

Méthodes d'analyse sémantique de corpus de décisions jurisprudentielles

Soutenance de thèse

présentée par Gildas TAGNY NGOMPÉ

le 24 janvier 2020

devant le jury composé de:

- 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
4. Identification du sens du résultat
5. Découverte des circonstances factuelles
6. Conclusions

1. Introduction

1.1 Contexte

1.2 État de l'art

1.3 Objectif de la thèse

2. Annotation des sections et entités judiciaires

3. Identification des demandes

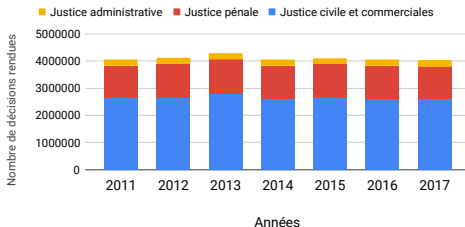
4. Identification du sens du résultat

5. Découverte des circonstances factuelles

6. Conclusions

- Utilité de la jurisprudence pour les juristes
 - elle est analysée pour comprendre l'application de la loi
- Motivation : limite de l'analyse manuelle

1. Gros volume de décisions



2. Limite des moteurs de recherche juridique :

- 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 Raghuv eer, 2012]

Objectif de la thèse

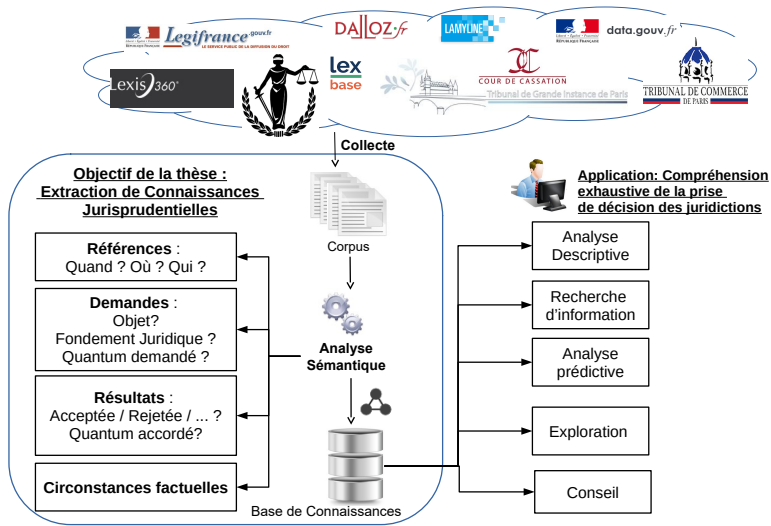


FIGURE – Objectifs et exemples d’application de la thèse.

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
€ .

PAR CES MOTIFS
La Cour , ...

▼ Original markups

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

Positionnement en fouille de texte

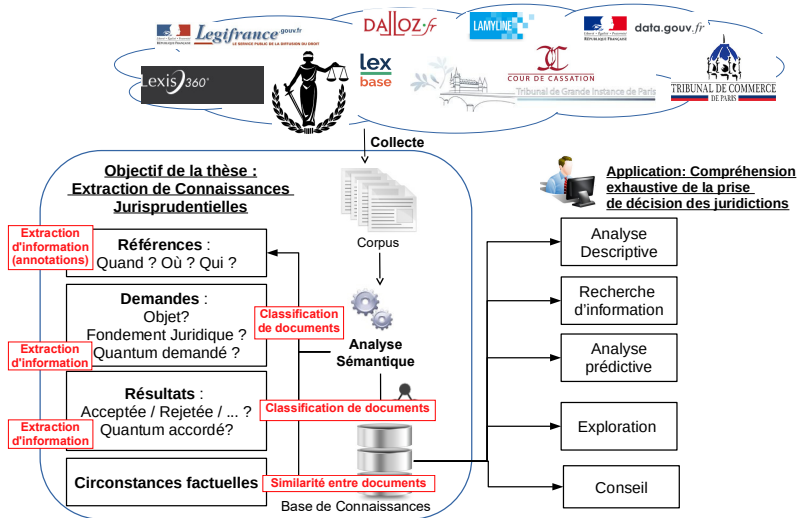


FIGURE – Tâches abordées en analyse de données textuelles.

Difficultés rencontrées par l'automatisation de ces tâches

- Les décisions sont des textes non-structurés
- Le langage juridique est complexe

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

SERVAIS, ...

COMPOSITION DE LA COUR LORS DU DÉLIBÉRÉ :

M. Dominique BRUZY, Président

M. Serge BERTHET, Conseiller

...

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,

PAR CES MOTIFS, LA COUR :

...

Vu l'article 809 du Code de Procédure Civile,

...

Débouté 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.

1. Introduction

2. Annotation des sections et entités judiciaires

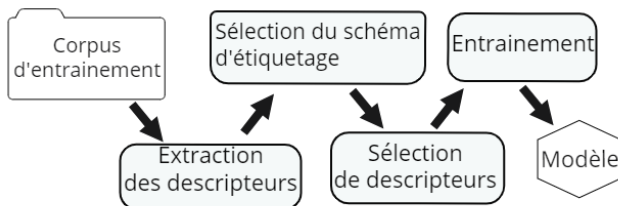
3. Identification des demandes

4. Identification du sens du résultat

5. Découverte des circonstances factuelles

6. Conclusions

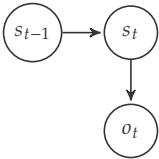
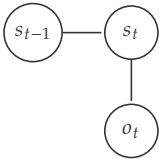
- Formulation du problème
 - Détection des sections par étiquetage de séquence de lignes
 - Détection des entités par étiquetage de séquence de mots
- Etude de l'application du HMM et du CRF
 1. sélection de modèle :



2. Analyse de l'impact de la quantité de données d'entraînement

Etudes de l'application du HMM et CRF

Modèles probabilistes d'étiquetage de séquences

HMM	CRF
un seul descripteur par observation	plusieurs descripteurs complexes par observation
	
$P_{\lambda}(S O) = \prod_{t=1}^T P(s_t s_{t-1}) * P(o_t s_t)$ <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

Entrainement fait sur des séquences préalablement étiquetées

1. Introduction

2. Annotation des sections et entités judiciaires

3. Identification des demandes

3.1 Objectif : identifier les informations sur les demandes

3.2 Méthode : identifier les passages, puis les informations

3.3 Expérimentations sur 6 catégories de demandes

4. Identification du sens du résultat

5. Découverte des circonstances factuelles

6. Conclusions

Objectif : identifier les informations sur les demandes

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**.

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 abusives	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

Retrouver les demandes à l'aide des termes clés

1. Identification de la catégorie par classification de textes
2. Exploiter la proximité entre **vocabulaire 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 3000,00 € **pour procédure abusive** et aux entiers dépens ;]_{demande_danais}
...

Section Dispositif : identification du résultat

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

3. Lier les informations relatives à la même demande

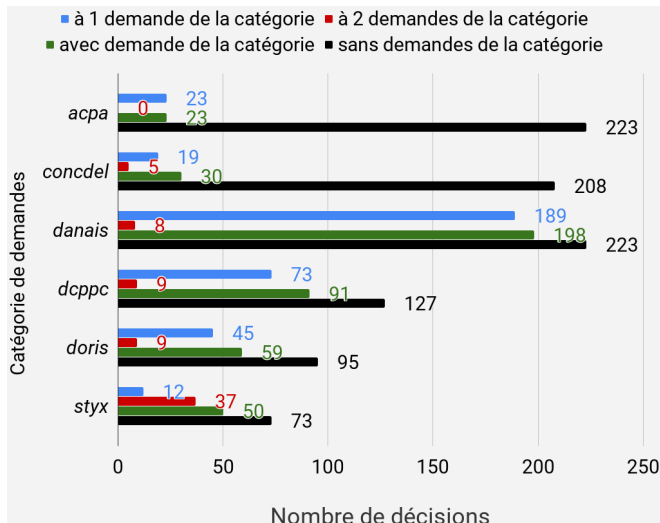
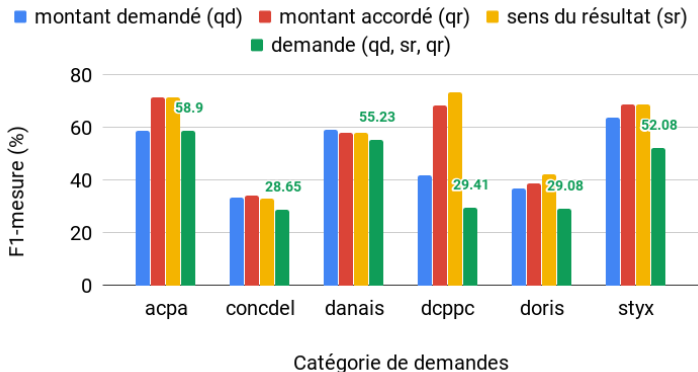


FIGURE — Répartitions des demandes dans les documents annotés.

Résultat : Extraction

- Détection de catégorie facile par des classifieurs traditionnels (K-plus-proches-voisins, SVM, Bayésien naïf, Arbre) : $98.8\% \leq F_1\text{-mesure} \leq 100\%$
- Extraction des quantas et sens du résultat



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

1. Méthodes

2. Résultats

1. Introduction
2. Annotation des sections et entités judiciaires
3. Identification des demandes
4. Identification du sens du résultat
5. Découverte des circonstances factuelles
 - 5.1 Objectifs
 - 5.2 Méthode : apprentissage d'une distance et utilisation pour du regroupement
 - 5.3 Sélection de la représentation des décisions
 - 5.4 Efficacité du regroupement

- 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
 - 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)

Sélection de la représentation : objectif

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

Sélection de la représentation : résultats

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>n gl</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 <unknown> 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
4. Identification du sens du résultat
5. Découverte des circonstances factuelles
6. Conclusions
 - 6.1 Bilan
 - 6.2 Perspectives

Contributions

- Etude de l'application du HMM et CRF pour détecter les sections et les entités juridiques
- Approche d'identification des demandes basée sur la proximité entre les termes-clés et les sommes d'argent
- 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.

Limites

- Évaluation sur de faibles quantité de données annotées ;
- Non expérimentation de méthodes récentes (réseaux de neurones)

Conclusions : perspectives

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

References I



Aletras, N., Tsarapatsanis, D., Preotiuc-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.



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.

References II



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 Raghuvver, 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.

References III



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.



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.



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.