

Analyse sémantique d'un corpus exhaustif de décisions jurisprudentielles pour l'élaboration d'un modèle prédictif du risque judiciaire

Début de thèse: 15 Décembre 2015

Doctorant: Gildas Tagny Ngompé

Direction de thèse:

- Jacky Montmain (École des mines d'Alès, LGI2P)
- Stéphane Mussard (Université de Nîmes, CHROME)

Encadrement de proximité:

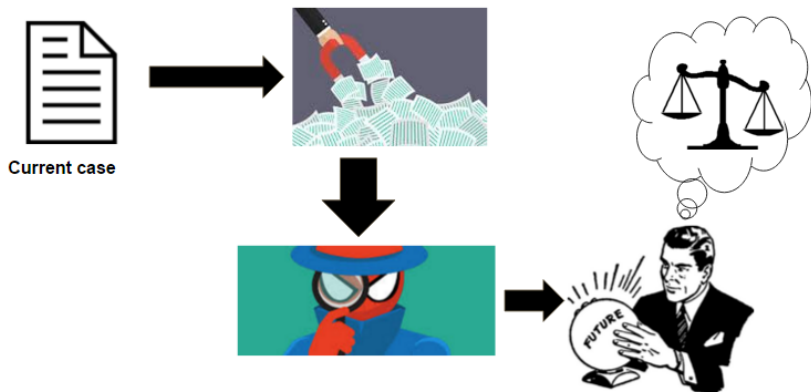
- Sébastien Harispe (Ecole des Mines d'Alès, LGI2P)
- Guillaume Zambrano (Université de Nîmes, CHROME)



1. Motivations et objectifs
2. Détection de sections et d'entités
3. Extraction d'informations sur les demandes
4. Conclusion
5. Questions ?

Motivations et objectifs

Les juristes analysent les décisions afin d'anticiper



Défis : grand volume de décisions

Plus de 4 millions de décisions prononcées / an

	2010	2011	2012	2013	2014
Justice civile	2 673 131	2 654 179	2 647 813	2 761 554	2 618 374
Justice pénale	1 173 242	1 180 586	1 251 979	1 303 469	1 203 339
Justice administrative	224 787	225 608	228 680	221 882	230 477

Source : <http://www.justice.gouv.fr/budget-et-statistiques-10054/chiffres-cles-de-la-justice-10303/>

TABLE – Nombre de décisions prononcées en France par an

Défis : Recherches et analyses sémantiques difficiles

Moteurs de recherche juridique à mots-clés

Pas d'analyse synthétique des décisions

☐ Recherche simple ☒ Recherche avancée

Mots ou expressions

Recherche

Ex : gérant **et** pouvoir, bail **s/5** résil!
[Aide à la recherche](#)

Gestion automatique des :

☒ Singulier / Pluriel ☒ Masculin / Féminin

☐ Verbes conjugués **avoir** cherche **ayons**

Sources ☒ *Toutes les sources

[Répertoire des sources](#)

ou

☒ Encyclopédies

☐ Codes et Lois

☐ JurisData

☐ Toute la jurisprudence

☐ Revues

☐ Bibliographies

☐ Actualités

☐ Bulletins Officiels

☐ Autorités administratives

☐ Parlement

☐ Europe

☐ Conventions Collectives

Période

Source : LexisNexis.com

Défis : Documents non-structurés et langage 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,

...

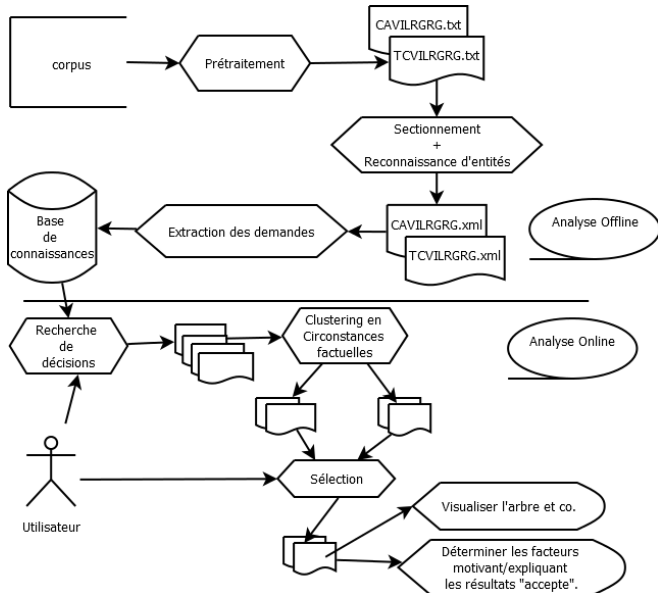
Déboute Madame A. de sa demande de
provision sur dommages-intérêts.

...

Vu l'article 700 du Code de Procédure
Civile,

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

Notre projet : Automatiser la structuration et l'analyse



Détection de sections et d'entités

Sectionner les décisions pour organiser l'extraction

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,

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 provi-
sion sur dommages-intérêts.

...

Vu l'article 700 du Code de Procédure Civile,
Condamne Madame JONES-B. à verser à Ma-
dame A. la somme de 2.500 euros.

Dispositif : résultats et normes

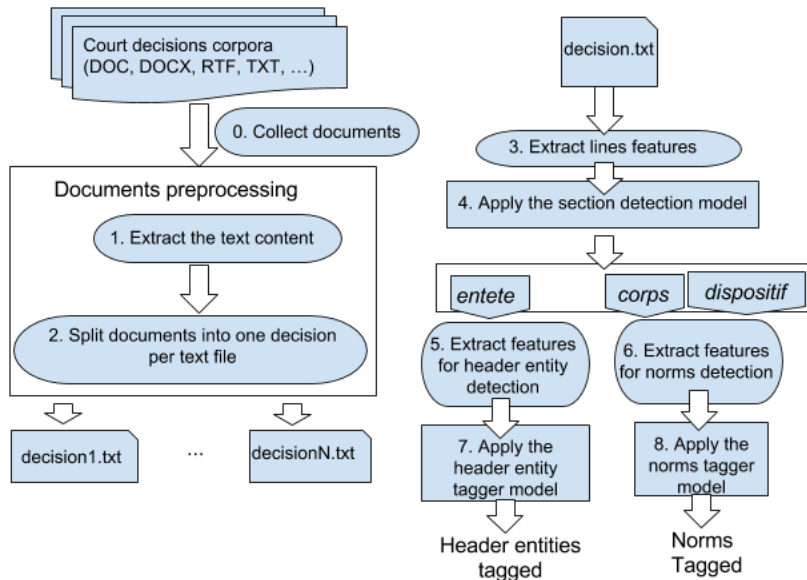
Entêtes : méta-données

Entités et sections à détecter

Entités	Labels	Exemples
Section entête (E)		
Numéro R.G.	RG	"10/02324", "60/JAF/09"
Ville	VL	"NÎMES", "Agen", "Toulouse"
Type de juridiction	JR	"COUR D'APPEL"
Formation	FM	"1re chambre", "Chambre économique"
Date	DT	"01 MARS 2012", "15/04/2014"
Partie appelante	AP	"SARL K.", "Syndicat ...", "Mme X ..."
Partie intimée	IM	- // -
Partie intervenante	IV	- // -
Avocat	AV	"Me Dominique A., avocat au barreau de Papeete"
Juge	JG	"Monsieur André R.", "Mme BOUSQUEL"
fonction du juge	FT	"Conseiller", "Président"
Corps (T) et dispositif (D)		
Norme	NO	"l' article 700 NCPC", "articles 901 et 903"
Élément à éviter	O	<i>tout élément ne faisant partie d'aucune entité ciblée</i>

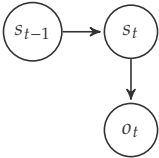
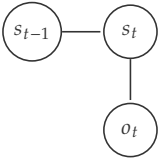
TABLE – Entités et leurs labels par section.

Architecture proposée



Approches probabilistes d'étiquetage de séquence

Modèles probabilistes à états et observations

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>[?]</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>[?]</p>

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

Entraînement sur des textes manuellement annotés

Premiers résultats [?]

	HMM			CRF-			CRF+		
labels	P(%)	R(%)	F1	P(%)	R(%)	F1	P(%)	R(%)	F1
<i>Section Entête (E)</i>									
AP	35.3	14.1	20.1	64.9	48.8	55.6	92.0	86.7	89.3
AV	83.8	98.3	90.5	96.4	97.5	96.9	97.6	98.1	97.9
DT	70.9	72.6	71.7	94.4	86.8	90.4	98.8	97.7	98.2
FM	87.6	93.7	90.5	98.8	98.4	98.6	98.9	99.3	99.1
FT	88.8	59.8	71.3	94.2	92.3	93.3	97.1	95.5	96.3
IM	53.1	57.4	55.1	67.2	64.6	65.8	89.3	88.1	88.7
IV	-	2.2	-	25.9	26.5	26.2	67.3	41.4	46.4
JG	68.0	85.7	75.7	96.2	95.7	96.0	98.1	97.7	97.9
JR	75.8	99.5	86.0	98.6	99.4	99.0	99.3	99.4	99.4
RG	-	0	-	83.7	46.1	59.4	98.6	97.4	98.0
VL	93.1	27.9	42.6	98.2	98.4	98.3	99.0	99.0	99.0
<i>Sections inférieures (T & D)</i>									
NO	92.9	90.9	91.9	96.0	93.8	94.9	97.9	96.5	97.2

5-fold validation croisée avec 500 documents annotées manuellement

TABLE – Précision (P), rappel (R), F1-mesure (F1) au niveau des mots.

Extraction d'informations sur les demandes

Extraction des informations sur les demandes

	A	B	C	D	F	H	L	N
1	IDENTIFICATION DE LA DECISION			DESCRIPTION DE LA PRETENTION			DESCRIPTION DU RESULTAT	
2	Type	Ressort	RG	OBJET	NORME	QUANTUM	RESULTAT	QUANTUM RESULTAT (obtenu)
	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
441	CA	Lyon	14/06911	dommages-intérêts	700 Code de Procédure Civile	3,500.00 €	rejette	0.00 €
442	CA	Lyon	14/06911	dommages-intérêts	700 Code de Procédure Civile	2,000.00 €	accepte	1,500.00 €

TABLE – Tableau des informations sur les demandes

Exemple : Informations pertinentes à extraire

- **Catégorie prédéfinie** : Dommages-intérêts pour procédure abusive
 - **Objet** : Dommages-intérêts
 - **Fondement** : Articles 1382 code civil et 32-1 code de procédure civile
- **Quantum demandé** : 20 000 euros
- **Sens du résultat** : "rejette"
- **Quantum accordé** : 0 euros

Difficultés (1)

Expressions non structurées, par **référence**, par **agrégation**

EXEMPLE (SUITE) : EXPRESSION DE DEMANDE

La société A. conclut à la confirmation du jugement entrepris sauf à former appel incident sur la disposition du jugement l'ayant déboutée de sa demande de **dommages intérêts pour abus de procédure** et elle demande à la cour de condamner l'appelante à lui payer la somme de **20 000 euros** à titre de dommages intérêts ...

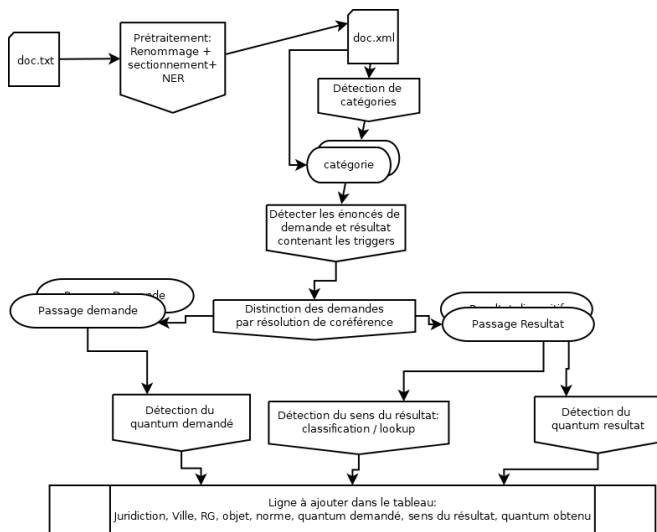
EXEMPLE (SUITE) : EXPRESSION DE RESULTAT

La cour, ...

Confirme **la décision entreprise** en **toutes ses dispositions**,

- Présence de plusieurs demandes de catégories similaires et/ou différentes dans une même décision
- Toutes les catégories ne sont pas connues d'avance

Proposition : Extraction des demandes par catégorie



- Méthode générique qui s'adapte aux spécificités de la catégorie traitée
- Définition incrémentale des catégories

Apprentissage des termes caractéristiques des catégories

Catégories	Termes			
<i>acpa</i>	amende civile	32-1 du code	amende	article 32-1
<i>concdel</i>	actes de concurrence	désorganisation	concurrence déloyale	salarié démissionnaire
<i>danais</i>	procédure abusive	abus	abusive	32-1
<i>dcppc</i>	créance déclarée	juge-commissaire	chirographaire	admet la créance
<i>doris</i>	trouble anormal	anormal	voisinage	trouble
<i>styx</i>	700 du code	article 700	700	lieu à application

TABLE – premiers termes pour chaque catégorie (de gauche à droite)

$$n_{gl}(w, c) = \frac{\sqrt{N}((N_{w,c}N_{\bar{w},\bar{c}})-(N_{w,\bar{c}}N_{\bar{w},c}))}{\sqrt{N_w N_{\bar{w}} N_c N_{\bar{c}}}} \quad [?]$$

Evaluation de l'extraction de demande-résultat

$$\forall D_j \in D_+ = \{D_1, D_2, \dots, D_{|D_+|}\}$$

avec D_+ : l'ensemble des documents annotés et contenant la catégorie

$$Precision_{I_s, j} = \frac{\#I_s \text{ bien identifiées dans } D_j}{\#I_s \text{ identifiées dans } D_j} = \frac{TP_{I_s, j}}{TP_{I_s, j} + FP_{I_s, j}}$$

$$Rappel_{I_s, j} = \frac{\#I_s \text{ bien identifiées dans } D_j}{\#I_s \text{ dans la vérité terrain de } D_j} = \frac{TP_{I_s, j}}{TP_{I_s, j} + FN_{I_s, j}}$$

avec $\#I_s$ = nombre de tuples d'informations dont le type appartient à I_s

$$Precision_{I_s} = \frac{\sum_{j=1}^{|D_+|} Precision_{I_s, j}}{|D_+|}$$

$$Rappel_{I_s} = \frac{\sum_{j=1}^{|D_+|} Rappel_{I_s, j}}{|D_+|}$$

$$F1_{I_s} = 2 \times \frac{Precision_{I_s} \times Rappel_{I_s}}{Precision_{I_s} + Rappel_{I_s}}$$

Résultats actuels de l'extraction de demande-résultat

Catég.	I_s	Train				Test			
		P	R	F1	Doc	P	R	F1	Doc
acpa	Q-DMD	0.607	0.643	0.624	0.571	0.611	0.667	0.638	0.556
	SENS RST	0.964	1.0	0.982	0.929	0.833	0.889	0.86	0.778
	Q-RST	0.964	1.0	0.982	0.929	0.833	0.889	0.86	0.778
concdel	Q-DMD	0.361	0.4	0.38	0.333	0.167	0.167	0.167	0.167
	SENS RST	0.833	0.596	0.695	0.444	0.819	0.715	0.764	0.417
	Q-RST	0.843	0.651	0.735	0.444	0.819	0.715	0.764	0.417
danais	Q-DMD	0.312	0.343	0.327	0.271	0.38	0.392	0.386	0.367
	SENS RST	0.88	0.922	0.901	0.797	0.905	0.911	0.908	0.861
	Q-RST	0.901	0.938	0.919	0.805	0.918	0.924	0.921	0.873
dcppe	Q-DMD	0.078	0.078	0.078	0.078	0.059	0.059	0.059	0.059
	SENS RST	0.961	0.873	0.915	0.824	0.912	0.797	0.85	0.735
	Q-RST	0.775	0.711	0.741	0.667	0.765	0.65	0.702	0.588
doris	Q-DMD	0.105	0.083	0.093	0.079	0.18	0.148	0.162	0.08
	SENS RST	0.553	0.456	0.5	0.368	0.5	0.385	0.435	0.28
	Q-RST	0.579	0.482	0.526	0.395	0.54	0.395	0.456	0.28
styx	Q-DMD	0.733	0.633	0.68	0.533	0.575	0.525	0.549	0.4
	SENS RST	0.983	0.783	0.872	0.567	0.825	0.642	0.722	0.35
	Q-RST	0.983	0.783	0.872	0.567	0.85	0.642	0.731	0.35

$I_s \subset I = \{Q-DMD, SENS RST, Q-RST\}$

$P, R, F1 = \text{resp. Precision}_{I_s}, \text{Rappel}_{I_s}, F1_{I_s}$

$Doc = \text{proportion de documents dans lesquels toutes les infos ont été bien identifiées}$

Code couleur : **bleu** : mesure ≥ 0.85 ; **rouge** : mesure < 0.5

Conclusion

Résumé des résultats actuels

- Bons résultats pour la détection d'entités et de sections à base de CRF
 - Difficultés :
 - Annotation manuelle d'un jeu suffisant d'exemples
 - Identification de bons descripteurs
 - Limite de l'approche :
 - Descripteurs définis manuellement
 - Etiquetage en plusieurs passes
- Premiers résultats encourageants sur l'extraction de demande par catégorie :
 - Identification facile de la présence d'une catégorie par classification binaire
 - Le sens et le quantum résultat semblent facile à identifier
 - Difficultés : identification des termes-clés d'expression des demandes et résultats

Résumé des tâches du projet

1. Extraction d'information : métadonnées, demande-résultat, situations factuelles d'une catégorie de demande (par ex. *licenciement* ou *divorce*)
2. Standardisation et représentation des informations extraites sous forme de base de connaissances
3. Détermination des facteurs associables aux décisions des juges

Questions ?

References I



Frank, E., Hall, M. A., and Witten, I. H. (2016).

The WEKA Workbench, chapter Online Appendix for "Data Mining : Practical Machine Learning Tools and Techniques".

Morgan Kaufmann.



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.



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

Learning hidden Markov model structure for information extraction.

AAAI-99 Workshop on Machine . . .



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

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

In *In Extraction et Gestion des Connaissances - EGC 2017, Revue des Nouvelles Technologies de l'Information, Grenoble, France*.