Détection d'anomalies sur des documents juridiques contractuels

Ahmed El Azhar Jebbari[*,**], Jean-Charles Lamirel[****,**]
Bart Lamiroy[***,**], Aurélie Vallereau[*]

[*]Batt & Associés, 29 Rue Mangin, 57000 Metz
[**]LORIA – Laboratoire Lorrain de Recherche en Informatique et ses Applications
[***]Université de Reims Champagne-Ardenne, CReSTIC
[****]Université de Strasbourg
elazhar.jebbari@batt.eu, jean-charles.lamirel@loria.fr,
bart.lamiroy@univ-reims.fr, aurelie.vallereau@batt.eu

Résumé. Cette étude vise à identifier différents types d'anomalies dans des corpus d'images de contrats juridiques de structure homogène, tels que des ensembles de contrats provenant d'une même source. Pour cela, nous nous appuyons sur une combinaison de méthodes d'analyse structurelle et de méthodes d'analyse sémantique. Les méthodes d'analyse structurelle proposées présentent l'avantage d'être adaptables à différents types de contrats et de ne nécessiter qu'un faible nombre de données annotées. A l'issue de l'analyse structurelle, nous proposons une étude préliminaire pour l'extraction des anomalies structurelles et celle des anomalies sémantiques en nous basant sur le contenu logique des documents et en exploitant des méthodes originales de catégorisation de textes basées sur le plongement. Les différentes étapes de ce processus font l'objet d'expérimentations détaillées sur des bases de contrats réels.

1 Introduction

Dans le domaine juridique, où la complexité et les nuances abondent, la révolution numérique a introduit une nouvelle ère où l'intelligence artificielle joue un rôle croissant dans l'analyse des textes juridiques. Notre recherche se concentre sur un aspect crucial mais encore peu exploré : la détection d'anomalies dans les documents contractuels. Les contrats, en tant que fondements des accords légaux, nécessitent une précision rigoureuse ; toute irrégularité, même mineure, peut avoir des conséquences significatives. En effet, des anomalies telles qu'un article manquant peuvent invalider un contrat ou nuire à la confiance entre les parties. Avec l'augmentation constante du volume des contrats, il devient impraticable de dépendre exclusivement de l'analyse humaine pour garantir leur exactitude et leur cohérence. Bien que la détection d'anomalies ait été étudiée dans divers domaines, son application aux contrats juridiques reste insuffisamment développée. Nous abordons deux types d'anomalies courantes : les anomalies structurelles, liées à la disposition des clauses, et les anomalies sémantiques, liées au contenu des clauses. Notre étude propose une approche méthodique pour l'analyse

structurelle des documents contractuels, essentielle pour préparer le terrain à une analyse sémantique approfondie. Notre objectif est de détecter les anomalies, afin de réduire la charge de travail des experts juridiques. Ainsi, nous espérons contribuer à l'amélioration de l'exactitude et de l'efficacité de l'analyse contractuelle dans le domaine juridique.

2 État de l'art

Pour extraire de l'information textuelle à partir de documents scannés, il est nécessaire de passer par des phases de segmentation physique identifiant des zones de pixels isolant ainsi le contenu d'intérêt. Dans le domaine de segmentation de documents scannés, on peut par exemple citer Forczmański et al. (2020) ou encore Josi et al. (2022), qui adresse spécifiquement la segmentation physique de documents juridiques allemands à partir d'un ensemble bien défini de caractéristiques de mise en page. Cette approche est circonscrite à des contrats ayant des mises en pages similaires et n'aborde pas les structures hétérogènes. Dans ce qui suit, nous introduisons un modèle supervisé ne nécessitant que peu d'annotations mais avec des performances de détection élevées. Cette segmentation physique permet ensuite d'obtenir une segmentation logique, notamment les titres et clauses. Dans la littérature, on distingue généralement deux grandes familles d'approches pour traiter cette segmentation : celles basées sur des règles et celles basées sur l'apprentissage automatique. Dans la première catégorie, on peut citer les travaux de Josi et al. (2022) et dans la seconde ceux de Chalkidis et al. (2019). Suite à l'obtention de la segmentation logique et l'extraction des titres et clauses, ceux-ci sont soumis à une analyse sémantique qui permettra d'identifier des anomalies. Cette analyse sémantique se fait par Plongements de Texte (PT) Mikolov et al. (2013) qui transforme le texte en vecteurs de taille fixe qui encapsulent leur sémantique. Ces représentations sont obtenues en pré-entraînant des modèles sur des corpus larges afin d'être suffisamment adaptables à des tâches spécifiques. L'utilisation de modèles de langage dynamiques et complexes ne s'avère cependant pas toujours nécessaire, car coûteuse. De ce fait, nous nous intéressons plus particulièrement à l'adaptation des représentations sémantiques obtenues à partir de modèles de langage statiques, plus spécifiquement Document to Vector (Doc2vec) de Mikolov et al. (2013) qui est simple et efficace.

3 Approche générale

Notre approche est donc la suivante : chaque document numérisé du corpus est composé d'une série d'images. Ces images subissent une segmentation physique en blocs de texte. Chaque bloc est ensuite soumis à une segmentation logique pour identifier les titres et les clauses.

A partir des clauses catégorisées par leur titre et que de leurs plongements, on procède à des catégorisations d'anomalies titre-contenu. On définit ainsi des **anomalies structurelles** se rapportent à l'organisation des éléments contractuels, détectables par analyse des séquences de clauses et éléments structuraux et les **anomalies sémantiques** qui concernent le sens et le contenu des clauses contractuelles. Ces anomalies sémantiques sont ensuite séparées en Minoritaires-isolés d'une part et en anomalies de Proximité Inter-groupes qui indiquent des écarts de contenu ou de proximité sémantique entre les catégories de clauses.

4 Détection des anomalies

Anomalies de Structure	Atypicités de Structure
Anomalies de numérotation	1. Numéros d'articles erronés (ex. 3 avant 2) 2. Numéros d'articles manquants 3. Numéros d'article en double 4. Le contrat inclut un autre contrat
Anomalies d'uniformité du style des clauses	5. Le contrat contient des titres d'articles avec leurs objets et d'autres non.
Atypicités de structure	6. Le contrat est composé d'une série d'articles unique ou rare

TAB. 1 – Les différents types d'anomalies structurelles prises en charge.

4.1 Données d'étude et méthodes

Notre approche de détection des anomalies repose sur une combinaison de techniques automatisées et de vérification manuelle par des experts. Cette stratégie permet de valider les anomalies structurelles et sémantiques identifiées, équilibrant l'efficacité de l'automatisation avec la rigueur de l'expertise humaine.

Le corpus analysé comprend 9996 clauses appartenant à 910 documents et réparties en 15 catégories. Pour les anomalies structurelles, nous utilisons l'extraction automatique des clauses, suivie d'une analyse des séquences de clauses pour identifier les contrats présentant des anomalies. Ces anomalies sont ensuite validées manuellement.

Pour identifier les anomalies **Minoritaires-isolées**, nous procédons comme suit : après réduction dimensionnelle et clustering des plongements de texte des clauses, nous identifions des clusters dans chaque catégorie de clause c_i. Un cluster K_j^i est considéré comme une anomalie Minoritaire-isolée si son effectif est significativement inférieur à celui du plus grand cluster dans c_i, selon une mesure spécifique δ :

$$\delta(K_j, c_i) = \frac{|K_{max}^{c_i}| - |K_j^i|}{|K_{max}^{c_i}|} \tag{1}$$

où $|K_{max}^{c_i}|$ est l'effectif du plus grand cluster dans c_i. Une anomalie minoritaire-isolée est identifiée si $\delta(K_j, c_i) > \gamma$ avec γ un paramètre de contrôle.

En ce qui concerne les anomalies de Proximité Inter-groupes, nous calculons les centroïdes (médians) pour chaque catégorie de clauses. Une clause est considérée comme une anomalie de proximité inter-groupes si elle est plus proche d'un centroïde d'une autre catégorie que de celui de sa propre catégorie.

Pour les anomalies **Minoritaires-isolées**, après avoir sélectionné des classes de clauses avec un nombre de données suffisant (ayant un minimum de 20 clauses), nous appliquons la réduction dimensionnelle via t-SNE et utilisons l'indice de Calinsky-Harabatz pour détecter les catégories de clauses avec multiples clusters. Ces clusters sont ensuite identifiés avec l'algorithme HDBSCAN. Les anomalies sont ensuite identifiés en se basant sur l'équation (1).

Pour les **anomalies de Proximité Inter-groupes**, nous projetons les plongements de texte via t-SNE et mesurons la proximité selon la méthode expliquée.

5 Segmentation Physique et Logique des Contrats

Nous abordons la segmentation des contrats en utilisant deux ensembles de données distincts : la base de données Gold, composée de 1059 images de pages de contrats de travail privés français annotées manuellement, et la base de données publique française des accord d'entreprise (ACCO) contenant plus de 150 000 contrats au format `docx`. Ces bases servent à la segmentation physique et logique, essentielles pour identifier les structures des contrats et pour la préparation des données, notamment avec l'utilisation de la bibliothèque PdfMiner et des critères de filtrage spécifiques.

Données Nous manipulons deux ensembles de données distincts en matière de segmentation de documents. Le premier, dénommé base de données Gold, est constitué de 1059 images de pages de contrats de travail français que nous avons annotés et vérifiés manuellement. Cette base de données sert à évaluer la capacité de généralisation de nos modèles en segmentation physique et logique. La seconde base de données publiques des contrats publics Accords d'entreprise (ACCO), contient plus de 150 000 contrats au format `docx`. Elle est utilisée spécifiquement en segmentation physique, pour pré-entraîner le modèle à identifier les structures des contrats. Pour l'étude courante, nous nous restreignons à 11 798 contrats qui représentent 75 348 pages de documents. présentant une richesse de structures et mise en page. Pour la préparation des données ACCO, les contrats sont d'abord convertis en format PDF puis en images, utilisant la bibliothèque PdfMiner pour obtenir simultanément des images des pages et des boîtes englobantes associées aux deux catégories : "Bloc de texte" et "Image". Un ensemble de critères de filtrage est appliqué pour écarter les pages inadéquatement segmentées, tels que l'analyse des intersections de boîtes englobantes et les variations d'interligne. De plus, nous enrichissons notre modèle en incorporant artificiellement des manuscrits à partir des bases de données utilisés par Marti et Bunke (1999), améliorant ainsi la variabilité et la complexité des données traitées.

Méthodes Dans la segmentation physique, nous identifions 6 classes de structures physiques dans les contrats : les images d'en-tête, les textes d'en-tête, les textes du corps, les images de pieds de page, les textes de pieds de page et les manuscrits. Cela est assuré par le modèle populaire de détection supervisé, le Faster R-CNN, avec l'architecture ResNet-50-FPN Wu et al. (2019). Ce modèle est pré-entraîné sur la base de données ACCO et affiné sur la base de données Gold, visant à exposer le modèle à une large gamme d'agencements structuraux.

En ce qui concerne la segmentation logique, nous nous concentrons sur la détection des blocs de texte "titre de clause", utilisant des approches basées sur des règles avec des expressions régulières pour des formats de titres standardisés et des approches supervisées pour des formats variables. Nous utilisons des classifieurs numériques entraînés avec des embeddings de texte tels que Doc2vec et CamemBERT.

6 Résultats

6.1 Segmentation des contrats

Segmentation physique Nos expériences sur la base de données Gold révèlent des performances élevées en segmentation physique. Avec seulement 5% des données d'entraînement, notre modèle a atteint un AP (5-95) de 66.76%. Sans pré-entraînement sur ACCO, l'AP (5-95) est légèrement inférieur, à 62%. L'AP50, un indicateur de détection de haut niveau, reste stable à environ 93%, même avec seulement 5% des données d'entrainement. L'AP75 montre une augmentation progressive avec plus de données d'entraînement, débutant à 77% pour 5% des données et augmentant avec l'ajout de données. Ces résultats indiquent que le pré-entraînement sur ACCO améliore significativement la performance de détection, même avec des ensembles de données d'entraînement réduits.

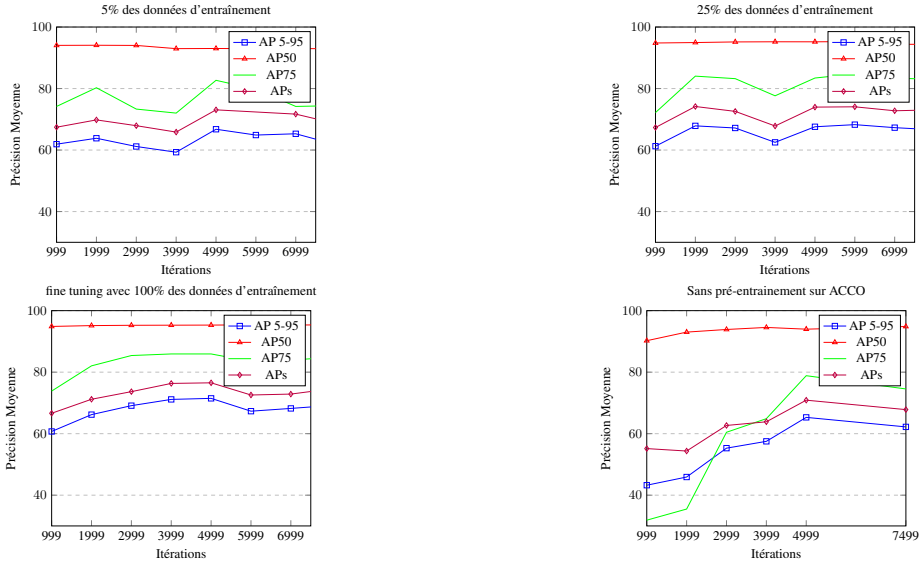

FIG. 1 – Évolution des performances de détection lors du fine-tuning sur 5%, 25% et 100% des données d'entraînement et également sans aucun pré-entrainement sur ACCO

Segmentation logique Nos résultats sur la base de données Gold, illustrés dans la figure 2, montrent l'efficacité des différentes approches de segmentation logique. Le classifieur RegEx, utilisant des règles prédéfinies, a démontré une performance élevée avec une AUC de 0.86 et une précision de 0.98, profitant de la structure standardisée des titres de clauses. Néanmoins, cette méthode présente des limites en termes de flexibilité, en particulier sur des contrats à structure variable. En comparaison, les classifieurs supervisés, particulièrement ceux utilisant des plongements CamemBERT, ont offert un équilibre optimal entre précision et rappel. Ces résultats soulignent l'importance de choisir une méthode de segmentation logique adaptée à la

nature du contrat traité, avec une préférence pour les approches supervisées dans des contextes plus variables.

Modèle	AUC	Préc.	Rappel	F1	CA
RegEx	**0.865**	**0.980**	0.870	**0.921**	**0.965**
SVM	0.628	0.827	0.507	0.583	0.507
RL	0.841	0.889	0.889	0.852	0.889
kNN	0.835	0.903	**0.907**	0.905	0.907
NB	0.690	0.957	0.690	0.866	0.742

Plongements Doc2vec

Modèle	AUC	Préc.	Rappel	F1	CA
Regex	0.865w	0.980	0.870	0.921	0.965
SVM	0.922	0.894	0.888	0.896	0.888
RL	**0.997**	0.985	0.985	0.985	**0.985**
kNN	0.984	**0.986**	**0.986**	**0.986**	0.936
NB	0.985	0.957	0.944	0.947	0.944

Plongements Camembert-Base

FIG. 2 – Performances de classification des titres de clauses

6.2 Les anomalies

Les anomalies structurelles Nous avons identifié 40 contrats présentant des anomalies structurelles potentielles, et 24 atypicités structurelles. Parmi les 40 anomalies structurelles identifiées, 13 le sont vraiment après vérification (32%). En résumé, les contrats anormaux constituent 7% du volume des contrats.

Les anomalies de Proximité Inter-groupes Nous constatons que la projection t-SNE dévoile une disposition spatiale intéressante des clauses. Globalement, les clusters de clauses se distinguent clairement, montrant une bonne cohérence sémantique au sein de chaque catégorie. La présence d'articles atypiques dispersés à l'écart de leur groupe principal, tout en étant plus proches d'autres groupes, indique des anomalies potentielles. Le tableau 2 résume l'ensemble des cas d'anomalies de Proximités Inter-groupes détectées et celles validées.

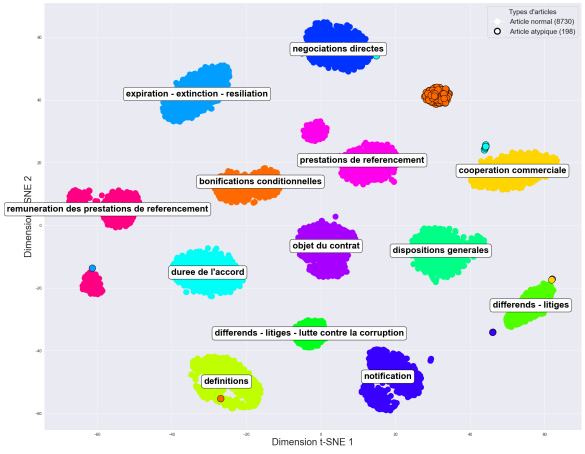

FIG. 3 – Visualisation des anomalies de Proximité Inter-groupes à travers une Projection t-SNE des plongements des textes de clauses contractuelles

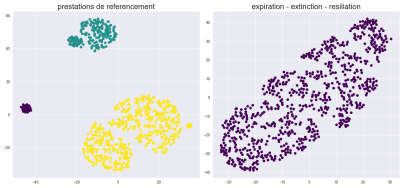

FIG. 4 – Visualisation des clusters de clauses dans le cas d'existence d'anomalies Minoritaires-isolées (à gauche) ou sans anomalie (à droite) pour une catégorie de clauses sélectionnée

Les anomalies Minoritaires-isolées Sur la Figure 4, nous avons montré deux cas-type de configuration de détection qui se manifestent sur les clauses sélectionnées suivantes : "prestations de référencement" et "expiration - extinction - résiliation". Le tableau 3 résume l'ensemble des cas d'anomalies Minoritaires-isolées détectées et celles validées.

6.3 Discussion

Notre étude a démontré une réduction significative du besoin d'analyse manuelle dans la détection d'anomalies structurelles des contrats, grâce à notre approche combinée d'extraction et d'analyse des clauses. Cette efficacité souligne l'intérêt de notre méthode pour traiter les incohérences structurelles, tout en allégeant considérablement le processus d'examen manuel. Dans le domaine des anomalies sémantiques, notre méthode a identifié avec succès des divergences notables au sein des catégories de clauses, mettant en évidence des anomalies de Proximité Inter-groupes et Minoritaires-isolées. L'inspection manuelle des anomalies détectées a confirmé leur pertinence, révélant des nuances sémantiques significatives. Cependant, l'utilisation de plongements CamemBERT pour la détection d'anomalies sémantiques a rencontré des limites dues à la sparsité des projections, suggérant la nécessité d'explorer d'autres techniques de plongement ou d'améliorer l'approche actuelle pour de meilleures performances. Nos résultats valident l'efficacité de la méthode proposée, tout en soulignant l'importance de choisir les bons outils et techniques pour l'analyse sémantique des contrats.

Catégorie de clauses	**Nb (validées)**
Bonifications conditionnelles	179 (179)
Durée de l'accord	13 (13)
Notification	2 (2)
Coopération commerciale	1 (1)
Expiration - Extinction - Résiliation	1 (1)
Objet du contrat	1 (1)
Rémunération des prestations de référencement	1 (1)

TAB. 2 – Occurrence des anomalies de proximité Inter-groupes

Catégorie de clauses	**Nb (validées)**
Définitions	22 (22)
Durée de l'accord	33 (33)
Durée de l'accord	14 (14)
Négociations directes	37 (37)
Notification	16 (16)
Objet du contrat	25 (25)
Prestations de référencement	31 (31)
Rémunération des prestations de référencement	33 (33)

TAB. 3 – Nombre de points dans les clusters Minoritaires-isolées. ($\gamma = 0.1$)

Conclusion

Dans notre étude, nous avons adopté une segmentation structurelle en deux phases avec pré-entraînement et affinage, ce qui a réduit la dépendance aux annotations manuelles et amélioré la précision de la détection des éléments structuraux. L'importance cruciale de la segmentation physique et logique dans l'identification des anomalies a été souligné. Cependant, la caractérisation et la compréhension des anomalies sémantiques demeurent des défis à surmonter. Dans nos travaux futurs, nous envisageons d'intégrer des méthodes d'extraction de sujets, telles que la Maximisation de Traits (MT) selon Anonymized (2017), avec des techniques avancées de parsing des textes et de reconnaissance d'entités nommées. Cette combinaison vise

à détecter les indicateurs clés de variabilité dans les clauses et à identifier leurs changements dans des contextes atypiques. La mise en œuvre de ces méthodes promet de réduire substantiellement la charge de travail manuelle des experts juridiques et d'affiner notre compréhension des anomalies dans les contrats.

Références

Anonymized (2017). Anonymized. *Advances in Anonymus Publications : Volume -$\pi/4$ 6*, $\sqrt{2}$.

Chalkidis, I., M. Fergadiotis, P. Malakasiotis, et I. Androutsopoulos (2019). Neural contract element extraction revisited. In *Workshop on Document Intelligence at NeurIPS 2019*.

Forczmański, P., A. Smoliński, A. Nowosielski, et K. Małecki (2020). Segmentation of scanned documents using deep-learning approach. In *Progress in Computer Recognition Systems 11*, pp. 141–152. Springer.

Josi, F., C. Wartena, et U. Heid (2022). Preparing legal documents for nlp analysis : Improving the classification of text elements by using page features. In *Computer Science & Information Technology (CS & IT)*, pp. 17–29. AIRCC Publishing Corporation.

Marti, U.-V. et H. Bunke (1999). A full english sentence database for off-line handwriting recognition. In *Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR'99 (Cat. No. PR00318)*, pp. 705–708. IEEE.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, et J. Dean (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems 26*.

Wu, Y., A. Kirillov, F. Massa, W.-Y. Lo, et R. Girshick (2019). Detectron2. `https://github.com/facebookresearch/detectron2`.

Summary

This study aims to identify different types of anomalies in image corpora of legal contracts of homogeneous structure, such as sets of contracts from the same source. To achieve this, we rely on a combination of structural and semantic analysis methods. The structural analysis methods proposed have the advantage of being adaptable to different types of contracts, and of requiring only a small amount of annotated data. Following the structural analysis, we propose a preliminary study for the extraction of structural anomalies and semantic anomalies, based on the logical content of the documents and exploiting original text categorization methods based on folding. The various stages of this process are the subject of detailed experiments on real contract databases.