Une approche hybride pour la segmentation automatique de documents juridiques

Filipo Studzinski Perotto, Fadila Taleb, Eric Trupin, Youssouf Saidali, Maryvonne Holzem, Jacques Labiche, Laurent Vercouter

Une approche hybride pour la segmentation automatique de documents juridiques

Filipo Studzinski Perotto[1] Fadila Taleb[2] Eric Trupin[1] Youssouf Saidali[1]
Maryvonne Holzem[2] Jacques Labiche[2] Laurent Vercouter[1]
(1) Normandie Université, UNIROUEN, UNIHAVRE, INSA Rouen, LITIS, 76000 Rouen, France
(2) Normandie Université, UNIROUEN, DYLIS, 76000 Rouen, France
filipo.perotto@litislab.fr

RÉSUMÉ

Cet article [1] propose une approche hybride pour la segmentation de documents basée sur l'agrégation de différentes solutions. Divers algorithmes de segmentation peuvent être utilisés dans le système, ce qui permet la combinaison de stratégies multiples (spécifiques au domaine, supervisées et non-supervisées). Un ensemble de documents étiquetés, segmentés au préalable et représentatif du domaine ciblé, doit être fourni pour être utilisé comme ensemble d'entraînement pour l'apprentissage des méthodes supervisées, et aussi comme ensemble de test pour l'évaluation de la performance de chaque méthode, ce qui déterminera leur poids lors de la phase d'agrégation. L'approche proposée présente de bonnes performances dans un scénario expérimental issu d'un corpus extrait du domaine juridique.

ABSTRACT

A hybrid approach for automatic text segmentation

This paper proposes a hybrid architecture for segmenting text documents, based on the aggregation of different solutions. Diverse segmentation algorithms can be incorporated into the system, allowing the combination of multiple strategies (domain-specific, supervised and unsupervised). A set of annotated documents is used for training the supervised methods, and for evaluating all the methods. The accuracy of each method determines its weight in the aggregation phase. A corpus extracted from a juridic domain was used for testing. The proposed approach presented good performances in such experimental scenario.

MOTS-CLÉS : segmentation linéaire automatique de texte.

KEYWORDS: automatic linear text segmentation.

1 Introduction

La *segmentation linéaire de texte* consiste à diviser un document en parties sémantiquement cohérentes, en identifiant des segments contigus et distincts en fonction de leurs caractéristiques communes (telle que la cohésion lexicale). L'automatisation de cette tâche est une étape aussi cruciale que problématique et les questionnements qu'elle soulève occupent une place importante au sein du *traitement automatique du langage naturel*. En effet, la segmentation constitue une porte d'entrée

1. Cet article est un résultat du projet *PlaIR 2.018*, cofinancé par l'Union Européenne à travers le Fonds Européen de Développement Régional (FEDER) et par la Région Normandie.

pour aborder d'autres problèmes, tels que la *fouille de textes*, la *synthèse de documents*, la *classification de documents*, la *recherche documentaire*, ainsi que la *visualisation*. La segmentation automatique est aussi l'étape préliminaire pour tout système d'aide à l'interprétation qui aurait pour but de faciliter l'accès à des documents complexes pour des lecteurs novices.

Tel est le cas du langage juridique utilisé dans les décisions de justice, qui en plus de la technicité et de l'arduité de leur vocabulaire, présentent une organisation textuelle particulière. Dans cet article, nous proposons une approche hybride pour la segmentation de documents issus d'un même corpus thématique. Plus spécifiquement, nous proposons un système dans lequel plusieurs algorithmes de segmentation peuvent être combinés automatiquement. Cela permet de faire collaborer différentes stratégies : des heuristiques fournies par un spécialiste du domaine, des mécanismes basés sur un apprentissage supervisé, et des méthodes non-supervisées. Pour pouvoir être intégré au système, un algorithme de segmentation doit être capable, lui étant donné un nouveau document à découper, de proposer un score pour chaque paragraphe en tant que candidat pour marquer le début d'un segment. L'architecture étant adaptative, des nouvelles méthodes peuvent être ajoutées au système à tout moment.

Un ensemble de documents préalablement segmentés (étiquetés avec les frontières correctes où un segment se termine et un autre commence) doit être fourni au système. Ces documents doivent être représentatifs du domaine ciblé. Ils servent comme exemples pour toutes les méthodes d'apprentissage supervisé, mais aussi comme ensemble de test permettant d'évaluer la précision de chaque méthode dans la tâche de segmentation. Cette évaluation de performance peut être interprétée comme une mesure de confiance sur chaque méthode, ce qui permet au système de leur attribuer un poids lors de l'agrégation des différentes solutions présentées dans une solution commune.

Cette approche a été testée sur un corpus extrait d'une base documentaire en ligne spécialisée dans la jurisprudence française en matière de transport. Nous avons segmenté un sous-ensemble de ces documents suivant l'usage rhétorique du domaine. Nous avons également défini un ensemble d'heuristiques, constituant un modèle linguistique qui est mis en œuvre par l'une des méthodes dans le système. L'approche hybride a permis d'identifier les segments corrects au sein des nouveaux documents avec une précision supérieure à celle de chaque méthode prise isolément.

Dans la suite de l'article, la section 2 réalise un bref aperçu des approches et des méthodes les plus importantes en segmentation automatique du texte. La section 3 présente le corpus de documents et la spécificité de la segmentation en question. La section 4 définie l'approche hybride que nous proposons. La section 5 compare expérimentalement notre architecture contre d'autres algorithmes classiques. Les conclusions et les travaux futurs sont discutés dans la section 6.

2 Travaux apparentés

La stratégie prédominante pour automatiser la segmentation linéaire de texte est l'utilisation des méthodes non-supervisées. Ces méthodes s'appuient sur la quantification de la cohésion lexicale entre différentes parties du document analysé (Koshorek *et al.*, 2018). La cohésion lexicale correspond à la manière dont les mots sont enchaînés dans le flux des phrases afin de créer des unités sémantiques (Morris & Hirst, 1991). Ces méthodes essaient d'identifier la cohésion lexicale dans une zone du texte, et ensuite de partitionner le document en un ensemble de segments thématiquement cohérents (Dadachev *et al.*, 2014). Les zones de texte avec un vocabulaire similaire sont susceptibles de faire

partie d'un même segment, et une variation lexicale peut être l'indicateur d'un changement de sujet.

`TextTiling` (Hearst, 1997) est le premier algorithme représentatif de cette approche. Il compare deux blocs de mots adjacents de longueur fixe en mesurant la répétition des mots entre eux. En déplaçant petit-à-petit les deux blocs tout le long du document, une fonction de similarité peut être dessinée selon la position (i.e. la frontière entre deux segments supposés). Les frontières qui affichent la plus petite similarité entre les blocs de texte qu'elles divisent sont sélectionnées comme candidates potentielles pour diviser le document en segments. Des méthodes plus performantes ont été proposées dans cette même approche grâce à l'utilisation des modèles statistiques plus fins (Choi, 2000; Brants *et al.*, 2002; Eisenstein & Barzilay, 2008; Chen *et al.*, 2009; Sakahara *et al.*, 2014), grâce à l'utilisation des relations sémantiques (ontologies) pour considérer la similarité entre les mots (Bayomi *et al.*, 2015; Ercan & Cicekli, 2016), ou des différentes façons d'extraire les caractéristiques du document (Utiyama & Isahara, 2001; Malioutov & Barzilay, 2006; Misra *et al.*, 2009; Dadachev *et al.*, 2014), ou par l'utilisation d'une représentation plus élaborée du lexique, s'éloignant du « bag-of-words » pour privilégier les « word embeddings » (Riedl & Biemann, 2012; Glavaš *et al.*, 2016).

Cependant, lorsqu'un ensemble représentatif de documents segmentés est disponible, l'utilisation de techniques d'apprentissage supervisé devient une opportunité intéressante. Pourtant peu de travaux suivant cette approche ont été publiés. (Beeferman *et al.*, 1999) utilise un algorithme de classification afin d'apprendre au système à détecter si une phrase donnée indique potentiellement le début ou la fin d'un segment. (Koshorek *et al.*, 2018) propose un modèle neuronal hiérarchique pour classifier l'appartenance d'une phrase à un segment spécifique. L'avantage d'utiliser l'apprentissage supervisé est que le critère de segmentation, même s'il est très complexe, est défini par extension. Si le jeu de données d'apprentissage est représentatif, il peut être « appris » au travers d'exemples, sans qu'une définition explicite soit nécessaire (Passonneau & Litman, 1997; Beeferman *et al.*, 1999).

3 Verrou scientifique

Dans le cadre d'un projet pluridisciplinaire [2], des chercheurs informaticiens et linguistes se sont posées la question de concevoir un système d'aide à l'interprétation d'un fond jurisprudentiel [3]. En amont de l'implémentation d'un tel système, un travail linguistique a été mené sur un corpus de plus de 300 décisions de justice dans le but de comprendre leur structure, le mécanisme argumentatif mis en œuvre, et surtout de mettre au jour des scénarios modaux [4] susceptibles de déclencher des parcours interprétatifs pouvant aider à la lecture de ces décisions (Taleb & Holzem, 2018). Dans cette perspective, une segmentation semi-manuelle a été effectué sur le corpus. Le découpage a servi à une première analyse textométrique différentielle [5].

Nous avons travaillé sur un corpus extrait d'une base documentaire en ligne de l'Institut du Droit

2. PlaIR (Plateforme d'Indexation Régionale - Normandie)

3. Parler d'aide à l'interprétation suppose d'adapter celle-ci aux pratiques professionnelles concernées et donc aux textes. L'enjeu est de pouvoir accéder finement à l'argumentaire juridique au sein de chacun des segments et comprendre leurs enchaînements intra et inter segmentales.

4. Scénario modal entendu comme l'expression du point de vue adopté par le magistrat en fonction des faits (modalité aléthique), puis de leur appréciation (modalité appréciative), autorisant un jugement de valeur de nature légal sur les actes en question (modalité axiologique), et le verdict (modalité déontique).

5. Repérage de constructions lexicales recourantes, qui marquent des moments clés du jugement, souvent corrélés à une transformation modale.

International des Transports (IDIT) [6]. Leur bibliothèque numérique, spécialisée dans la jurisprudence en matière de transport, dispose d'environ 40000 documents, dont 3000 en accès libre. Nous avons analysé un sous-ensemble de 300 arrêts [7], écrits en français, produits à différents moments, par différents juges, dans diverses *cours d'appel* [8] françaises. Chacun de ces documents a été manuellement partitionné en 4 segments, suivant l'usage en pratique dans le domaine. Ce corpus annoté constitue un ensemble de données d'entraînement permettant l'application de techniques d'apprentissage supervisé.

En raison de la spécificité du domaine, la structure des segments dans les documents du corpus est régulière. Chaque document présente 4 segments distincts et contigus, apparaissant toujours dans le même ordre : (1) le *header*, où sont déclarées des informations telles que le nom du tribunal, le juge, la ville, la date et le nom des parties en litige ; (2) les *faits*, où le contexte du désaccord est rappelé sous la forme de récit, ainsi que les prétentions des parties appelantes et intimées, et le verdict prononcé par la juridiction d'instance inférieure (le tribunal) ; (3) les *motifs*, où le juge expose les arguments qui justifient sa décision ; et (4) la *conclusion*, où le juge énonce son verdict final. Ces segments constituent des unités de plusieurs paragraphes à l'intérieur du document. Les frontières entre les segments se trouvent toujours dans le passage d'un paragraphe au suivant [9]. Chaque paragraphe appartient nécessairement à un seul segment.

Formellement, le corpus est constitué d'un ensemble de n documents $D = \{d_1, d_2, ..., d_i, ..., d_n\}$. Chaque document d_i est une séquence de m_i paragraphes $\{p_1, p_2, ..., p_x, ..., p_{m_i}\}$. La tâche consiste à rechercher les limites $\{b_1, b_2, b_3\}$ qui séparent correctement les 4 segments cherchés $\{s_1, s_2, s_3, s_4\}$ (*header*, *faits*, *motifs*, *conclusion*), où b_j correspond à l'index du paragraphe qui marque le début du segment $j + 1$. Chaque segment doit être constitué d'au moins un paragraphe. Le premier segment commence toujours au premier paragraphe, et le dernier segment se termine toujours par le dernier paragraphe. On observe donc la contrainte $1 < b_1 < b_2 < b_3 \leq m_i$. Le défi est celui de pouvoir segmenter des nouveaux documents, jamais vus auparavant, uniquement à partir de l'ensemble de documents d'entraînement.

4 Méthode Proposée

La première approche que nous avons adoptée a été l'analyse textométrique du sous-ensemble de 300 documents étiquetés avec les frontières réelles entre les segments, pour en déduire un système d'expressions régulières capables de capturer la majorité des cas (table 1). Même si ces documents juridiques issues des la cour d'appel française ne sont pas des formulaires mais des textes libres, il est fréquent d'observer l'usage des marqueurs explicites pour indiquer le début de chaque segment. Ces règles grammaticales permettent de retrouver correctement environs 90% des frontières entre les segments dans le jeu de documents en question. C'est une performance assez correcte, mais qui signifie plutôt que dans la majorité des documents il y a une intention explicite de bien démarquer le passage entre les segments. En plus, on peut s'attendre à ce que le taux de précision baisse dès lors que les règles seront appliquées à d'autres documents, en dehors du sous-ensemble qui leur a donné

6. `www.idit.fr`

7. Le contenu des documents est disponible en format texte brut, généralement extrait des fichiers PDF. Après prétraitement, chaque document est représenté comme une collection ordonnée de paragraphes.

8. Cette juridiction de second degré est sollicitée par l'une des parties d'un litige ayant fait appel d'un précédent jugement rendue par une juridiction de premier degré.

9. Un changement de segment est contraint de coïncider avec un changement de paragraphe.

Segment	Expression régulière (insensible à la casse)	Confiance
Faits	^(\s*)(*exposé* (*du litige* | *des faits*) | *faits*)([\s :]*)$	1.0
Faits	^(\s*)(*vu que* | *par actes* | *la cour* (.*) *attendu que*)	0.9
Motifs	^(\s*)(*motifs*)([\s :]*)$	1.0
Motifs	^(\s*)(*sur ce* | (*ce*) ? *sur quoi* | (*ceci* | *cela*) *étant* | *étant exposé*)	0.9
Conclusion	^(\s*)(*décision*)([\s :]*)$	1.0
Conclusion	^(\s*)(*par ces motifs*)	0.9

TABLE 1 – Expressions régulières construites après analyse textométrique pour identifier les expressions-repères indiquant le début de chaque segment dans les documents de décision de cours d'appel françaises.

origine. Finalement, la faiblesse de cette méthode vient des 10% de segmentations incorrectes. Ces sont les cas pour lesquels aucune règle ne peut être appliquée, et donc aucune suggestion ne peut être retournée, ou les cas où plusieurs correspondances sont trouvées.

Le défi posé par ce problème est donc de développer une méthode plus précise que celle basée sur les heuristiques apportées par un expert humain. Les méthodes non-supervisées, comme on peut le constater dans les résultats (section 5), ont une précision assez faible (inférieure à 10%). Ce n'est pas surprenant, vu qu'il s'agit de documents longs, segmentés d'une manière spécifique au domaine, assez éloignée du principe qui régit la segmentation thématique, qui consiste à trouver des segments homogènes d'un point de vue sémantique. Le corpus en question demande plutôt une tâche de structuration, chaque document étant composé d'un nombre identique de segments, qui suivent une structure canonique bien précise. Cela laisse supposer que l'on retrouve un vocabulaire commun tout au long du document dans ce type de décision juridique, alors que quelques mots structurants (repérés par l'approche textométrique) viennent marquer les changement de segment, de sorte que les approches basées sur la cohésion lexicale ne peuvent pas être très performantes.

Finalement, les méthodes basées sur l'apprentissage supervisé s'approchent de la précision rendue par la méthode d'expressions régulières, restant pourtant moins performantes. Dans cet article, l'apprentissage supervisée a été réalisé de la façon suivante : un classificateur du type *Naive Bayes* a été entraînée pour identifier les paragraphes initiaux des différents segments. Les exemples d'entraînement étant constitués de la liste de *tokens* (mots après filtrage de *stop-words*) contenus dans le paragraphe, et l'étiquette du segment qu'il débute ($1 =$ *entête*, $2 =$ *faits*, $3 =$ *motifs*, $4 =$ *conclusions*, ou 0 s'il n'est au début d'aucun segment). Cette méthode a pu atteindre un taux de précision d'environs 80%.

La solution que nous proposons consiste à faire collaborer plusieurs méthodes de segmentation dans un système hybride. Lorsque nous disposons d'un ensemble représentatif d'exemples indiquant comment les documents à l'intérieur du corpus doivent être segmentés, il est possible d'estimer la précision de chaque méthode. Cette valeur (la précision de chaque méthode évaluée contre le jeu de documents préalablement segmentés) indique la confiance que le système a en la méthode. Cette confiance servira ensuite à pondérer ses réponses lors d'une phase ultérieure d'agrégation de la solution. Il s'agit d'un problème d'*agrégation de préférences* (Conitzer, 2006; Sen & Yang, 1994). Il faut combiner de manière adéquate les différents résultats provenant des différentes méthodes. Dans le cas de la segmentation, une moyenne pondérée de différentes positions proposées ne semblerait pas avoir beaucoup de sens. Nous proposons de suivre la majorité pondérée par deux facteurs : la confiance qui le système a dans la méthode, et la confiance que la méthode a en ses réponses.

Formellement, étant donné un ensemble de documents $D = \{d_1, d_2, ..., d_i, ..., d_n\}$, où les frontières

$B_i = \{b_1, b_2, b_3\}$ entre les segments sont connues pour chaque document d_i, et un ensemble de z méthodes de segmentation $A = \{a_1, a_2, ..., a_k, ..., a_z\}$, nous calculons la qualité w_k de la méthode a_k dans la segmentation de l'ensemble de documents. Les métriques classiques de précision, telles que celles proposées par (Beeferman *et al.*, 1999) et (Hearst, 1997), considèrent la distance de chaque phrase par rapport à son segment correct. Nous avons choisi un score plus simple et plus strict. La précision w_k d'une méthode a_k correspond à la proportion de frontières correctement trouvées, et est calculée comme suit :

$$w_k = \frac{\sum_{i=1}^{n} \sum_{j=1}^{3} c_{k,i,j}}{3n} \tag{1}$$

où n est le nombre de documents dans D, 3 est le nombre de frontières à trouver dans chaque document, $c_{k,i,j} = 1$ si la frontière b_j dans le document d_i est correctement proposée par la méthode a_k, et 0 sinon. Cette formule peut être utilisée telle quelle pour évaluer les méthodes non-supervisées, ou basées sur des heuristiques. Pour les méthodes supervisées, considérant que l'ensemble de test est utilisé comme ensemble d'entraînement, on utilise une validation croisée du type 5-*fold* [10].

Après avoir analysé un document donné d_i, une méthode a_k doit indiquer une valeur de confiance $v_{k,i,j,x}$ indiquant sa croyance que la frontière b_j se trouve au début de chaque paragraphe p_x du document. Une confiance combinée peut être en suite calculée à partir des valeurs de confiance de chaque méthode individuelle, comme suit :

$$\forall i \; \forall j \; \forall x \quad : \quad s_{i,j,x} = \sum_{k=1}^{z} {w_k}^2 \; v_{k,i,j,x} \tag{2}$$

où $s_{i,j,x}$ est le score du paragraphe p_x en tant que candidat pour la frontière b_j du document d_i, w_k est la confiance accordée à la méthode a_k, et $v_{k,i,j,x}$ est la confiance de la méthode a_k sur le fait que la frontière b_j du document d_i se trouve au paragraphe p_x. Dans la formule, on utilise le carré de la confiance pour favoriser les réponses données par les meilleures méthodes. La segmentation finale d'un document donné d_i est choisie en sélectionnant la combinaison de frontières qui obtient le plus grand score moyen, en respectant la contrainte d'ordre entre les segments, i.e. celle qui optimise :

$$\max \frac{1}{3} \sum_{j=1}^{3} s_{i,j,x} \quad \text{sujet à} \quad b_1 < b_2 < b_3 \tag{3}$$

5 Résultats

Nous avons essayé notre approche hybride en utilisant 3 différentes méthodes :

(1) La première méthode implémente l'ensemble d'expressions régulières issues d'une analyse textométrique sur l'ensemble de documents d'entraînement (table 1). Ces règles sont conçues pour rechercher l'occurrence de certaines expressions-repères qui indiquent le début de chaque segment. Chacune de ces règles est associée à un niveau de fiabilité (aussi fourni par l'expert du domaine). Lorsque l'expression est trouvée dans le texte, cette fiabilité de la prédiction est retournée, associée

10. L'ensemble de documents d'entraînement est reparti en 5 différents échantillons. L'apprentissage en utilise 4 et la validation se fait sur l'échantillon restant. La démarche est répétée pour chaque échantillon.

au paragraphe en question, pour la frontière en question. Si aucune règle ne s'applique au paragraphe, la valeur 0 est retournée.

(2) La deuxième méthode est une version adaptée de l'algorithme non-supervisé classique `TextTiling` (Hearst, 1997). Comme les limites de segment sont contraintes de coïncider avec les limites de paragraphe, les comparaisons de similarité ne sont faites que sur ces positions. `TextTiling` retourne, pour chaque paragraphe, une valeur qui correspond à la probabilité que ce paragraphe soit le début d'un segment, sans pouvoir préciser de quel segment s'agit-il.

(3) La dernière méthode utilise un classificateur bayésien naïf, entraîné pour prévoir la probabilité qu'un paragraphe donné soit le paragraphe initial d'un segment cherché.

Les deux premières méthodes n'ont pas besoin d'être entraînés sur les documents préalablement étiquetés : la première est constituée d'heuristiques, et la deuxième est non-supervisée. Leur évaluation peut être faite directement en regardant leur précision dans la segmentation du jeu de documents d'entraînement, ici servant à tester leur performance. La dernière méthode utilise l'apprentissage supervisé. Dans ce cas, même si l'entraînement se fait avec l'intégralité de la base de documents étiquetés, sa précision doit être évaluée par validation croisée.

Dans la table 2, nous pouvons comparer la précision de chacune des méthodes séparément, puis la précision de notre système hybride, qui met en place une collaboration entre eux. Nous pouvons constater que l'approche hybride conduit à une amélioration de la performance, en comparaison avec chacune des autres méthodes isolées.

Stratégie	**Méthode**	**Précision**
heuristique	`RegEx`	0.91
non-supervisée	`TextTilling`	0.08
supervisée	`NaiveBayes`	0.81
combinée	**`Hybride`**	**0.96**

TABLE 2 – Comparaison de performance entre les méthodes expérimentées.

6 Conclusions et Travaux Futurs

Dans cet article, nous proposons une architecture hybride pour la segmentation linéaire de documents texte. Dans cette architecture, chaque méthode peut implémenter un algorithme différent, ce qui permet de combiner la puissance de plusieurs stratégies : spécifiques au domaine, supervisées et non-supervisées. Un ensemble de documents extraits du domaine juridique préalablement segmentés a été utilisé pour l'entraînement des méthodes supervisées et pour l'évaluation de toutes les méthodes. Nous avons pu démontrer que l'agrégation des différentes solutions à l'aide d'une règle de majorité pondérée a permis d'améliorer la précision de la segmentation automatique si comparé à la performance de chaque méthode isolée. Comme chaque méthode cherche à identifier des caractéristiques différentes pour déterminer les changements de segments, même une méthode qui n'est pas très performante en moyenne peut venir contribuer à la solution combinée dans les cas où les méthodes plus performantes échouent. Nous souhaitons poursuivre cette recherche en réalisant une analyse comparative plus approfondie entre notre méthode et d'autres algorithmes de segmentation, ainsi qu'en étendant les tests à d'autres jeux de documents.

Références

BAYOMI M., LEVACHER K., GHORAB M. R. & LAWLESS S. (2015). Ontoseg : A novel approach to text segmentation using ontological similarity. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, p. 1274–1283 : IEEE.

BEEFERMAN D., BERGER A. L. & LAFFERTY J. D. (1999). Statistical models for text segmentation. *Machine Learning*, **34**(1-3), 177–210.

BRANTS T., CHEN F. & TSOCHANTARIDIS I. (2002). Topic-based document segmentation with probabilistic latent semantic analysis. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, CIKM '02, p. 211–218, New York, NY, USA : ACM.

CHEN H., BRANAVAN S. R. K., BARZILAY R. & KARGER D. R. (2009). Global models of document structure using latent permutations. In *Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, p. 371–379, Stroudsburg, PA, USA : Association for Computational Linguistics.

CHOI F. Y. Y. (2000). Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, p. 26–33, Stroudsburg, PA, USA : Association for Computational Linguistics.

CONITZER V. (2006). *Computational Aspects of Preference Aggregation*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA. AAI3232648.

DADACHEV B., BALINSKY A. & BALINSKY H. (2014). On automatic text segmentation. In *Proceedings of the 2014 ACM Symposium on Document Engineering*, DocEng '14, p. 73–80, New York, NY, USA : ACM.

EISENSTEIN J. & BARZILAY R. (2008). Bayesian unsupervised topic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, p. 334–343, Stroudsburg, PA, USA : Association for Computational Linguistics.

ERCAN G. & CICEKLI I. (2016). Topic segmentation using word-level semantic relatedness functions. *J. Inf. Sci.*, **42**(5), 597–608.

GLAVAŠ G., NANNI F. & PONZETTO S. P. (2016). Unsupervised text segmentation using semantic relatedness graphs. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, p. 125–130 : Association for Computational Linguistics.

HEARST M. A. (1997). Texttiling : Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, **23**(1), 33–64.

KOSHOREK O., COHEN A., MOR N., ROTMAN M. & BERANT J. (2018). Text segmentation as a supervised learning task. *CoRR*, **abs/1803.09337**.

MALIOUTOV I. & BARZILAY R. (2006). Minimum cut model for spoken lecture segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, p. 25–32, Stroudsburg, PA, USA : Association for Computational Linguistics.

MISRA H., YVON F., JOSE J. M. & CAPPE O. (2009). Text segmentation via topic modeling : An analytical study. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, p. 1553–1556, New York, NY, USA : ACM.

MORRIS J. & HIRST G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Comput. Linguist.*, **17**(1), 21–48.

PASSONNEAU R. J. & LITMAN D. J. (1997). Discourse segmentation by human and automated means. *Computational Linguistics*, **23**(1), 103–139.

RIEDL M. & BIEMANN C. (2012). Topictiling : A text segmentation algorithm based on lda. In *Proceedings of ACL 2012 Student Research Workshop*, ACL '12, p. 37–42, Stroudsburg, PA, USA : Association for Computational Linguistics.

SAKAHARA M., OKADA S. & NITTA K. (2014). Domain-independent unsupervised text segmentation for data management. In *2014 IEEE International Conference on Data Mining Workshop*, p. 481–487 : IEEE.

SEN P. & YANG J.-B. (1994). Design decision making based upon multiple attribute evaluations and minimal preference information. *Mathl. Comput. Modelling*, **20**(3), 107–124.

TALEB F. & HOLZEM M. (2018). Exploration textométrique d'un corpus de motifs juridiques dans le droit international des transports. In *Proceedings of the* 14^{th} *Int. Conf. on Statistical Analysis of Textual Data (JADT 2018)*.

UTIYAMA M. & ISAHARA H. (2001). A statistical model for domain-independent text segmentation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, p. 499–506, Stroudsburg, PA, USA : Association for Computational Linguistics.