

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

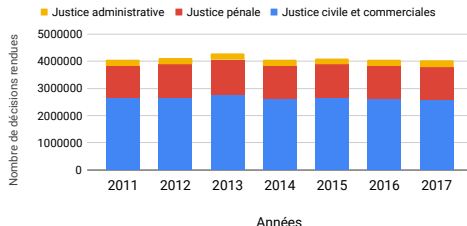
4. Identification du sens du résultat

5. Découverte des circonstances factuelles

6. Conclusions

Motivations

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



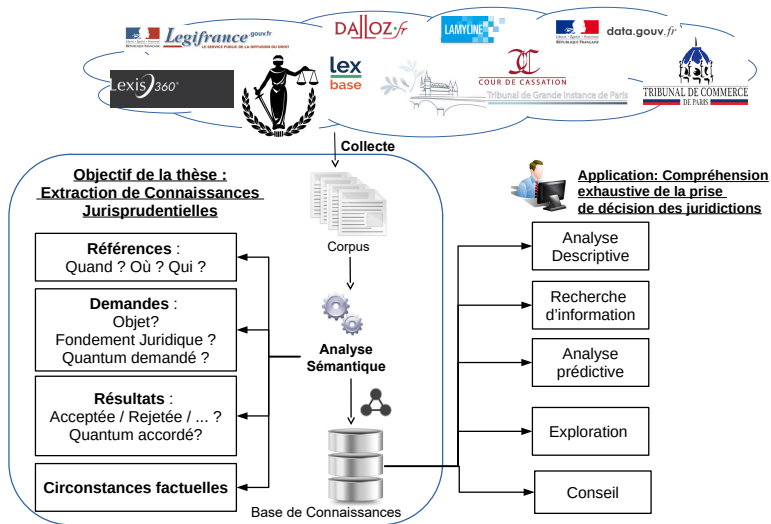
2. Les moteurs de recherche juridique limités :
 - Pas de critère de recherche sémantique (catégorie de demande, type de faits, etc.)
 - pas d'analyse synthétique de corpus

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 (Chine) [Ma et al., 2018]
- Similarité entre décisions
 - décisions qui citent les mêmes lois et précédents [Nair and Wagh, 2018]
 - recherche d'affaires antérieures pertinentes [Thenmozhi et al., 2017]
 - identifier la sentence prononcée (Chine) [Ma et al., 2018]
 - similarité basée sur la question discutée et les faits sous-jacents (Inde) [Kumar et al., 2011]
 - regroupement non-supervisé [Ravi Kumar and Raghuveer, 2012]

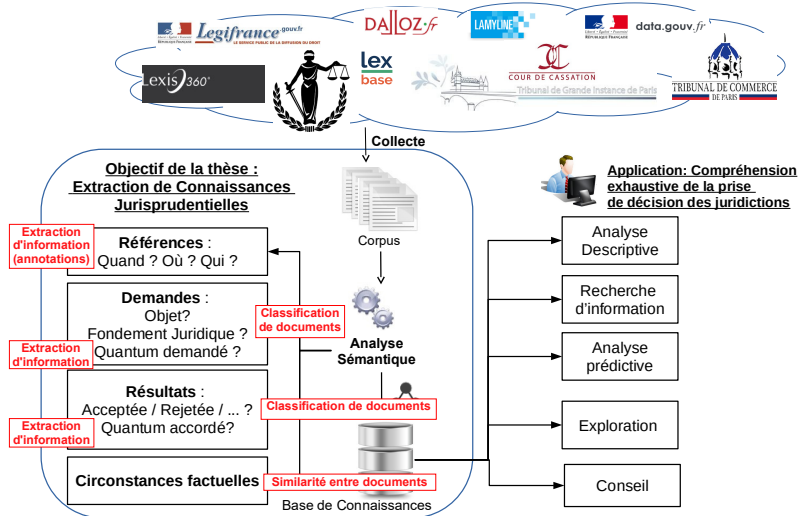
1.2 Objectif de la thèse

Tâches et exemples d'applications



1.2 Objectif de la thèse

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 modèle

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 utilisées

The screenshot shows a legal document on the left and a legend titled "Original markups" on the right. The document text includes:

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
intimés et il convient de leur allouer à ce titre la somme de 1.000
€ .

The legend "Original markups" lists the following categories with checkboxes:

- ☒ appellant
- ☒ 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

ARRÊT N°
R.G : 11/03924
COUR D'APPEL DE NÎMES
CHAMBRE CIVILE
1ère Chambre A
ARRÊT DU 20 MARS 2012
APPELANTE :
Madame Michèle A. ...
assistée de la SELARL VAJOU, ...
INTIMES :
Monsieur Martial B ...
assisté de la SCP MARION GUIZARD PATRICIA SER-
VAIS, ...
COMPOSITION DE LA COUR LORS DU DÉLIBÉRÉ :
M. Dominique BRUZY, Président
M. Serge BERTHET, Conseiller
...

Entêtes : méta-données

FAITS, PROCEDURE, ...

Madame Michèle A. demande :

...

- de condamner Madame JONES-B. à lui payer la somme de
2.500 euros au titre de l'article 700 du Code de Procédure Civile,

Corps : demandes, arguments et normes

PAR CES MOTIFS, LA COUR :

...

Vu l'article 809 du Code de Procédure Civile,

...

Déboute Madame A. de sa demande de provision sur
dommages-intérêts.

...

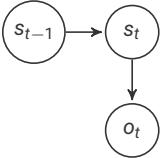
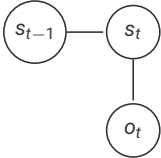
Vu l'article 700 du Code de Procédure Civile,

Condamne Madame JONES-B. à verser à Madame A. la somme
de 2.500 euros.

Dispositif : résultats et normes

2.2 Approches probabilistes de détection d'entités

Modèles probabilistes à états et observations

| 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> |

Objectif : Trouver la séquence la plus probable d'étiquetage pour l'ensemble du texte

Entraînement sur des séquences préalablement étiquetées

Méthodes explorées

- Descripteurs de lignes pour les sections : longueur ? position ? etc.
- Descripteurs de mots pour les entités : est-ce une initiale ("B.") ? est-ce un mot clé de citation de loi ? etc.
- Schéma d'étiquetage : distinction des parties d'une entité

| | <i>composée</i> | <i>de</i> | <i>Madame</i> | <i>Martine</i> | <i>JEAN</i> | <i>,</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 | ... |

- sélection du sous-ensemble de descripteurs court et aux meilleurs résultat (recherche par BDS et SFFS)

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'entité de moins de 3%
- sélection des descripteurs
 - Lenteur des algorithmes BDS et SFFS (plus de 15h)
 - 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

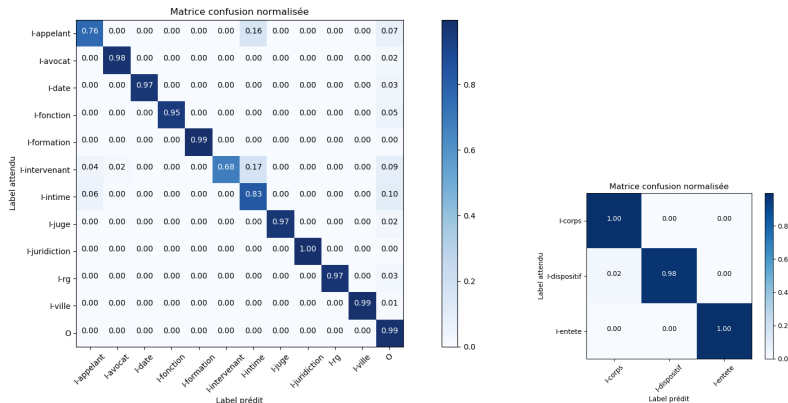


FIGURE – Matrice de confusion des modèles basés CRF

Impact de la quantité de décisions d'entraînement

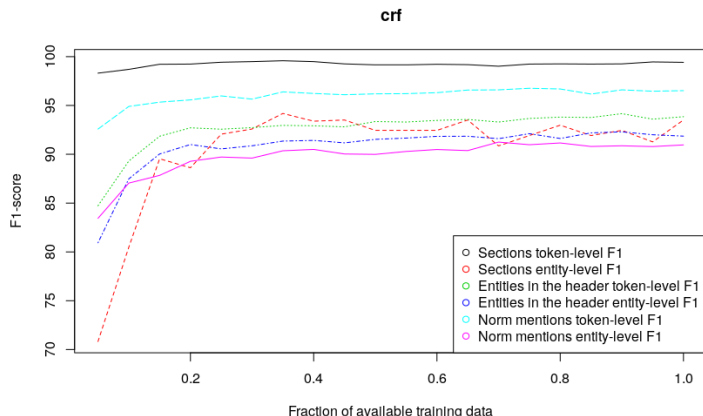


FIGURE – Evolution de la F1-mesure en fonction de la fraction utilisée

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 |

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

3.3 Résultats expérimentaux

4. Identification du sens du résultat

5. Découverte des circonstances factuelles

6. Conclusions

3.1 Objectif de la tâche

Exemple : dommage-intérêts pour procédure abusive (danais)

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; ...

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

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

| IDENTIFICATION DE LA DECISION | | | DESCRIPTION DE LA PRETENTION | | | DESCRIPTION DU RESULTAT | |
|-------------------------------|-------------|----------|------------------------------|--|------------|-------------------------|------------------------------|
| Type | Ressort | RG | OBJET | NORME | QUANTUM | RESULTAT | QUANTUM RESULTAT (obtenu) |
| CA | Saint Denis | 14/01082 | dommages-intérêts | 1382 code civil + 32-1 code de procédure civile : en procédure abusive | 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
- Énoncés non structurés, avec des références, et des agrégations

3.2 Méthode proposée

1. Détermination automatique de la terminologie (déclencheurs) de la catégorie

$$ngl(t, c) = \frac{\sqrt{N}(N_{t,c}N_{\bar{t},\bar{c}}) - (N_{t,\bar{c}}N_{\bar{t},c})}{\sqrt{N_t N_{\bar{t}} |D_c| |D_{\bar{c}}|}}.$$

2. Détection de la présence de la catégorie par classification de la décision (c vs. \bar{c})
3. Identification des passages de demandes et résultats
4. Exploiter la proximité entre les déclencheurs de la catégorie et sommes d'argent pour extraire les quanta :

Section Litige : identification de la demande

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}

...

Section Dispositif : identification du résultat

La cour ...

CONFIRME le jugement entrepris en toutes ses dispositions.

5. Mise en correspondance des informations relatives à la même demande

3.3 Résultats expérimentaux

Données

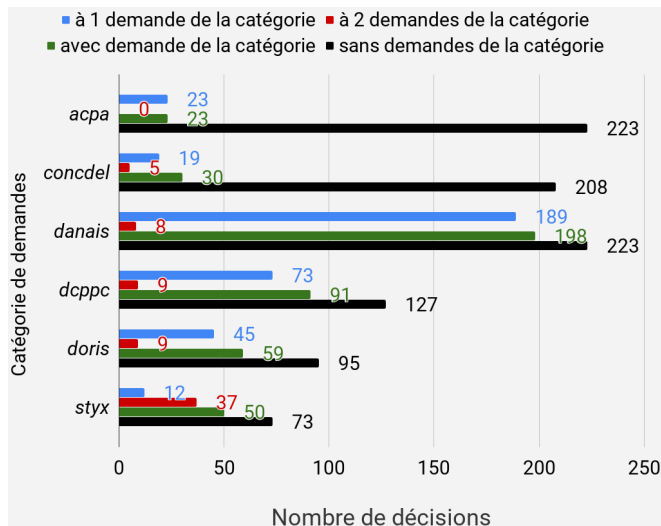
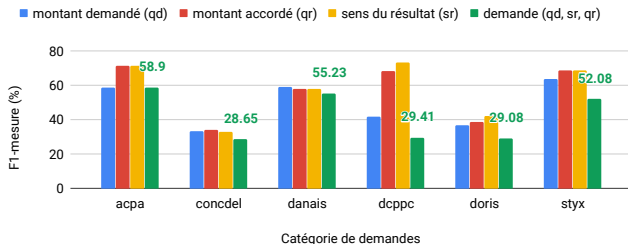


FIGURE — Répartitions des demandes dans les documents annotés.

Efficacité de la méthode

- Détection de catégorie facile par des classifieurs traditionnels (K-plus-proches-voisins, SVM, Bayésien naïf, Arbre) : $98.8\% \leq F_1\text{-mesure} \leq 100\%$
- Résultat plus facile à extraire que le montant



- Source d'erreurs :
 - Sélection difficile de déclencheurs rares
 - Non exploitation des références aux jugements antérieurs
 - Certains quanta sont absents des sections Litige et Dispositif
 - Mauvaise méthode de mise en correspondance

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 Méthode : adaptations de la régression Gini-PLS

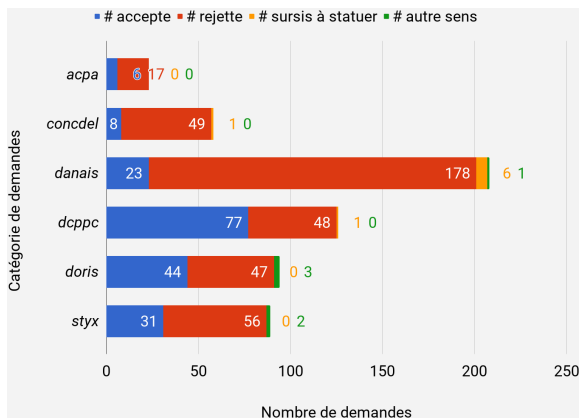
4.3 Résultats expérimentaux

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**



Algorithmes classiques existants

- Classifieur bayésien naïf
- K-plus-proches-voisins
- SVM
- Arbre de décision
- Analyse discriminante linéaire et quadratique
- NBSVM [Wang and Manning, 2012]
- fastText [Grave et al., 2017]
- etc.

4.2 Méthode : adaptations de la régression Gini-PLS

PLS standard (Régression partielle des moindres carrés)

Réduction supervisée des dimensions x_1, x_2, \dots, x_p en
composantes orthogonales t_1, \dots, t_h

$$t_h = w_{h1}x_1 + \dots + w_{hj}x_j + \dots + w_{hp}x_p$$

$$\text{avec } w_{hj} = \frac{\text{cov}(u_{(h-1)j}, \epsilon_h)}{\sqrt{\sum_{p=1}^j \text{cov}^2(u_{(h-1)j}, \epsilon_h)}}, y = c_1 t_1 + \dots + c_h t_h + \epsilon_h,$$

$$\text{et } x_j = \beta_{1j} t_1 + \dots + \beta_{hj} t_h + u_{(h-1)j}$$

4.2 Méthode : adaptations de la régression Gini-PLS

1. Gini-PLS : élimination de la sensibilité au *outliers* en remplaçant la covariance $\text{cov}(x_j, y)$ par la covariance de Gini $\text{cog}(y; x_j) := \text{cov}(y; R(x_j))$ pour l'estimation des résidus $u_{(h)j}$ et des poids w_{hj} [Mussard and Souissi-Benrejab, 2018]
2. Logit-PLS : $\forall j > 1$, les w_{hj} sont les coefficients de la régression logistique de y sur les composantes $t_1, \dots, t_{h-1}, u_{(h-1)j}$ [Tenenhaus, 2005]
3. Gini-Logit-PLS : covariance Gini pour $u_{(h)j}$ et coefficient Logit pour les w_{hj}

Classifieurs PLS vs. aux classifieurs classiques

| Représentation | Algorithme | F_1 | $F_{1_{\text{arbre}}} - F_1$ | $F_{1_{\text{max}}} - F_{1_{\text{min}}}$ |
|---------------------|--------------|-------|------------------------------|---|
| $tf - gss$ | Arbre | 0.668 | 0 | 0.42 |
| $tf - avg_{global}$ | LogitPLS | 0.648 | 0.02 | 0.263 |
| $tf - avg_{global}$ | StandardPLS | 0.636 | 0.032 | 0.346 |
| $tf - \Delta_{DF}$ | GiniPLS | 0.586 | 0.082 | 0.426 |
| $tf - \Delta_{DF}$ | GiniLogitPLS | 0.578 | 0.09 | 0.547 |
| - | NBSVM | 0.494 | 0.174 | 0.434 |
| - | fastText | 0.412 | 0.256 | 0.127 |

TABLE – Comparaison des combinaisons représentation+algorithme proposées avec les arbres, fastText et NBSVM pour la détection du sens du résultat.

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 |
| | litige.motifs_dispositif | <i>tf</i> – <i>dbidf</i> | StandardPLS | 0.697 |
| | litige.motifs_dispositif | <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 |
| | motifs | <i>tf</i> – <i>idf</i> | GiniLogitPLS | 0.703 |
| | context | <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 |
| | demande_resultat.a_resultat.context | <i>atf</i> – <i>avg_{global}</i> | LogitPLS | 0.721 |
| | demande_resultat.a_resultat.context | <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 |
| | demande_resultat.a_resultat.context | <i>tf</i> – χ^2 | LogitPLS | 0.94 |
| | litige.motifs_dispositif | <i>tp</i> – <i>mar</i> | StandardPLS | 0.934 |
| <i>doris</i> | litige.motifs_dispositif | <i>tp</i> – <i>dsidf</i> | GiniPLS | 0.806 |
| | litige.motifs_dispositif | <i>tp</i> – <i>dsidf</i> | GiniLogitPLS | 0.806 |
| | litige.motifs_dispositif | <i>atf</i> – <i>ig</i> | StandardPLS | 0.772 |
| <i>styx</i> | motifs | <i>tf</i> – <i>dsidf</i> | Arbre | 1 |
| | demande_resultat.a_resultat.context | <i>logave</i> – <i>dsidf</i> | GiniLogitPLS | 0.917 |
| | litige.motifs_dispositif | <i>tf</i> – <i>rf</i> | GiniPLS | 0.833 |

TABLE – Impact de la restriction des documents à certains passages sur l'identification du sens du résultat.

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 du regroupement

6. Conclusions

5.1 Objectif de la tâche

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

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 regroupement non supervisé des décisions

- Apprentissage d'une distance basé sur la transformation d'un document en l'autre
 - Formulation de la distance pour un ensemble de modifications connues

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

- Génération d'un corpus d'entraînement
$$B_M = \{((d_1, d_2), Dis(d_1, d_2))_i\}_{1 \leq i \leq |B_M|}$$
- 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 |

TABLE – Terminologies de la catégorie *arcpa* et de ses cas

| 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.

TABLE – Meilleures représentations sur la catégorisation manuelle.

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 |

TABLE — Evaluation de la catégorisation par K-moyennes et K-medoïdes sur D_{arcpa} avec détermination du nombre de clusters basée sur la silhouette.

Regroupement des catégories non annotées

| | | | | |
|---------------------|-----------------|------------|---|-------|
| D_{doris} (59) | Dis_M | K-medoïdes | 2 | 0.509 |
| | Dis_M | K-moyennes | 3 | 0.527 |
| | Dis_{cosine} | K-medoïdes | 5 | 0.549 |
| | Dis_{cosine} | K-moyennes | 4 | 0.586 |
| | $Dis_{jaccard}$ | K-medoïdes | 3 | 0.600 |
| | $Dis_{jaccard}$ | K-moyennes | 4 | 0.645 |

TABLE — Evaluation non-supervisée des K-moyennes et K-medoïdes sur D_{doris} .

| Cluster | Terminologie (<i>ngl</i>) |
|---------|---|
| 0 | excéder inconvenient, inconvenient normal, excéder inconvenient normal, normal voisinage, inconvenient normal voisinage, inconvenient, trouble excéder inconvenient, trouble excéder, excéder, normal |
| 1 | copropriétaire, syndicat copropriétaire, syndicat, condamner in, anormal voisinage, trouble anormal voisinage, in, trouble anormal, syndic, jouissance subir |
| 2 | deux fond—fonds, séparatif deux fond—fonds, limite séparatif deux, ordonner démolition, séparatif deux, implanter, condamner démolir, devoir établir toit, devoir établir, toit manière |
| 3 | manière plus, chose manière plus, chose manière, usage prohiber loi, prohiber loi règlement, prohiber loi, absolu, usage prohiber, manière plus absolu, plus absolu |
| 4 | situer zone, hauteur @card@ mètre, hauteur dépasser, appel contester, vitrer, dont hauteur dépasser, urbaniser, recevabilité junknown; appel, cahier charge lotissement, charge lotissement |

TABLE — Terminologies des circonstances factuelles découvertes en combinant les K-medoïdes et la distance cosinus sur D_{doris} .

1. Formulation comme problème de regroupement non supervisé de décisions de la catégorie
2. Méthode d'apprentissage d'une distance de dis-similarité au sein d'une catégorie
3. Sélection de la représentation des textes qui reflète la notion subjective de similarité de l'expert
4. Expérimentation des propositions sur 7 catégories de demandes dont 1 annotées

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 problèmes importants d'analyse de corpus de décisions
 - Formulation en tâches de fouille de textes
 - 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
 - Approche d'identification des demandes par catégorie basée sur la proximité entre des termes-clés appris et les sommes d'argent
 - Proposition et évaluation d'extensions du Gini-PLS pour identifier le sens du résultat
 - 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

- Amélioration des propositions
 - Désambiguïser les entités détectées pour indexer les décisions
 - Expérimentation des approches récentes pour l'identification des **demandes formalisées comme relation entre montant demandé et montant accordé**
 - Découverte des circonstances factuelles vue comme **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.



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.



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].



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.



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



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



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.



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.



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.



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.

References IV



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.



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.