

Méthodes d'analyse sémantique de corpus de décisions jurisprudentielles

Soutenance de thèse

Gildas TAGNY NGOMPÉ

24 janvier 2020

Jury:

- Stéphane MUSSARD, Professeur, Université de Nîmes (Directeur de thèse)
- Jacky MONTMAIN, Professeur, IMT Mines Alès (Co-directeur de thèse)
- Sandra BRINGAY, Professeur, Université Paul Valéry Montpellier (Rapporteur)
- Mohand BOUGHANEM, Professeur, Université Toulouse III Paul Sabatier (Rapporteur)
- Françoise SEYTE, Maître de Conférences (HDR), Université de Montpellier (Examineur)
- Fabrice MUHLENBACH, Maître de Conférences, Université Jean Monnet de Saint-Étienne (Examineur)
- Guillaume ZAMBRANO, Maître de Conférences, Université de Nîmes (Encadrant de proximité)
- Sébastien HARISPE, Maître Assistant, IMT Mines Alès (Encadrant de proximité)



1. Introduction
2. Annotation des sections et entités judiciaires
3. Identification des demandes des parties
4. Identification du sens du résultat
5. Découverte des circonstances factuelles
6. Conclusions

1. Introduction

1.1 Contexte

1.2 Objectif de la thèse

2. Annotation des sections et entités judiciaires

3. Identification des demandes des parties

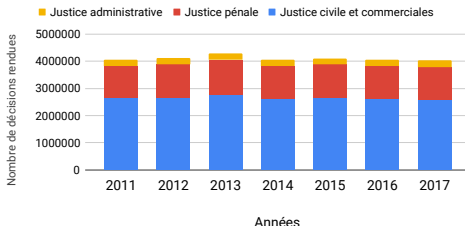
4. Identification du sens du résultat

5. Découverte des circonstances factuelles

6. Conclusions

Motivations

- La jurisprudence est analysée par les juristes pour comprendre l'application de la loi
- Difficultés de l'analyse manuelle
 - Existence d'un gros volume de décisions



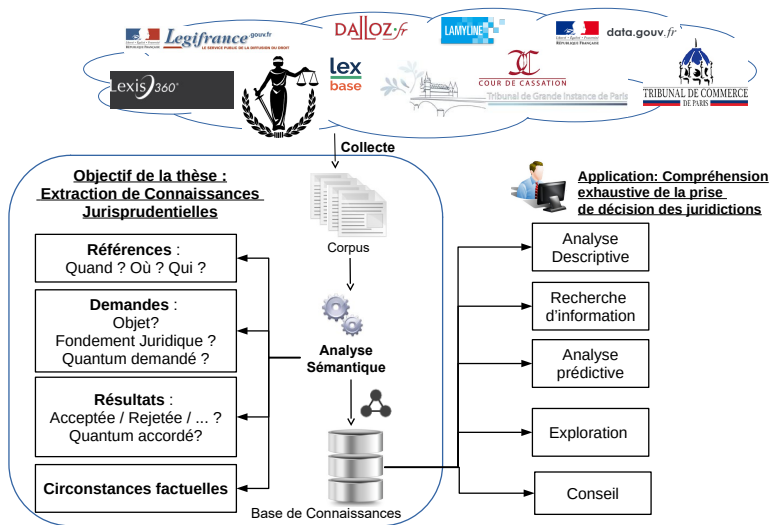
- Les moteurs de recherche juridique limités :
 - Pas de critère de recherche sémantique (catégorie de demandes, type de faits, etc.)
 - pas d'analyse synthétique de corpus

Activités en analyse automatique de décisions judiciaires

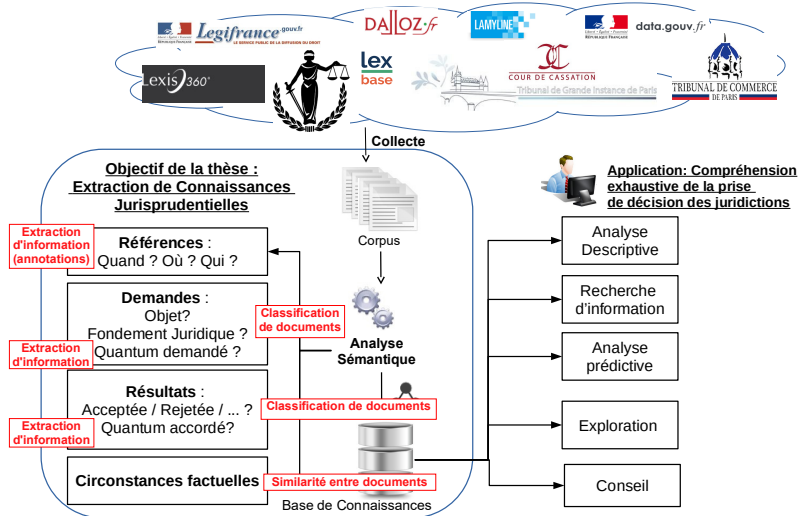
- Extraction d'information dans les décisions
 - Entités juridiques [Waltl et al., 2016, Andrew and Tannier, 2018]
 - Faits [Wyner, 2010, Wyner and Peters, 2010, Shulayeva et al., 2017]
 - Définitions de concept juridiques [Waltl et al., 2016, Waltl et al., 2017]
 - Arguments [Moens et al., 2007]
- Classification de décisions
 - Prédiction des décisions de justice [Ashley and Brüninghaus, 2009, Aletras et al., 2016]
 - Identification de la formation et la période [Şulea et al., 2017b, Şulea et al., 2017a]
 - Identifier la sentence prononcée [Ma et al., 2018]
- Similarité entre décisions
 - Décisions qui citent les mêmes lois et précédents [Nair and Wagh, 2018]
 - Similarité basée sur la question discutée et les faits sous-jacents [Kumar et al., 2011]
 - Recherche d'affaires antérieures pertinentes [Thenmozhi et al., 2017]
 - Classification supervisée des décisions [Ma et al., 2018]
 - Regroupement non-supervisé [Ravi Kumar and Raghuv eer, 2012]

1.2 Objectif de la thèse

Tâches et exemples d'applications



Formulation en analyse de données textuelles



1. Introduction

2. Annotation des sections et entités judiciaires

2.1 Objectif de la tâche

2.2 Approches probabilistes de détection d'entités

2.3 Sélection de modèles

2.4 Discussions des résultats

3. Identification des demandes des parties

4. Identification du sens du résultat

5. Découverte des circonstances factuelles

6. Conclusions

2.1 Objectif de la tâche

Détecter les méta-données de référence et les normes

Cour d'appel
Lyon
6e chambre
17 Mars 2016
Répertoire Général : 14/06777
APPELANTE :
Mme Monique V. ...
Représentée par Me Chrystelle P. , avocat au ...
INTIMES :
Mme Sylvianne C. ...
Composition de la Cour ... :
- Claude VIEILLARD , président ...
FAITS, PROCÉDURE, MOYENS ET ...
Suite à un prêt de 10.000 € ...
Par jugement en date du 4 avril 2013, ...
Dans leurs conclusions ..., Mme Sylvianne C. , M. ...
demandent à la cour de :
- condamner Mme V. à leur payer ... au titre de l'article
700 du code ... , ...
MOTIFS DE LA DÉCISION
La cour constate au préalable que le jugement n' est pas
remis en causes ...
...
La Cour estime par contre que ... application
de l' article 700 du code de procédure civile en cause d'
appel au profit des

▼ Original markups

- ☒ appelant
- ☒ avocat
- ☐ corps
- ☒ date
- ☐ decision
- ☐ dispositif
- ☐ entete
- ☒ fonction
- ☒ formation
- ☒ intime
- ☒ juge
- ☒ juridiction
- ☐ litige
- ☐ motifs
- ☒ norme
- ☒ rg
- ☒ ville

2.1 Objectif de la tâche

Sectionner pour organiser l'extraction des connaissances

Cour d'appel

Lyon

6e chambre

17 Mars 2016

Répertoire Général : 14/06777

APPELANTE :

Mme Monique V. ...

Représentée par Me Chrystelle P. , avocat au

INTIMES :

Mme Sylvianne C. ...

Composition de la Cour ... :

- Claude VIEILLARD , président ...

Entête : méta-données de référence de l'affaire

PAR CES MOTIFS

La Cour , ...

Confirme le jugement entrepris en toutes ses dispositions sauf en ce qu' il a ...

Statuant à nouveau de ce chef ,

Condamne Mme Monique V. à payer ... aux consorts C. la somme de

MILLE euroS (1.000 €) au titre de l' **article 700 du code de procédure civile** ...

Dispositif : résultats et normes

FAITS, PROCÉDURE, MOYENS ET ...

Suite à un prêt de 10.000 € ...

Par jugement en date du 4 avril 2013, ...

Dans leurs conclusions ..., Mme Sylvianne C. , M. ... demandant à la cour de :

- condamner Mme V. à leur payer ... au titre de l'**article 700 du code ...** , ...

Litige : normes, faits, jugements antérieurs, prétentions et arguments des parties

MOTIFS DE LA DÉCISION

La cour constate au préalable que le jugement n' est pas remis en causes ...

...

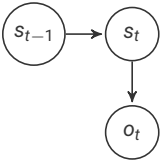
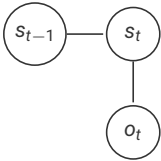
La Cour estime par contre que ... application de l' **article 700 du code de procédure civile** en cause d' appel au profit des intimés et il convient de leur allouer à ce titre la somme de **1.000 €** .

Motifs : normes, raisonnement des juges, réponses des juges

Corps

2.2 Approches probabilistes de détection d'entités

Modèles probabilistes d'étiquetage de séquences

HMM	CRF
un seul descripteur par observation	plusieurs descripteurs complexes par observation
	
$P(S, O) = \prod_{t=1}^T P(s_t s_{t-1}) P(o_t s_t)$ <p>[Seymore et al., 1999]</p>	$P_{\lambda}(S O) = \frac{1}{Z(O)} \exp \left(\sum_{t=1}^T \sum_k \lambda_k f_k(s_{t-1}, s_t, o_t) \right)$ <p>[Peng and McCallum, 2006]</p>

2.3 Sélection de modèles

Méthodes explorées

- Définition de descripteurs de lignes pour les sections :
 - forme : mots, longueur, etc.
 - contexte : position, lignes voisines, etc.
- Définition de descripteurs de mots pour les entités :
 - forme : est-ce une initiale ("B.")?, un mot clé de norme?, etc.
 - contexte : mots voisins, position, etc.
 - syntaxe : rôle grammaticale
- Réduction du nombre de descripteurs
 - recherche bidirectionnelle (BDS) [Liu and Motoda, 2012]
 - sélection séquentielle avant à flottement (SFFS) [Pudil et al., 1994]
- Sélection du schéma d'étiquetage

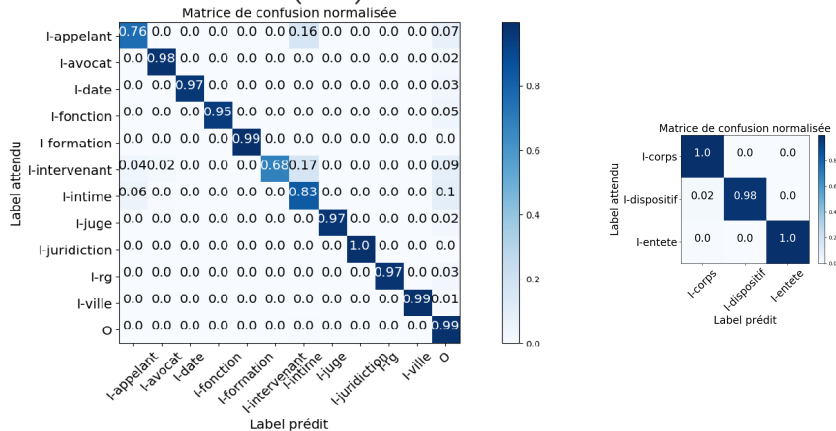
	<i>composée</i>	<i>de</i>	<i>Madame</i>	<i>Martine</i>	<i>JEAN</i>	,	<i>Président</i>	<i>de</i>	...
IO	0	0	I-JUGE	I-JUGE	I-JUGE	0	I-FONCTION	I-FONCTION	...
BIO	0	0	B-JUGE	I-JUGE	I-JUGE	0	B-FONCTION	I-FONCTION	...
IEO	0	0	I-JUGE	I-JUGE	E-JUGE	0	I-FONCTION	I-FONCTION	...
BIEO	0	0	B-JUGE	I-JUGE	E-JUGE	0	B-FONCTION	I-FONCTION	...

Résultats (CRF)

- sélection du schéma d'étiquetage
 - Les schémas plus complexes que IO rendent l'entraînement plus long
 - Les schémas complexes ne semblent pas améliorer la détection des sections (baisse de F_1 de près de 20%)
 - Les schémas complexes améliorent légèrement la détection d'entités de moins de 3%
- sélection des descripteurs
 - Lenteur des algorithmes BDS et SFFS
 - BDS réduit de moitié
 - SFFS réduit beaucoup plus
 - Pas d'amélioration ou détérioration considérable de la détection

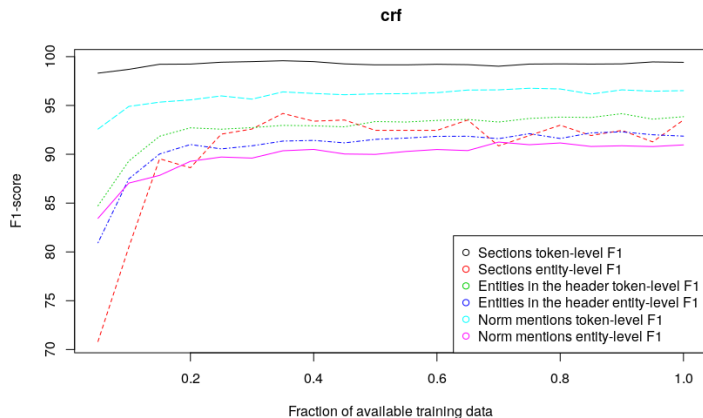
2.4 Discussions des résultats

Confusions de labels (CRF)



2.4 Discussions des résultats

Impact de la quantité de décisions d'entraînement



2.4 Discussions des résultats

Description manuelle vs. représentation apprise

	CRF + descripteurs manuels			BiLSTM-CRF		
	<i>Precision</i>	<i>Rappel</i>	F_1	<i>Precision</i>	<i>Rappel</i>	F_1
appellant	82.49	69.42	74.72	80.26	71.53	75.04
avocat	90.15	89.02	89.56	84.93	87.88	86.36
date	95.34	91.46	93.12	95.04	90.79	92.63
fonction	95.87	95.08	95.44	92.69	93.48	93.03
formation	96.91	91.31	93.7	91.05	89.47	89.84
intervenant	51.42	32.71	36.8	31.48	20	23.11
intime	76.01	79.15	77.22	67.7	75.43	70.83
juge	95.67	94.07	94.84	95.44	95.56	95.46
juridiction	98.55	98.25	98.33	97.95	99.22	98.57
rg	95.46	95.29	95.27	91.13	97.26	93.92
ville	98.33	93.01	94.71	91.43	95.34	93.3
norme	91.08	90.27	90.67	91.43	92.65	92.03
Evaluation globale	92.2	90.09	91.12	89.21	90.43	89.81

Sectionnement en 4 sections pour l'extraction des demandes

	CRF (%)		
	<i>Precision</i>	<i>Rappel</i>	F_1
entete	99.80	99.54	99.67
litige	96.10	97.66	96.87
motifs	97.31	95.96	96.62
dispositif	99.00	98.49	98.72
Evaluation globale	97.55	97.55	97.55

1. Introduction

2. Annotation des sections et entités judiciaires

3. Identification des demandes des parties

3.1 Objectif de la tâche

3.2 Méthode proposée : approche par catégorie de demandes

3.3 Résultats expérimentaux

4. Identification du sens du résultat

5. Découverte des circonstances factuelles

6. Conclusions

3.1 Objectif de la tâche

Extraction de données sur les demandes des parties

- Exemple : demande de dommage-intérêts pour procédure abusive

- Extrait de décision relatif à la demande :

Jennifer M. et Catherine M. ... demandent à la Cour de :

- **infirmen le dit jugement** en **toutes ses dispositions**; ...

Statuant à nouveau ...

- les condamner au paiement d'une somme de **3 000,00 € pour procédure abusive** et aux entiers dépens; ...

La cour ... CONFIRME **le jugement entrepris** en **toutes ses dispositions**.

Légende : référence au jugement antérieur en **rouge**, énoncés fusionnés en **bleue**

- Données à extraire

IDENTIFICATION DE LA DECISION			DESCRIPTION DE LA PRETENTION			DESCRIPTION DU RESULTAT	
Type ▾	Ressort ▾	RG ▾	OBJET ▾	NORME ▾	QUANTUM DEMANDE ▾	SENS DU RESULTAT ▾	QUANTUM RESULTAT ▾
CA	Saint Denis	14/01082	dommages et intérêts	1382 code civil + 32-1 code de procédure civile	3,000.00 €	rejette ▾	0.00 €

- Difficultés

- Présence de plusieurs demandes de catégories similaires et/ou différentes dans une même décision
 - Toutes les catégories ne sont pas connues d'avance (+500 catégories)
 - Difficile d'annoter une base d'évaluation pour toutes les couvrir

3.2 Méthode proposée : approche par catégorie de demandes

Extraction d'une seule catégorie (c) à l'aide de sa terminologie

- Détection de la présence de la catégorie par classification de la décision (c vs. \bar{c})
- Identification des quantas : montants à proximité des termes-clés de c dans les énoncés explicites de demandes et résultats

Jennifer M. et Catherine M. ... demandent à la Cour de :
- infirmer le dit jugement en toutes ses dispositions; ...
Statuant à nouveau ...
- [les condamner au paiement d' une somme de 3 000,00 € pour
procédure abusive et aux entiers dépens;]_{demande_danais} ...
La cour ... CONFIRME le jugement entrepris en toutes ses dispositions.

- Identification du sens du résultat
 - soit en fonction du verbe introductif de l'énoncé du résultat

Résultat (par polarité)		
accepte	sursis à statuer	rejette
<i>accorde, admet, condamne, ...</i>	<i>réserve, surseoit, ...</i>	<i>déboute, rejette, ...</i>

- soit "rejette" si pas d'énoncé explicite du résultat
- Mise en correspondance des informations relatives à la même demande
 - énoncé demande et énoncé résultat similaires
 - quantum demandé et quantum accordé apparaissant dans le même ordre

Apprentissage de la catégorie à extraire

- Entraînement d'un algorithme de classification pour détecter sa présence
- Détermination automatique de sa terminologie à l'aide d'une méthode de pondération de termes :

- $idf(t) = \log_2 \left(\frac{N}{N_t} \right)$ [Sparck Jones, 1972]
- $\Delta_{DF}(t, c) = DF_{t|c} - DF_{t|\bar{c}}$
- $\chi^2(t, c) = \frac{N((N_{t,c}N_{\bar{t},\bar{c}}) - (N_{t,\bar{c}}N_{\bar{t},c}))^2}{N_t N_{\bar{t}} |D_c| |D_{\bar{c}}|}$ [Schütze et al., 1995]
- $ngl(t, c) = \frac{\sqrt{N}(N_{t,c}N_{\bar{t},\bar{c}}) - (N_{t,\bar{c}}N_{\bar{t},c})}{\sqrt{N_t N_{\bar{t}} |D_c| |D_{\bar{c}}|}}$ [Ng et al., 1997]
- etc.

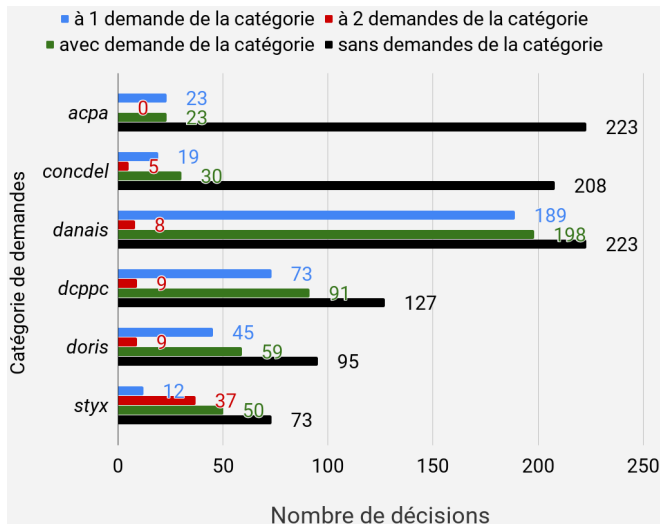
3.3 Résultats expérimentaux

Catégories sur lesquelles la méthode a été expérimentée

Label	Nom	Objet	Fondement
<i>acpa</i>	amende civile pour abus de procédure	amende civile	Articles 32-1 code de procédure civile + 559 code de procédure civile
<i>concdel</i>	dommages-intérêts pour concurrence déloyale	dommages-intérêts	Article 1382 du code civil
<i>danais</i>	dommages-intérêts pour abus de procédure	dommages-intérêts	Articles 32-1 code de procédure civile + 1382 code de procédure civile
<i>dcppc</i>	déclaration de créance au passif de la procédure collective	déclaration de créance	L622-24 code de commerce
<i>doris</i>	dommages-intérêts pour trouble de voisinage	dommages-intérêts	principe de responsabilité pour trouble anormal de voisinage
<i>styx</i>	frais irrépétibles	dommages-intérêts	Article 700 du code de procédure civile

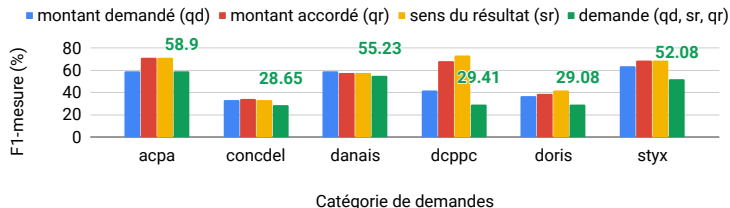
3.3 Résultats expérimentaux

Quantité de décisions annotées manuellement par l'expert



Efficacité de la méthode

- Les algorithmes traditionnels (KNN, SVM, naïf bayésien, arbre) donnent de bons résultats pour la détection de la catégorie : $98.8\% \leq F_1 \leq 100\%$
- Extraction :



- Le résultat est plus accessible
- Certaines catégories (*acpa*, *danais*, *styx*) sont plus accessibles que d'autres (*concdel*, *dcppc*, *doris*)
- Sources d'erreurs :
 - Difficulté à identifier les termes-clés rares
 - Absence de certains quanta dans les énoncés de demandes et résultats
 - Erreur de mise en correspondance des données extraites

1. Introduction

2. Annotation des sections et entités judiciaires

3. Identification des demandes des parties

4. Identification du sens du résultat

4.1 Contexte

4.2 Méthodes proposées : adaptations de la régression Gini-PLS

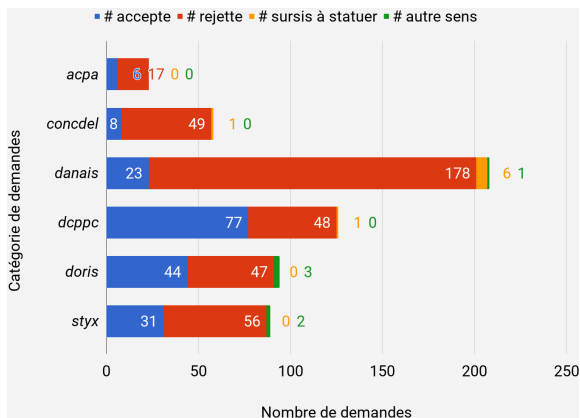
4.3 Résultats expérimentaux

5. Découverte des circonstances factuelles

6. Conclusions

Restriction du problème d'identification des demandes

- Uniquement les décisions à une demande de la catégorie
 - Raison : plus de 50% des documents dans la majorité des catégories
- Classification binaire (éviter les subtilités de rédaction)
 - Raison : les demandes sont en majorité **acceptées** ou **rejetées**



Plusieurs algorithmes classiques existent

- Classifieur bayésien naïf [Duda et al., 1973]
- K-plus-proches-voisins [Cover and Hart, 1967]
- SVM [Vapnik, 1995]
- Arbre de décision
- Analyse discriminante linéaire [Fisher, 1936] et quadratique [McLachlan, 1992]
- Logit-PLS [Tenenhaus, 2005]
- NBSVM [Wang and Manning, 2012]
- fastText [Grave et al., 2017]
- etc.

Régression Gini-PLS

- PLS [Wold, 1966] (régression partielle des moindres carrés)
 - Réduction supervisée de dimensions x_1, x_2, \dots, x_m en composantes orthogonales t_1, \dots, t_h

$$t_h = \sum_{k=1}^m w_{hk} \cdot \hat{U}_{(h-1)k}$$

avec $\hat{U}_{0k} = x_k, \forall h > 0, \hat{U}_{hk}$ est le résidu de la régression de x_k sur t_1, \dots, t_{h-1}

et $w_{hj} = \frac{\text{cov}(\varepsilon_h, \hat{U}_{(h-1)j})}{\sqrt{\sum_{j=1}^m \text{cov}^2(\varepsilon_h, \hat{U}_{(h-1)j})}}$ (solution de $\max \text{cov}(y, w_h X)$ s.c. $\|w_h\| = 1$)

- Régression de la variable dépendante y dans l'espace réduit

$$y = c_1 t_1 + \dots + c_h t_h + \varepsilon_h$$

- Gini-PLS [Mussard and Souissi-Benrejab, 2018]
 - Remplacement de la covariance $\text{cov}(x_j, y)$ par la covariance de Gini $\text{cog}(y; x_j) := \text{cov}(y; R(x_j))$
 - Élimination de la sensibilité du PLS aux valeurs aberrantes

Algorithmes proposés (en cours de rédaction pour la revue **Stats**)

○ Gini-PLS généralisé

- Utilisation de l'opérateur co-Gini généralisé :

$$\text{cog}_\nu(x_\ell, x_k) := -\nu \text{cov}(x_\ell, r_k^{\nu-1}); \nu > 1$$

où r_k est le vecteur rang décroissant de la variable x_k

- ν contrôle le compromis entre l'atténuation de la variabilité des variables ($\nu \rightarrow 1$) et l'influence des queues de distributions ($\nu \rightarrow \infty$)

○ Logit-Gini-PLS généralisé :

- Estimation des w_{hj} à partir du paramètre β de

$$P(y_i = 1/X = X_i) = \frac{\exp\{X_i\beta\}}{1+\exp\{X_i\beta\}}, \text{ où } X_i \text{ étant une observation}$$

$$w_j = \frac{\beta_j}{\|\beta\|}$$

- les $\hat{U}_{(h)j}$ toujours estimés avec le co-Gini généralisé

Comparaison des classifieurs PLS aux classifieurs classiques

Représentation	Algorithme	F_1	$F_{1_{\text{arbre}}} - F_1$
$tf - gss$	Arbre	0.668	0
$tf - avg_{global}$	LogitPLS	0.648	0.02
$tf - avg_{global}$	StandardPLS	0.636	0.032
$tf - \Delta_{DF}$	GiniPLS	0.586	0.082
$tf - \Delta_{DF}$	GiniLogitPLS	0.578	0.09
-	NBSVM	0.494	0.174
-	fastText	0.412	0.256

4.3 Résultats expérimentaux

Amélioration de la classification par restriction du document

Catégorie	Zone	Représentation	Algorithme	F_1
<i>acpa</i>	demande_resultat_a_resultat_context	<i>tf</i> – <i>dbidf</i>	Arbre	0.846
	<i>litige_motifs_dispositif</i>	<i>tf</i> – <i>dbidf</i>	StandardPLS	0.697
	<i>litige_motifs_dispositif</i>	<i>tf</i> – <i>avg_{global}</i>	LogitPLS	0.683
<i>concdel</i>	litige_motifs_dispositif	<i>tf</i> – <i>gss</i>	Arbre	0.798
	<i>motifs</i>	<i>tf</i> – <i>idf</i>	GiniLogitPLS	0.703
	<i>context</i>	<i>logave</i> – <i>dbidf</i>	StandardPLS	0.657
<i>danais</i>	demande_resultat_a_resultat_context	<i>avg_{local}</i> – χ^2	Arbre	0.813
	<i>demande_resultat_a_resultat_context</i>	<i>atf</i> – <i>avg_{global}</i>	LogitPLS	0.721
	<i>demande_resultat_a_resultat_context</i>	<i>atf</i> – <i>avg_{global}</i>	StandardPLS	0.695
<i>dcppc</i>	demande_resultat_a_resultat_context	<i>tf</i> – χ^2	Arbre	0.985
	<i>demande_resultat_a_resultat_context</i>	<i>tf</i> – χ^2	LogitPLS	0.94
	<i>litige_motifs_dispositif</i>	<i>tp</i> – <i>mar</i>	StandardPLS	0.934
<i>doris</i>	litige_motifs_dispositif	<i>tp</i> – <i>dsidf</i>	GiniPLS	0.806
	<i>litige_motifs_dispositif</i>	<i>tp</i> – <i>dsidf</i>	GiniLogitPLS	0.806
	<i>litige_motifs_dispositif</i>	<i>atf</i> – <i>ig</i>	StandardPLS	0.772
<i>styx</i>	motifs	<i>tf</i> – <i>dsidf</i>	Arbre	1
	<i>demande_resultat_a_resultat_context</i>	<i>logave</i> – <i>dsidf</i>	GiniLogitPLS	0.917
	<i>litige_motifs_dispositif</i>	<i>tf</i> – <i>rf</i>	GiniPLS	0.833

1. Introduction

2. Annotation des sections et entités judiciaires

3. Identification des demandes des parties

4. Identification du sens du résultat

5. Découverte des circonstances factuelles

5.1 Objectif de la tâche

5.2 Méthode

5.3 Sélection de la représentation des décisions

5.4 Efficacité de la méthode proposée

6. Conclusions

Déterminer les situations distinctes où sont formulées les demandes d'une catégorie données

- Exemple : catégorie "action en responsabilité civile professionnelle contre les avocats" (*arcpa*)
 - cas *a* : un avocat négligent qui envoie son assignation de manière tardive ;
 - cas *b* : un avocat qui n'a pas donné un conseil opportun, qui n'a pas soulevé le bon argument ;
 - cas *c* : un avocat qui n'a pas rédigé un acte valide ou réussi à obtenir un avantage fiscal ;
 - cas *d* : un avocat attaqué par son adversaire et non par son propre client.
- Formulation comme un problème de regroupement non supervisé des décisions

Apprentissage et utilisation d'une distance basée sur la transformation d'un document en un autre

- Formulation de la distance pour un ensemble de modifications connues

$$Dis_M(d, d') = f(M_{(d, d')}) = \frac{\sum_{(d[k], d'[k]) \in M_{(d, d')}} Dis_{cos}(\vec{d[k]}, \vec{d'[k]})}{|d|}$$

- Génération d'un corpus d'entraînement $B_M = \{((d_1, d_2), Dis(d_1, d_2))\}_{1 \leq i \leq |B_M|}$
- Entraînement d'un modèle de régression pour prédire la distance entre deux documents

$$Dis_M(d_i, d_j) = Reg_M(\vec{d}_i - \vec{d}_j)$$

- Utilisation de la distance dans un algorithmes de regroupement (K-moyennes et K-medoides)

5.3 Sélection de la représentation des décisions

Trouver la représentation qui discrimine les cas sur leur champ sémantique

Corpus	Terminologie
<i>arcpa</i>	chance, perte chance, avocat, perte, diligence, chance obtenir, perdre, client, devoir conseil, manquement
<i>cas a</i>	chance, perte chance, chance succès, perte, client, préjudice indemnisable, article code commerce, indemnisable, condamnation emporter, emporter nécessairement rejet
<i>cas b</i>	défense intérêt, intérêt client, avocat, contractuel égard, responsabilité contractuel droit, responsabilité professionnel avocat, contractuel droit commun, assurer défense intérêt, civil avocat, grief articuler
<i>cas c</i>	rédacteur acte, rédacteur, avocat rédacteur acte, avocat rédacteur, qualité rédacteur acte, rédaction acte, qualité rédacteur, projet acte, prendre initiative conseiller, initiative conseiller
<i>cas d</i>	revêtir aucun, revêtir aucun caractère, article code, article code procédure, faire référence aucun, fautif madame, civil profit autre, civil depuis, mention expresse, moyen dont

5.3 Sélection de la représentation des décisions

Meilleures représentations sur la catégorisation manuelle

Distance	Base ^a	Silhouette optimale (pondération, réduction, dim.)
$Dis_{jaccard}$	0.001	0.212 (TP-NGL, FNM, 4)
Dis_{cos}	0.002	0.202 (TP-NGL, FNM, 4)
Dis_M	-0.049	0.195 (TP-NGL, FNM, 4)
$Dis_{braycurtis}$	0.002	0.182 (TP-NGL, FNM, 4)
$Dis_{euclidienne}$	0.001	0.168 (TP-NGL, FNM, 4)
$Dis_{manhattan}$	-0.019	0.17 (TP-NGL, FNM, 4)
$Dis_{pearson}$	0.014	0.057 (TP-CHI2, aucune, 19763)
Dis_{wmd}	-0.096	-

^a occurrence de mots pour Dis_{wmd} , et TF-IDF pour les autres distances.

5.4 Efficacité de la méthode proposée

Regroupement pour la catégorie annotée

Distance	Algorithme	K	Silhouette	ARI	NMI	R	P	F_1
Dis_M	K-moyennes	3	0.438	0.407	0.423	0.552	0.654	0.599
Dis_M	K-medoïdes	6	0.453	0.359	0.395	0.298	0.669	0.413
$Dis_{braycurtis}$	K-moyennes	4	0.473	0.383	0.407	0.446	0.658	0.532
$Dis_{braycurtis}$	K-medoïdes	5	0.448	0.344	0.375	0.331	0.645	0.437
Dis_{cosine}	K-moyennes	4	0.528	0.383	0.407	0.446	0.658	0.532
Dis_{cosine}	K-medoïdes	4	0.526	0.398	0.421	0.464	0.680	0.551
$Dis_{euclidean}$	K-moyennes	5	0.478	0.365	0.395	0.341	0.670	0.452
$Dis_{euclidean}$	K-medoïdes	5	0.456	0.313	0.346	0.335	0.619	0.434
$Dis_{jaccard}$	K-moyennes	4	0.570	0.367	0.391	0.439	0.643	0.522
$Dis_{jaccard}$	K-medoïdes	4	0.560	0.389	0.412	0.451	0.666	0.538
$Dis_{manhattan}$	K-moyennes	4	0.482	0.376	0.400	0.452	0.657	0.535
$Dis_{manhattan}$	K-medoïdes	5	0.452	0.368	0.397	0.345	0.675	0.456
$Dis_{pearson}$	K-moyennes	2	0.611	0.054	0.072	0.746	0.453	0.564
$Dis_{pearson}$	K-medoïdes	2	0.171	0.152	0.166	0.598	0.482	0.534
Dis_{wmd}	K-medoïdes	2	0.332	-0.016	0.002	0.545	0.397	0.459

5.4 Efficacité de la méthode proposée

Regroupement des catégories non annotées

\mathcal{D}	Distance	Algorithme	K	\bar{s}_K
$\mathcal{D}_{\text{concdel}}$ (30)	$Dis_{\mathcal{M}}$	K-medoïdes	4	0.639
	$Dis_{\mathcal{M}}$	K-moyennes	4	0.624
	Dis_{cosine}	K-medoïdes	4	0.675
	Dis_{cosine}	K-moyennes	4	0.632
	Dis_{jaccard}	K-medoïdes	5	0.719
	Dis_{jaccard}	K-moyennes	5	0.702

Cluster	Terminologie
0	distinctif, reproduire, code propriété intellectuel, contrefaçon, attaque, dépôt marque, circonstance intervenir, intervenir jugement, intervenir jugement déferer, caractère distinctif
1	publicitaire, action concurrence, défaut qualité agir, action concurrence déloyal, qualité agir, fichier client, force chose juger, fonder demande titre, date transfert, celui-ci fonder
2	acte concurrence, acte concurrence déloyal, clause non concurrence, non concurrence, clause non, entreprise concurrent, démarchage, démarcher, salarié, massif
3	non concurrencer, clause non concurrencer, tout droit, résilier contrat, marcher, préjudice invoquer, détournement, compte entre partie, contrat @card@ juillet, compte entre

1. Introduction
2. Annotation des sections et entités judiciaires
3. Identification des demandes des parties
4. Identification du sens du résultat
5. Découverte des circonstances factuelles
- 6. Conclusions**
 - 6.1 Bilan
 - 6.2 Perspectives

- Définition de tâches importantes pour l'analyse de corpus de décisions
 - Formulation en problèmes d'analyse de données textuelles
 - Production avec un expert de données annotées d'apprentissage
- Proposition et évaluation d'approches d'extraction de connaissances jurisprudentielles :
 - Application du HMM et CRF pour détecter les sections et les entités juridiques [**Tagny Ngompé** et al., 2017, **Tagny Ngompé** et al., 2019]
 - Approche d'identification des demandes par catégorie basée sur la proximité entre des termes-clés appris et les sommes d'argent
 - Deux extensions du Gini-PLS pour identifier le sens du résultat (article en préparation pour **Stats**)
 - Approche d'apprentissage d'une distance de similarité pour regrouper les décisions suivant les circonstances factuelles.
- Démonstration d'applications en analyse descriptive sur un grand corpus de décisions

○ Extensions des propositions

- Désambiguïsation des entités détectées pour indexer les décisions
- Expérimentation d'approches d'extraction des événements et relations pour l'identification des demandes
- Découverte des circonstances factuelles par modélisation thématique

○ Applications

- Anonymisation des décisions : confidentialité des informations
- Analyse prédictive : identifier les raisons qui poussent les juges à accepter une demande

Questions



Aletras, N., Tsarapatsanis, D., Preoțiu-Pietro, D., and Lampos, V. (2016).
Predicting judicial decisions of the European Court of Human Rights : A Natural Language Processing perspective.
PeerJ Computer Science, 2 :e93.



Andrew, J. J. and Tannier, X. (2018).
Automatic Extraction of Entities and Relation from Legal Documents.
In *Proceedings of the Seventh Named Entities Workshop*, pages 1–8.



Ashley, K. D. and Brüninghaus, S. (2009).
Automatically classifying case texts and predicting outcomes.
Artificial Intelligence and Law, 17(2) :125–165.



Cover, T. and Hart, P. (1967).
Nearest Neighbor Pattern Classification.
IEEE Transactions on Information Theory, 13(1) :21–27.



Duda, R. O., Hart, P. E., et al. (1973).
Pattern Classification And Scene Analysis, volume 3.
John Wiley & Sons, New York.



Fisher, R. A. (1936).
The use of multiple measurements in taxonomic problems.
Annals of Eugenics, 7(2) :179–188.



Grave, E., Mikolov, T., Joulin, A., and Bojanowski, P. (2017).

Bag of tricks for efficient text classification.

In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2, pages 427–431, Valencia, Spain.



Kumar, S., Reddy, P. K., Reddy, V. B., and Singh, A. (2011).

Similarity analysis of legal judgments.

In *Proceedings of Compute 2011 - Fourth Annual ACM Bangalore Conference*, page 17. ACM.



Liu, H. and Motoda, H. (2012).

Feature selection for knowledge discovery and data mining, volume 454.

Springer Science & Business Media.



Ma, Y., Zhang, P., and Ma, J. (2018).

An Efficient Approach to Learning Chinese Judgment Document Similarity Based on Knowledge Summarization.

arXiv preprint arXiv :1808.01843 [cs.AI].



McLachlan, G. J. (1992).

Discriminant Analysis and Statistical Pattern Recognition.

John Wiley & Sons.



Moens, M.-F., Boiy, E., Palau, R. M., and Reed, C. (2007).

Automatic detection of arguments in legal texts.

In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 225–230. ACM.



Mussard, S. and Souissi-Benrejab, F. (2018).
Gini-PLS Regressions.
Journal of Quantitative Economics, pages 1–36.



Nair, A. M. and Wagh, R. S. (2018).
Similarity Analysis of Court Judgements Using Association Rule Mining on Case Citation Data - A Case Study.
International Journal of Engineering Research and Technology, 11(3) :373–381.



Ng, H. T., Goh, W. B., and Low, K. L. (1997).
Feature selection, perceptron learning, and a usability case study for text categorization.
In *ACM SIGIR Forum*, volume 31, pages 67–73. ACM.



Peng, F. and McCallum, A. (2006).
Information extraction from research papers using conditional random fields.
Information processing & management, 42(4) :963–979.



Pudil, P., Novovičová, J., and Kittler, J. (1994).
Floating search methods in feature selection.
Pattern recognition letters, 15(11) :1119–1125.



Ravi Kumar, V. and Raghuveer, K. (2012).
Legal documents clustering using latent dirichlet allocation.
International Journal of Applied Information Systems (IJAIS), 2(6) :34–37.



Schütze, H., Hull, D. A., and Pedersen, J. O. (1995).

A comparison of classifiers and document representations for the routing problem.

In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, pages 229–237. ACM.



Seymore, K., McCallum, A., and Rosenfeld, R. (1999).

Learning hidden Markov model structure for information extraction.

AAAI-99 workshop on machine learning for information extraction.



Shulayeva, O., Siddharthan, A., and Wyner, A. (2017).

Recognizing cited facts and principles in legal judgements.

Artificial Intelligence and Law, 25(1) :107–126.



Sparck Jones, K. (1972).

A statistical interpretation of term specificity and its application in retrieval.

Journal of Documentation, 28(1) :11–21.



Şulea, O.-M., Zampieri, M., Malmasi, S., Vela, M., P. Dinu, L., and van Genabith, J. (2017a).

Exploring the Use of Text Classification in the Legal Domain.

In Proceedings of 2nd Workshop on Automated Semantic Analysis of Information in Legal Texts, page 5, London, United Kingdom. ASAIL'2017.



Şulea, O.-M., Zampieri, M., Vela, M., and van Genabith, J. (2017b).

Predicting the Law Area and Decisions of French Supreme Court Cases.

In Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2017, pages 716–722.



Tenenhaus, M. (2005).

La regression logistique PLS.

In Dreesbeke, Jean-Jacques and Lejeune, Michel and Saporta, Gilbert, editor, *Modèles statistiques pour données qualitatives*, chapter 12, pages 263–276. Editions Technip.



Tagny Ngompé, t., Harispe, S., Zambrano, G., Montmain, J., and Mussard, S. (2019).

Detecting Sections and Entities in Court Decisions Using HMM and CRF Graphical Models.

In Pinaud, B., Guillet, F., Gandon, F., and Largeron, C., editors, *Advances in Knowledge Discovery and Management : Volume 8*, pages 61–86. Springer International Publishing.



Tagny Ngompé, t., Harispe, S., Zambrano, G., Montmain, J., and Mussard, S. (2017).

Reconnaissance de sections et d'entités dans les décisions de justice : application des modèles probabilistes HMM et CRF.

In *proceedings of Extraction et Gestion des Connaissances - EGC 2017, Revue des Nouvelles Technologies de l'Information*, pages 201–212.

Grenoble, France.



Thenmozhi, D., Kannan, K., and Aravindan, C. (2017).

A Text Similarity Approach for Precedence Retrieval from Legal Documents.

In *Proceedings of Forum for Information Retrieval Evaluation - FIRE (Working Notes)*, pages 90–91.



Vapnik, V. N. (1995).

The Nature of Statistical Learning Theory.

Springer.



Waltl, B., Landthaler, J., Scepankova, E., Matthes, F., Geiger, T., Stocker, C., and Schneider, C. (2017). Automated extraction of semantic information from German legal documents. In *IRIS : Internationales Rechtsinformatik Symposium. Association for Computational Linguistics*.



Waltl, B., Matthes, F., Waltl, T., and Grass, T. (2016). LEXIA - A Data Science Environment for Semantic Analysis of German Legal Texts. In *IRIS : Internationales Rechtsinformatik Symposium*. Salzburg, Austria.



Wang, S. and Manning, C. D. (2012). Baselines and bigrams : Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Short Papers-Volume 2*, pages 90–94. Association for Computational Linguistics.



Wold, H. (1966). Estimation of principal components and related models by iterative least squares. *Multivariate Analysis*, pages 391–420.



Wyner, A. and Peters, W. (2010). Lexical Semantics and Expert Legal Knowledge towards the Identification of Legal Case Factors. In *JURIX*, volume 10, pages 127–136.



Wyner, A. Z. (2010). Towards annotating and extracting textual legal case elements. *Informatica e Diritto : special issue on legal ontologies and artificial intelligent techniques*, 19(1-2) :9–18.