

Extraction d'information à partir de décisions judiciaires

Journées des Doctorants du LGI2P – 22 juin 2017

TAGNY NGOMPE Gildas

Début de thèse: 15 Décembre 2015

Direction de thèse:

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

Encadrement de proximité:

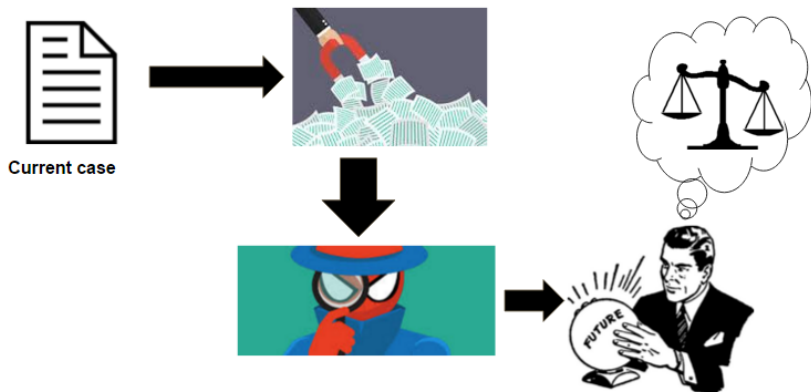
- Sébastien Harispe (Ecole des Mines d'Alès, LGI2P)
- Guillaume Zambrano (Univ. Nîmes, CHROME)



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



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 trouvées

☐ 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

Pas de restriction de date

Source : LexisNexis.com

Défis : Documents non-structurés

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,

...

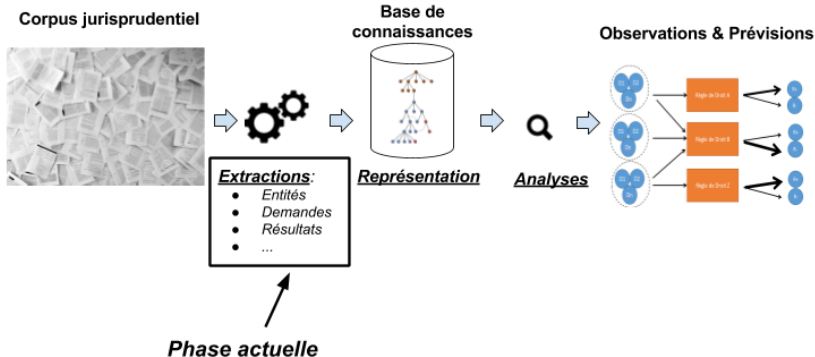
Déboute Madame A. de sa demande de
provision sur dommages-intérêts.

...

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



Elaboration et mise en oeuvre de techniques de :

- Traitement du langage naturel
- Représentation des connaissances
- Analyse prédictive

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.

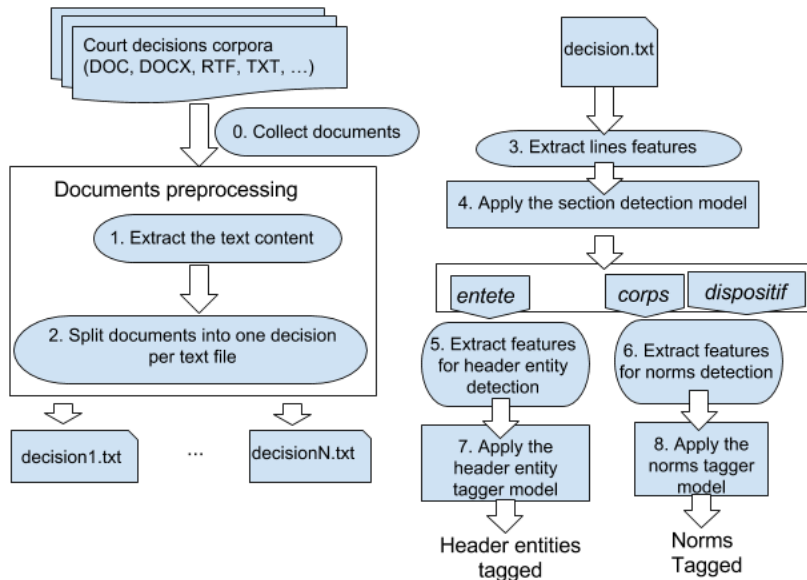
...

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

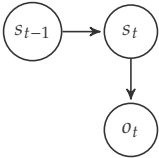
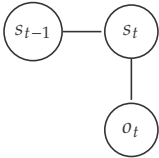
Entêtes : méta-données

Architecture à 2 passes



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>[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

- Utilité de la prise en compte des particularités des textes
 - forme : le mot est-il en majuscule, lemmes, longueur de la ligne, ...
 - contexte : mots voisins, position par rapport à un mot-clé, ...
- Certaines entités restent difficiles à détecter

Comment améliorer les résultats ?

Définir plus de caractéristiques :

- 14 pour les sections
- 35 pour les entêtes
- 28 pour les normes

Sélection des caractéristiques

Algorithm 1: Bidirectional search algorithm

[Liu and Motoda, 2012]

Data: annotated dataset, X list of all the candidate features

Result: optimal feature subset

```
1 Start SFS with  $Y_{F_0} = \emptyset$ ;  
2 Start SBS with  $Y_{B_0} = X$ ;  
3  $k = 0$ ;  
4 while  $Y_{F_k} \neq Y_{B_k}$  do  
5      $x^+ = \operatorname{argmax}_{\substack{x \notin Y_{F_k} \\ x \in Y_{B_k}}} F1measure(Y_{F_k} + x)$ ;  
6      $Y_{F_{k+1}} = Y_{F_k} + x^+$ ;  
7      $x^- = \operatorname{argmax}_{\substack{x \notin Y_{F_{k+1}} \\ x \in Y_{B_k}}} F1measure(Y_{F_k} - x)$ ;  
8      $Y_{B_{k+1}} = Y_{B_k} - x^-$ ;  $k = k + 1$ ;  
9 return  $Y_{F_k}$ ;
```

Réduit de **moitié** le nombre de caractéristiques

Améliore légèrement les résultats

Mais **très lent** (plus de 10 h lors de nos tests)

Nombre nécessaire de données d'entraînement

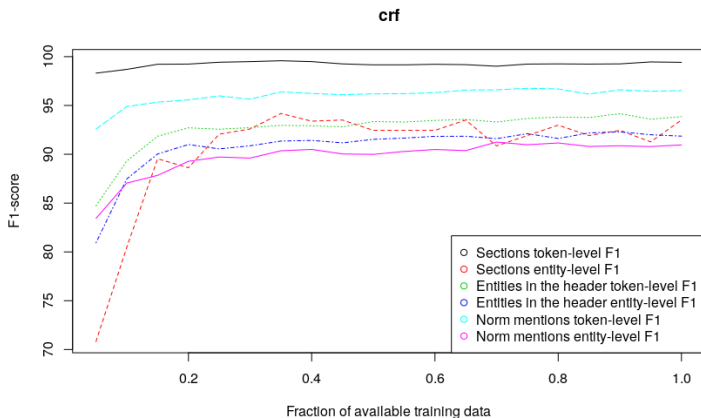


FIGURE — Résultats en fonction du nombre de données d'entraînement (fractions d'environ 380 décisions)

Extraction d'informations sur les demandes

Informations pertinentes à extraire

- **Position de la partie** : Intimé
- **Catégorie de demande** : 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
- **Résultat** : Rejet
- **Quantum accordé** : 0 euros

Difficultés (1)

Expressions explicite / implicite

EXEMPLE : 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 ...

...

EXPRESSION DE RESULTAT

La cour, ...

Confirme la décision entreprise en toutes ses dispositions,

- Plusieurs catégories similaires ou différentes dans une décision
- Toutes les catégories ne sont pas connues d'avance

Simplification du probleme

- On suppose qu'une décision ne comprend qu'au plus une demande d'une catégorie donnée
- Méthode générique qui s'adapte aux spécificités de la catégorie traitée
- Définition incrémentale des catégories

Détection d'une catégorie par classification

(1) Sélection de termes caractéristiques

DOMMAGES-INTERETS POUR ABUS DE PROCEDURE

Terme (n-gram)	Poids global (NGL)
procédure abusive	15.710
pour procédure abusive	15.007
pour procédure	14.890
abusive	13.721
intérêts pour procédure	10.306
abus	10.288
intérêts pour procédure abusive	9.984
32-1	9.534
...	...

$$n gl(w, c) = \frac{\sqrt{N}((N_{w,c}N_{\bar{w},\bar{c}})-(N_{w,\bar{c}}N_{\bar{w},c}))}{\sqrt{N_w N_{\bar{w}} N_c N_{\bar{c}}}} \text{ [Ng et al., 1997]}$$

Détection d'une catégorie par classification

(2) Représentation vectorielle & Classification des décisions

$$poids(w^*, t) = poids_{local}(w^*, t) * poids_{global}(w^*) * facteur_{normalisation}$$

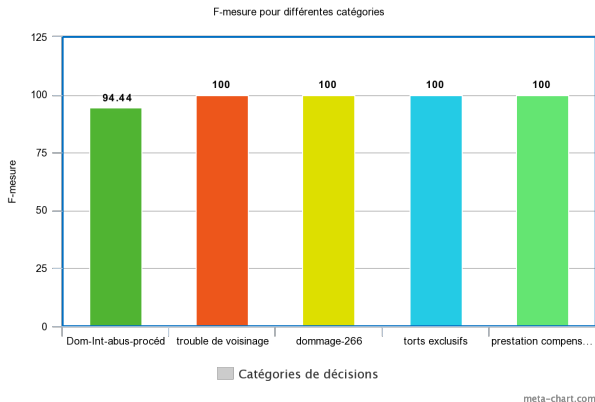


FIGURE — Résultats des meilleures configurations (taille des vecteurs, poids global, poids local, modèle de classifieur)

Interprétation des résultats pour une catégorie

Tentative par classification des décisions

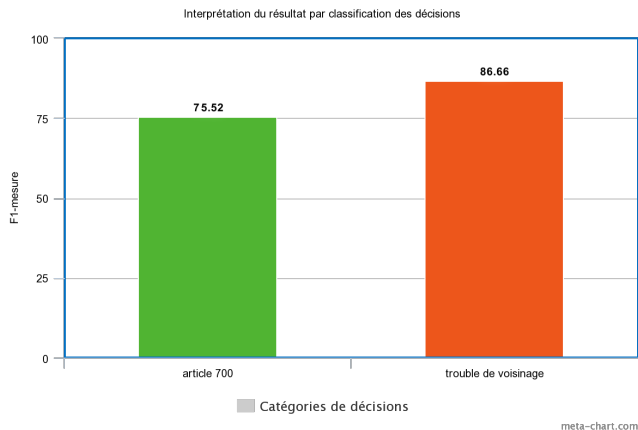


FIGURE – Résultats des meilleures configurations (taille des vecteurs, poids global, poids local, modèle de classifieur)

Conclusion

- Détection d'entités et de sections basée HMM / CRF
 - Bons résultats même avec un peu de données annotées
 - Difficultés :
 - Annotation manuelle d'un jeu suffisant d'exemples
 - Identification de bons descripteurs
 - Lenteur de la sélection de caractéristiques
 - Limite de l'approche :
 - Descripteurs définis manuellement
 - Etiquetage en plusieurs passes
- Détection de termes propres aux catégories de demandes
- Détection des catégories par classification
- Détection moins triviale du sens du résultat

1. Extraction et catégorisation non-supervisée des demandes
2. Désambiguïsation et représentation des informations
3. Détermination des facteurs "*expliquant*" la décision des juges

Questions ?

References I



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

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

Springer Science & Business Media.



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*.

Structure dans la base de connaissances

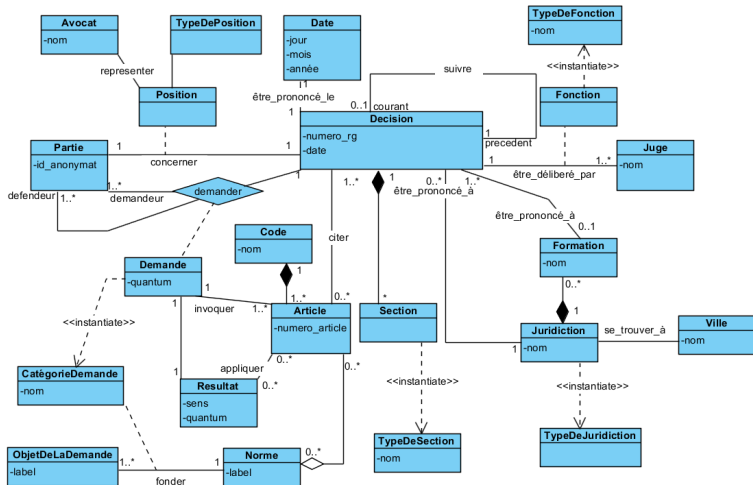


FIGURE – Modèle des données

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.