Institut de Recherche en Informatique de Toulouse
CNRS - INP - UT3 - UT1 - UT2J

Rapport sur le mémoire de thèse de Gilgas TAGNÉ NGOMPÉ, intitulé « Méthodes d'analyse sémantique de corpus de décisions jurisprudentielles »

1 Contexte du travail présenté

Les travaux décrits dans ce mémoire rentrent dans le contexte de l'extraction d'information dans des textes et s'intéressent plus précisément à des textes jurisprudentiels. Ces textes, décrivant des décisions de justice, rapportent notamment les faits, les demandes des plaignants, les procédures judiciaires, le verdict des juges. Ces informations, disséminées dans le document, sont souvent collectées manuellement par différentes parties (juges, avocats, plaignants, ...) pour les exploiter dans différents cadres, par exemple identifier des situations similaires à leur situation d'intérêt.

L'objectif de ces travaux est, précisément, d'élaborer des approches permettant l'extraction automatique de ces informations. Le candidat propose différentes approches pour extraire différents types d'informations tels que les entités (avocat, juges, dates, ...), les demandes et les résultats. Il a élaboré ces approches en s'appuyant sur différentes hypothèses. Sans être exhaustif, dans la première il considère que bien qu'il n'existe pas de structure explicitée, les documents sont, dans leur majorité, rédigés avec une certaine structure latente, on retrouve en effet, un entête contenant les entités, suivi d'un corps détaillant les faits, les procédures, les conclusions et enfin les résultats (les réponses aux demandes). La deuxième stipule qu'il existe des indicateurs (des termes-clés) ou des caractéristiques qui permettent d'identifier le changement de sections dans le document ainsi que des informations fines telles les montants d'argent (quantum) demandés/accordés, le sens de la décision (acceptée/rejetée/sans). Enfin, la troisième considère que des décisions similaires d'une catégorie d'intérêt donnée (i.e dommages et intérêts) partagent les mêmes mots-clés. Le candidat exploite ces différentes hypothèses pour proposer quatre approches permettant respectivement, de détecter les sections d'un document et reconnaître les entités, d'extraire les demandes et les résultats correspondants, d'extraire le sens du résultat et enfin de catégoriser les décisions. Ces différentes approches sont mises à l'épreuve sur un jeu de de données annotées par des experts dans l'environnement du candidat et plusieurs résultats sont listés et étayés par des analyses pertinentes.

2 Analyse du mémoire

Le mémoire comporte 6 chapitres, une introduction et une conclusion. Le chapitre 1 liste quelques classes d'approches d'analyse de textes, les chapitres 2, 3, 4 et 5 sont consacrés aux contributions proprement dites, le chapitre 6 donne quelques statistiques descriptives extraites d'un corpus de décisions issues de de la base *CAPP*.

L'introduction aborde le contexte et la problématique de l'étude. Elle définit le document judiciaire, met en exergue les différentes informations utiles qu'il comporte puis souligne l'intérêt et la difficulté de les extraire automatiquement. Les contributions et le plan de la thèse sont ensuite listés à la fin de cette introduction.

Le chapitre **un** traite de l'analyse automatique de textes judiciaires. Il se focalise sur les trois tâches d'analyse qui y sont exploitées tout au long de la thèse. Il s'agit de l'annotation et l'extraction d'information, la classification et la similarité entre décisions. Pour chacune de ces tâches un bref état de l'art spécifique aux décisions judiciaires est décrit. Ce chapitre donne une bonne synthèse des différents classes d'approches d'analyse de corpus judiciaires.

Les chapitres contributions suivent la même structure, ils décrivent tout d'abord les travaux connexes et les cadres théoriques qui y sont exploités, puis détaillent la contribution ainsi que les expérimentations.

Le chapitre **deux** se focalise sur la première contribution, il s'agit de définir des approches pour la segmentation du document judiciaire en sections et l'extraction des entités juridiques qui y sont référencées. On rappelle que ces décisions sont des textes bruts sans structure prédéfinie, le candidat considère que bien que la structure ne soit pas explicitée, ces décisions suivent tout de même (assez souvent) une certaine structure de surface. Elles comportent, en effet, un entête, un corps et un dispositif (synthèse du résultat final). De plus, il existe des termes clés dont les occurrences dans le texte permettent d'identifier le changement de sections et reconnaître les entités judiciaires. Le candidat propose d'appliquer deux modèles graphiques probabilistes, en l'occurrence *HMM* et *CRF* pour répondre à ces deux tâches. Deux ensembles de descripteurs, un par type de tâche, sont proposés et testés sur une collection de décisions judiciaires annotées dans l'environnement du candidat. Plusieurs expérimentations évaluant différents impacts (des deux modèles HMM et CRF, des algorithmes de sélection de descripteurs) sont présentées. De même, les descripteurs manuels sont comparés à des descripteurs construits par une approche neuronale. Il en ressort de manière claire, sur les modèles, la supériorité des CRF vis-à-vis de HMM, ces modèles restent performants même en prenant un ensemble de descripteurs très réduit. Au delà de ces résultats, le candidat a réalisé une analyse fine et pertinente des différents résultats pour mieux comprendre les limites des approches proposées vis-à-vis par exemple de la quantité d'exemples annotés.

Le chapitre **trois** traite de l'identification des éléments structurant une demande. Il s'agit précisément d'extraire la catégorie de la demande (i.e dommage et intérêts,...) ainsi que la demande, ce que réclament les parties (quantum demandé) et la réponse du juge (quantum obtenu et le sens de la décision). La problématique majeure dans cette tâche réside dans le fait qu'une décision judiciaire peut comporter plusieurs demandes. Donc au delà de la difficulté d'identifier individuellement les différents attributs (quanta), il faut aussi arriver à les relier (i.e relier le quantum demandé, le quantum obtenu et le sens de la décision (accepté ou rejeté)). Le cœur de la contribution ici concerne l'identification des attributs (quanta). L'approche proposée s'appuie sur la proximité potentielle, dans les texte, de ces attributs avec des termes-clés. Ne disposant pas explicitement de ces termes, le candidat propose de les extraire directement à partir de textes annotés en se basant sur des techniques de pondération de termes. Le lien entre la demande et le résultat (le quantum demandé, le quantum accordé et le sens du résultat) est réalisé en fonction de la similarité des énoncés de la demande et du résultat, et en fonction de l'ordre de grandeurs des quanta (relier ceux qui sont dans le même ordre). Plusieurs expérimentations mettant à l'épreuve les propositions ont été réalisées sur la collection annotée (mentionnée ci dessus) aussi bien pour la phase de catégorisation que l'extraction des paires demandes-résultats. Il en ressort que l'approche d'extraction des attributs arrive à identifier les attributs individuellement, l'identification des paires est plus ardue. Les résultats dépendent fortement des données de la catégorie considérée. On constate en effet, que les 6 catégories de documents sont très disparates en particulier, en termes de nombre de demandes par document. l'approche s'en sort

mieux avec quand des documents comportant une seule demande. Le candidat analyse les raisons de ces erreurs et souligne par la même la difficulté de la tâche.

Dans le chapitre précédent, le candidat a bien montré la difficulté d'extraire les demandes et les résultats d'une décision. Dans ce chapitre **quatre**, il se limite à à extraire uniquement le sens de la décision, acceptée ou rejetée. Le candidat revient sur les limites de l'approche basée sur les termes-clés relatives en particulier au fait qu'elle soit dépendante des données d'apprentissage, donc difficilement généralisable. Il s'oriente vers des approches classiques de classification binaire. En préambule, il rappelle brièvement quelques algorithmes de classification (i.e NB, SVM,KNN, Arbre de décisions...) les plus répandus dans la littérature. Il propose ensuite d'adapter un classifieur produit dans son environnement de recherche, en l'occurrence GINI-PLS, pour une classification textuelle. Il a mis à l'épreuve et comparé une panoplie de classifieurs (une douzaine) utilisant différentes pondérations des termes. Ces expérimentations sont réalisées sur des documents du jeu de données comportant uniquement une demande. Les résultats montrent des performances variables, mais globalement les approches arbres de décisions sont les plus performantes quelle que que soit la catégorie. Dans son analyse le candidat souligne également que les performances dépendent de la partie textuelle fournie au classifieur. Il montre dans ce cas que même si les approches basées sur les arbres restent performantes, les algorithmes proposés par le candidat n'en sont pas loin.

Le chapitre **cinq** aborde une question transverse importante portant sur la catégorisation des décisions judiciaires. En effet, regrouper des décisions similaires permet, entre autres, à un juge, un avocat, ou un individu de retrouver des décisions proches de sa situation d'intérêt. La majorité des approches de catégorisation textuelle se base sur la comparaison de textes. C'est précisément, là où se situe la contribution décrite dans ce chapitre. Le candidat propose une mesure de similarité textuelle construite par apprentissage automatique. Cette mesure, qui rassemble dans l'esprit à une distance de *levenshtein*, consiste à mesurer le nombre de transformations nécessaires pour passer d'un document à un autre. Comme cette distance est apprise automatiquement, et qu'il n'existe pas de jeu d'entraînement, la solution judicieuse proposée est de construire des documents synthétiques par mutation de documents originaux. Plusieurs expérimentations sont listées, tout d'abord pour identifier la meilleure représentation des textes. La distance proposée est ensuite comparée à plusieurs métriques de la littérature dans le cadre d'une tâche de catégorisation. Les résultats montrent clairement sa supériorité dans la majorité des cas considérés. Le résultat important ici concerne la capacité de cette distance, utilisée avec des Kmoyennes, à trouver automatiquement le bon nombre de clusters.

En conclusion, ces quatre chapitres listent des contributions et des résultats à la fois pertinents et intéressants. Tout au long de ces chapitres, le candidat a eu le souci de décrire le cadre et les modèles qu'il exploite avant de détailler ses approches. Chaque proposition a été expérimentée et comparée avec des travaux connexes. Les résultats ont toujours été discutés et analysés pour comprendre les raisons des erreurs ou des faiblesses des performances. Ces expérimentations sont réalisées sur un jeu de données annotées par des experts. La limite de ces travaux, comme ceci a été souligné par le candidat, réside dans la taille de ce jeu de données qui est assez réduite, en particulier quand les données sont traitées par catégorie. Par conséquent, les conclusions que l'on peut tirer peuvent en être affectées. Je suis bien évidemment conscient de la difficulté d'obtenir ce type de collection, il existe effectivement des corpus de plus grande taille tel que celui décrit dans le chapitre 6 mais il n'est pas annoté. Le chapitre **six** présente en effet quelques statistiques descriptives sur un grand corpus de décisions issues de la *CAPP* de la *DILA*.

La conclusion générale résume les contributions apportées, leurs limites et liste les perspectives. Le document est complété par une bibliographie riche et récente.

3 Conclusion

Les travaux décrits dans ce mémoire portent sur une question fondamentale, l'extraction d'information dans des documents judiciaires. Tout au long du document le candidat a clairement mis en exergue la multitude d'informations que peut comporter une décision judiciaire et l'intérêt de les extraire de manière automatique. Plusieurs approches ont été proposées chacune d'elle est élaboré pour extraire et identifier une partie particulière de l'information recherchée. La majorité de ces approches exploitent des cadres théoriques pertinents. Il est également à noter l'important travail d'expérimentation et de validation qui a été réalisé. Sur le plan plan rédactionnel, le document est clair et bien structuré.

En conséquence, compte tenu de l'ensemble de ces remarques, j'émets un avis favorable à la soutenance de la thèse de Gildas Tagny NGompé en vue de l'obtention du Doctorat en Informatique.

Fait, à Toulouse, le 9 Janvier 2020

M. Boughanem

Professeur à l'université Paul Sabatier de Toulouse

Université Paul. Sabatier
IRIT
118 route de Narbonne
31062 TOULOUSE CEDEX 9
Tél. 33 (0)5 61 55 67 65