Classification One-Shot faiblement supervisée par réseaux de neurones récurrents avec attention : application à la détection de résultat juridique

Charles Condevaux [1*], Sébastien Harispe[**]
Stéphane Mussard[*] Guillaume Zambrano[*]

[*]CHROME Univ. Nîmes, France
charles.condevaux@unimes.fr
[**]LGI2P, IMT Mines Alès, Univ. Montpellier, Alès, France
sebastien.harispe@mines-ales.fr

Résumé. Déterminer si une demande juridique est acceptée à l'aide d'arguments énoncés par un juge est une tâche importante dans l'analyse de décisions de justice. L'application de techniques modernes d'apprentissage automatique peut toutefois s'avérer inappropriée pour résoudre ce type de problème car dans le domaine juridique, les jeux de données étiquetés sont le plus souvent de très petite taille, rares, et coûteux à construire. Cet article présente un modèle d'apprentissage profond et une méthodologie pour résoudre des tâches de classification en traitement du langage avec peu d'exemples étiquetés. Nous montrons en particulier que le fait de combiner un apprentissage *one-shot* avec des réseaux de neurones récurrents et un mécanisme d'attention permet d'obtenir des modèles performants. Les résultats présentés portent sur le traitement de plusieurs catégories de réclamations devant les tribunaux français et par le biais de différents processus de vectorisation pour la représentation des phrases.

1 Introduction

La classification de textes est un thème prédominant de la littérature relative au traitement automatisé de corpus de documents légaux (arrêts, décisions, contrats, règlements…) (Gonçalves et Quaresma, 2005). Elle est par exemple utilisée pour faciliter la recherche dans les corpus juridiques en regroupant les décisions passées suivant une organisation prédéfinie – les avocats recherchent souvent des affaires similaires lors de l'étude d'un cas spécifique (Brüninghaus et Ashley, 1999). Cet article présente plusieurs architectures de modèle d'apprentissage automatique ainsi qu'une méthodologie adaptées à l'étude de tâches complexes de classification de textes lorsque seuls quelques exemples étiquetés sont disponibles. Une application des modèles et de l'approche proposés est effectuée dans un contexte de prédiction de résultats juridiques dans des décisions de justice françaises. Dans ce contexte, nous étudions la manière de combiner un apprentissage *one-shot* avec des réseaux de neurones récurrents et un mécanisme d'attention afin d'obtenir des modèles de classification efficaces.

1. Granted by Région Occitanie: project PREMAT-TAJ.

2 Travaux antérieurs

La littérature dédiée à la classification de texte distingue classiquement les approches dites à base de règles de celles dites à base d'apprentissage machine (Waltl et al., 2017). Nous nous focaliserons sur celles à base d'apprentissage automatique par la suite. Elles répondent aux limitations de l'identification manuelle de règles en permettant l'identification automatique de modèles prédictifs capables de réaliser la classification. Une approche supervisée est adoptée la plupart du temps [2] : à partir d'un jeu de données labélisé (e.g. décisions classées), des algorithmes d'apprentissage sont utilisés pour distinguer des modèles prédictifs qui permettront par la suite de classer de données non-annotées.

Parmi les modèles les plus fréquemment utilisés en classification de textes ces vingt dernières années, nous pouvons citer : le classifieur bayésien naïf, la régression logistique, les forêts aléatoires, les machines à vecteur de support (SVM), ou le perceptron multicouche. Ces modèles requièrent une représentation vectorielle de textes qui sera considérée comme donnée d'entrée du système. Des informations statistiques peuvent être utilisées pour pallier la difficulté de définir manuellement des caractéristiques pour construire ces représentations. Des simples modèles de type sac-de-mots et vectoriels de la fin des années 60, aux schémas de pondérations capables de modéliser la pertinence de la prise en compte de certains termes pour la classification (e.g. TF-IDF), ces approches ont amené des modèles capables de résoudre des problèmes de classification dans de nombreux contextes. Malgré ces succès encourageants, des problématiques de classification restent hors de portée de ces méthodes.

Les récents développements en Apprentissage Profond (*Deep Learning*) ont produit une diversité d'approches novatrices basées sur des réseaux de neurones. Les réseaux de neurones récurrents, tels que les LSTM (*Long Short-Term Memory*) sont par exemple performants pour le traitement de données séquentielles. L'apprentissage profond s'intéresse aussi aux techniques d'apprentissage de représentations (*embeddings*) afin d'encoder l'information portée par des entités sémantiques et ainsi les représenter dans des espaces de tailles réduites (e.g. BERT, ELMo, FastText). Des mécanismes d'attention sont aussi développés afin de mieux identifier et incorporer des informations importantes lors du processus décisionnel intervenant dans la tâche de classification. Ils ont en particulier permis l'obtention d'améliorations intéressantes de la performance des systèmes, et cela pour de nombreux challenges offerts à la classification de textes (Bahdanau et al., 2014; Luong et al., 2015; Vaswani et al., 2017).

Cependant, les modèles à base d'apprentissage profond nécessitent de larges quantités de données (labélisées) pour être entraînés (Slingerland et al., 2018). Cet écueil est central et limite sévèrement leur potentielle utilisation dans le domaine légal où il est souvent difficile de mobiliser des experts. Des recherches actives en apprentissage machine travaillent sur la réduction de cette dépendance aux volumétries importantes de données labélisées, notamment en tirant parti autant que possible de l'information véhiculée par les annotations disponibles (e.g. *one-shot learning*). Une stratégie expérimentée dans ce papier propose de tirer parti de l'apprentissage de type *one-shot* pour résoudre une tâche de classification sur petits échantillons (Fei-Fei et al., 2006). Au lieu d'apprendre directement une correspondance entre une donnée fournie en entrée et une classe, l'approche *one-shot* basée sur des réseaux siamois se focalise sur l'estimation d'une fonction de similarité entre deux observations (Bromley et al., 1994).

2. Nous nous concentrons ici sur la modélisation commune du problème de classification de textes sur la base du paradigme de l'apprentissage machine supervisé ; des approches alternatives sont étudiées à la marge.

Ce problème peut être réduit à une tâche de classification binaire en considérant que l'objectif vise à reconnaître des données d'entrée similaires. Le développement de ce type d'approches doit selon nous être encouragé dans le domaine légal de manière à ce que ce dernier puisse pleinement tirer parti des récentes avancées en apprentissage machine.

3 Apprentissage one-shot par réseaux siamois avec attention

Le modèle proposé implémente un apprentissage *one-shot* par le biais de réseaux récurrents siamois dotés d'un mécanisme d'attention afin d'obtenir de bonnes représentations de phrases. Celles-ci seront ensuite réutilisées conjointement à des variables préalablement sélectionnées afin de résoudre le problème de classification, dans notre cas la prédiction du résultat binaire associé à une demande. Nous présentons dans un premier temps l'architecture générale pour par la suite nous attarder sur son utilisation pour répondre à la tâche de classification étudiée.

3.1 Architecture générale

L'architecture générale du modèle est présentée en Figure 1. Nous nous concentrons dans un premier temps sur la partie haute de la figure qui illustre le réseau siamois proposé. L'objectif de ce réseau est d'estimer une distance qui permettra de distinguer si deux observations appartiennent à une même classe.

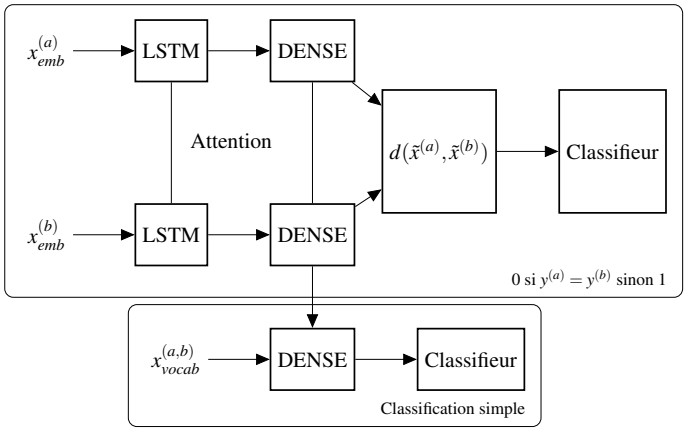

FIG. 1 – *Architecture en réseaux siamois présentée.*

Le réseau siamois est composé de deux sous-réseaux symétriques ; ceux-ci partagent les mêmes poids mais prennent différentes données en entrée. L'architecture choisie repose sur une couche de cellules LSTM bidirectionnelles prenant en entrée des phrases pré-vectorisées ($x^{(a)}_{emb}$ et $x^{(b)}_{emb}$). Un mécanisme d'attention est ajouté afin de permettre au réseau d'identifier des portions d'intérêt. Chaque mot est pondéré selon la version concaténée ou multiplicative de Luong (Luong et al., 2015) en fonction des données et de l'architecture choisie. Nous utilisons une approche d'attention dite *many-to-one* afin d'obtenir une matrice indépendante de la

longueur de la séquence. La projection obtenue correspond à la représentation qui sera par la suite exploitée dans la tâche de classification dite principale. L'apprentissage *one-shot* nécessite le calcul d'une fonction de distance entre paires d'observations analysées dans le réseau siamois, i.e. $d(\tilde{x}^{(a)}, \tilde{x}^{(b)})$. Deux fonctions ont été sélectionnées parmi les nombreuses testées. Ces fonctions sont appliquées indépendamment sur chaque variable : la première est basée sur la distance en valeur absolue, la seconde utilise une version modifiée de la distance de cosinus. Si les sommes pondérées de ces distances sont minimales, la sigmoïde renverra des valeurs proches de 0 : les observations partageront alors la même étiquette.

Puisque les jeux de données utilisés pour cette tâche sont petits (moins de 100 observations), le nombre de paramètres du réseau doit être limité pour contenir le surapprentissage. Des couches de faibles tailles et une technique de *dropout* (0.25) ont été utilisées. La couche de LSTM est de taille (16 x 2) couplée à celle d'attention composée de 16 entrées. [3]

3.2 Classification

La partie haute du réseau peut être utilisée de façon indépendante pour de la prévision. Une nouvelle observation peut en effet être comparée à d'autres observations issues du jeu d'entraînement. La même étiquette que celle associée à la phrase ou aux phrases considérées comme les plus proches peut ainsi être appliquée. L'utilisation d'un second classifieur est cependant plus stable et permet l'obtention de meilleures performances.

La sortie de la couche dense du réseau précédant la fonction du distance peut s'apparenter à une vectorisation de taille fixe de la phrase proposée en entrée. Dans notre architecture, nous enrichissons cette représentation en la concaténant à un ensemble de mots jugés discriminants pour la classification ($x_{emb}^{(a,b)}$). La sélection des mots est effectuée en comparant leur fréquence d'apparition dans chaque classe. Comme le problème est binaire (le juge accepte ou rejette une demande), les fréquences sont définies comme ceci, f_c^w est la fréquence du mot w dans les phrases issues de la catégorie c. Les mots discriminants maximisent la différence $|f_i^w - f_j^w|$.

4 Données et représentions des mots

Dans cette section, nous décrivons les jeux de données utilisés pour la tâche de classification et la configuration de l'expérimentation. Le pipeline de prétraitement et la façon dont les mots spécialisés sont construits sont également présentés.

4.1 Les données

Cinq jeux de données quasi-équilibrés ont été choisis pour couvrir différents types de demande. Ils ont été annotés manuellement par des juristes qui ont étiqueté la zone où le juge présente ses arguments pour la demande spécifique et le résultat associé (acceptation ou rejet de la demande). Les ensembles de données sont équilibrés par rapport aux demandes de changement de nom (600. NOM, 74 observations), aux dettes impayées (600.DEC, 96 observations), à la responsabilité des avocats (500.RES, 400.RES, 100 observations chacune) et aux actions en dommages et intérêts pour blessures graves (300.DOM, 98 observations).

3. Ce nombre est doublé lorsqu'une stratégie d'augmentation de données est appliquée par la suite.

4.2 Représentations des mots

Nous comparons différentes méthodes de représentation allant de la simple approche fréquentielle TF-IDF aux modèles plus avancés. Nous avons pour cela construit des représentations de mots spécialisées allant de 32 à 128 dimensions ; toutes sont issues de l'analyse d'un vaste corpus composé de 670 millions de mots issus de décisions de justice (françaises) et de textes de lois. Quatre approches récentes de représentation de mots ont été étudiées : FastText (Bojanowski et al., 2016), ELMo (Peters et al., 2018), Flair (Akbik et al., 2018) ; notons toutefois que du fait des temps de calcul induits par BERT, les représentations utilisées pour ce modèle proviennent du modèle multilingue public. [4] Puisque les textes juridiques ont des structures similaires, un vocabulaire de niche et des expressions redondantes, des modèles très compacts peuvent générer une faible perplexité (< 20 pour ELMo avec 64 dimensions de base) - cela souligne le caractère prévisible du langage juridique français.

4.3 L'augmentation de données

L'augmentation de données, qui vise à enrichir artificiellement la base d'apprentissage, est considérée comme une tâche difficile en TAL car la modification d'un seul mot peut modifier radicalement le sens d'une phrase. Différentes approches ont été étudiées afin de vérifier si cette technique peut être appliquée à notre contexte d'étude impliquant du langage juridique. Parmi les stratégies étudiées, nous avons testé la substitution aléatoire de mots par leurs synonymes à l'aide d'un thésaurus général et la traduction par plusieurs langues intermédiaires. La première méthode n'offre aucun bénéfice à cause du vocabulaire spécifique employé par les juristes. La traduction offre quant à elle des gains significatifs, permettant d'obtenir des variations intéressantes tout en conservant à divers degrés la cohérence et le sens des phrases.

5 Résultats et protocoles expérimentaux

Nous comparons différentes approches afin d'évaluer l'influence de la sélection du vocabulaire et de l'approche *one-shot* sur la qualité des prévisions. Nous examinons les performances des algorithmes standards appliqués à deux types de vectorisations : TF-IDF et la moyenne des plongements couplée au vocabulaire sélectionné (Table 1). Nous discutons ensuite l'apport d'une part du modèle *one-shot* et d'autre part de l'augmentation de données (Table 3). Tous les résultats sont basés sur une validation croisée de 10 sous-échantillons et une F-mesure moyenne pondérée par la taille des classes.

5.1 Algorithmes standards

La forêt aléatoire a de meilleures performances sur chaque catégorie de demande, à l'exception de la responsabilité des avocats (500.RES) pour laquelle les classifieurs SVM et les classifieurs logistiques fournissent de meilleures F-mesures, s'élevant respectivement à 0,659 et à 0,645. Les représentations de mots sont moyennées afin de fournir le plongement lexical de la phrase, celle-ci étant concaténée avec le vocabulaire sélectionné. Cette première approche permet déjà l'obtention d'un gain de performance significatif.

4. https ://github.com/google-research/bert

	SVM		Forêt aléatoire		Logistique	
	P	F	P	F	P	F
600.NOM*	0.798	0.748	0.782	**0.786**	0.743	0.728
600.NOM**	0.820	0.781	0.867	**0.813**	0.752	0.739
600.DEC*	0.669	0.788	0.747	**0.795**	0.658	0.767
600.DEC**	1.000	0.908	0.931	**0.959**	0.889	0.902
500.RES*	0.591	0.568	0.651	0.619	0.674	**0.645**
500.RES**	0.716	**0.659**	0.623	0.636	0.639	0.570
400.RES*	0.943	0.795	0.826	**0.837**	0.810	0.828
400.RES**	0.783	0.789	0.924	**0.893**	0.709	0.736
300.DOM*	0.827	0.834	0.847	**0.854**	0.847	0.825
300.DOM**	0.963	0.931	1.000	**0.952**	1.000	0.910

* Vectorisation TF-IDF — Précision (P) et F-mesure (F)

** Plongement moyen + sélection de mots

TAB. 1 – *Comparaisons des performances de classification avec différentes entrées.*

5.2 Apport de l'apprentissage *one-shot*

Les résultats du modèle avec réseau siamois sont présentés dans la Table 2. Celui-ci exploite une couche intermédiaire du modèle *one-shot* concaténée au vocabulaire discriminant sélectionné. Cette approche sur-performe la forêt aléatoire indépendamment du modèle de plongement utilisé. BERT montre des lacunes du fait de son absence de pré-entraînement sur un corpus juridique. Les modèles sont entraînés sur des plongements de 32 dimensions, avec attention concaténée et fonction de distance ℓ_1. Le passage à une représentation de 64 dimensions n'améliore pas les performances (surapprentissage).

	FastText		ELMo		Flair		BERT	
	P	F	P	F	P	F	P	F
600.NOM	0.842	0.818	0.867	**0.846**	0.842	0.843	0.755	0.789
600.DEC	0.945	0.967	0.975	0.986	0.986	**0.992**	0.986	0.983
500.RES	0.756	0.701	0.774	**0.734**	0.805	0.709	0.610	0.686
400.RES	0.868	0.882	0.907	**0.908**	0.828	0.854	0.921	0.872
300.DOM	1.000	**1.000**	1.000	0.990	0.983	0.982	0.963	0.971

TAB. 2 – *Comparaisons des plongements lexicaux sur la classification one-shot.*

Enfin, le tableau 3 présente les résultats de la contribution de l'augmentation de données.

	Avec augmentation		Sans augmentation		Gain total	
	P	F	P	F	ΔF*	ΔF**
600.NOM	0.870	**0.880**	0.867	0.846	+0.034	+0.094
600.DEC	0.986	**0.992**	0.986	**0.992**	+0.000	+0.197
500.RES	0.794	**0.817**	0.774	0.734	+0.083	+0.172
400.RES	0.918	**0.940**	0.907	0.908	+0.032	+0.107
300.DOM	1.000	**1.000**	1.000	**1.000**	+0.000	+0.146
Moyenne					+0.030	+0.142

TAB. 3 – *Performance des modèles avec et sans augmentation. ΔF : différences de F-mesure (*) avec et sans augmentation, (**) avec augmentation et TF-IDF naïf*

L'ajout de données permet d'augmenter significativement le nombre de paires d'observations exploitables et ainsi d'autoriser plus de flexibilité dans le choix de l'architecture du réseau. Les plongements sémantiques passent à 64 dimensions, les tailles des couches sont doublées, l'attention multiplicative remplace sa version concaténée et la fonction de distance est modifiée (cosinus). Ces changements apportent des gains dans 3 des 5 catégories. Globalement, les jeux de données relatifs à la responsabilité des avocats sont les plus complexes à exploiter du fait de l'absence d'un vocabulaire discriminant. Ainsi la structure des phrases et l'agencement des mots y sont d'autant plus importants. Enfin, le passage à des plongements de 32 ou 128 dimensions ou leur fine-tuning directe impacte négativement les performances quelque soit l'approche considérée puisque le surapprentissage y est instantané.

6 Conclusion

Le réseau récurrent siamois *one-shot* proposé dans cet article, dont le but est de prédire les décisions des juges, s'avère être prometteur par rapport aux algorithmes traditionnels de la littérature. Les résultats obtenus avec les mécanismes d'attention et d'augmentation des données semblent contribuer aux bonnes performances du modèle. Ce travail pourrait ouvrir une nouvelle voie dans l'utilisation d'architectures réseaux récentes en jurimétrie, telles que les réseaux adverses, qui offrent un bon potentiel pour trouver des mots et des expressions discriminants.

Références

Akbik, A., D. Blythe, et R. Vollgraf (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, pp. 1638–1649. Association for Computational Linguistics.

Bahdanau, D., K. Cho, et Y. Bengio (2014). Neural machine translation by jointly learning to align and translate. *ArXiv 1409*.

Bojanowski, P., E. Grave, A. Joulin, et T. Mikolov (2016). Enriching word vectors with subword information. *CoRR abs/1607.04606*.

Bromley, J., I. Guyon, Y. LeCun, E. Säckinger, et R. Shah (1994). Signature verification using a "siamese" time delay neural network. In J. D. Cowan, G. Tesauro, et J. Alspector (Eds.), *Advances in Neural Information Processing Systems 6*, pp. 737–744. Morgan-Kaufmann.

Brüninghaus, S. et K. D. Ashley (1999). Toward adding knowledge to learning algorithms for indexing legal cases. In *Proceedings of the 7th International Conference on Artificial Intelligence and Law*, ICAIL '99, New York, NY, USA, pp. 9–17. ACM.

Fei-Fei, L., R. Fergus, et P. Perona (2006). One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell. 28*(4), 594–611.

Gonçalves, T. et P. Quaresma (2005). Is linguistic information relevant for the classification of legal texts ? In *Proceedings of the 10th International Conference on Artificial Intelligence and Law*, ICAIL '05, New York, NY, USA, pp. 168–176. ACM.

Luong, M., H. Pham, et C. D. Manning (2015). Effective approaches to attention-based neural machine translation. *CoRR abs/1508.04025*.

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, et L. Zettlemoyer (2018). Deep contextualized word representations. *CoRR abs/1802.05365*.

Slingerland, R., A. Boer, et R. Winkels (2018). Analysing the impact of legal change through case classification. In *Proc. of the 31st International Conference on Legal Knoweledge and Information Systems (JURIX)*, pp. 121–1230.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, et I. Polosukhin (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc.

Waltl, B., J. Muhr, I. Glaser, E. S. Georg Bonczek, et F. Matthes (2017). Classifying legal norms with active machine learning. In *Proc. of the 30st International Conference on Legal Knoweledge and Information Systems (JURIX)*, pp. 11– 20.

Summary

Determining if a claim has been accepted based on the arguments given by judges is an important task for analyzing Law. Applying recent machine learning techniques to automatize this task is however challenging because of the difficulty to obtain labeled dataset in the legal domain - datasets are indeed most often rare, small and expensive in that domain. This article introduces a deep learning model and a methodology to tackle classification tasks in Natural Language Processing when only few labeled examples are available. We show in particular that combining one-shot learning with memory-augmented recurrent neural networks enabled obtaining efficient classification models in such constraining supervised settings. Experimental results and empirical evaluations are proposed using different approaches to represent sentences dealing with several categories of claims expressed in French courts.