rapport sur les enjeux éthiques des algorithmes et de l'intelligence artificielle, *Rapport CNIL* [en ligne]
Disponible sur : www.cnil.fr/fr [consulté le 11 novembre 2018]

ANGWIN Julia, KIRCHNER Lauren, LARSON Jeff and MATTU Surya, 23 may 2016, Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks [en ligne]
Disponible sur : www.propublica.org/ [consulté le 11 novembre 2018]

O'NEIL Cathy, 2016, Weapons of Math Destruction, *Broadway books*.

HARDT Moritz, VIÉGAS Fernanda and WATTENBERG Martin, 2016, Attacking discrimination with smarter machine learning. An interactive visualization, *research.google.com* [en ligne]
Disponible sur : research.google.com [consulté le 11 novembre 2018]

Executive Office of the President, 2016, Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights, *The White House* [en ligne]
Disponible sur : obamawhitehouse.archives.gov/ [consulté le 11 novembre 2018]

Les partenaires et participants

Rédaction et édition : Camille Le Douaron et Michaël Benesty

Conception graphique et maquette : Mathilde Roussillat Sicsic

Les données statistiques de l'enquête sont disponibles en Open data sur la page ressources d'Open Law à l'adresse suivante : https://openlaw.fr/ressources-open-law

Cet ouvrage est disponible en version numérique sur le site www.openlaw.fr

Diffusion sous licence Cc by SA 4.0