Impact de la structure logique des documents sur les modèles distributionnels : expérimentations sur le corpus TALN

Ludovic Tanguy, Cécile Fabre, Yoann Bard

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Impact de la structure logique des documents sur les modèles distributionnels : expérimentations sur le corpus TALN

Ludovic Tanguy, Cécile Fabre, Yoann Bard
CLLE : Université de Toulouse & CNRS, France
ludovic.tanguy@univ-tlse2.fr, cecile.fabre@univ-tlse2.fr,
yoann.bard@hotmail.fr

RÉSUMÉ

Nous présentons une expérience visant à mesurer en quoi la structure logique d'un document impacte les représentations lexicales dans les modèles de sémantique distributionnelle. En nous basant sur des documents structurés (articles de recherche en TAL) nous comparons des modèles construits sur des corpus obtenus par suppression de certaines parties des textes du corpus : titres de section, résumés, introductions et conclusions. Nous montrons que malgré des différences selon les parties et le lexique pris en compte, ces zones réputées particulièrement informatives du contenu d'un article ont un impact globalement moins significatif que le reste du texte sur la construction du modèle.

ABSTRACT

Impact of document structure on distributional semantics models: a case study on NLP research articles

We present an experiment which aims at measuring the impact of document structure on distributional semantics models. Based on a structured corpus of French research articles in NLP, we have built different models by removing specific parts of the corpus: section headers, abstracts, introductions and conclusions. We show that removing these parts has different effects, depending on the target words and the nature of the removed part. Most importantly, we show that these different parts of a research article that are considered particularly informative of its content have a significantly lesser effect than the rest of the text on the resulting distributional models.

MOTS-CLÉS : structure de document, analyse distributionnelle, corpus spécialisé.

KEYWORDS: Document structure, distributional semantics, specialised corpus.

1 Introduction

Les modèles de sémantique distributionnelle sont généralement construits à partir de très grands corpus rassemblant des textes hétérogènes, afin de tirer bénéfice de la masse des données, dont l'impact sur la qualité du modèle généré a été maintes fois démontré, par exemple par Sahlgren & Lenci (2016). Or, de nombreuses applications nécessitent de construire des représentations sémantiques propres à un domaine de spécialité dans le cadre d'un travail terminologique qui vise en particulier la construction de vocabulaires contrôlés et d'ontologies. C'est le cas par exemple de Bernier-Colborne (2014) pour le domaine de l'environnement, ou de Cohen & Widdows (2009) dans le domaine bio-médical. Dans le cadre de ces travaux à visée applicative où la nature précise des données importe,

Volume 2 : Traitement Automatique des Langues Naturelles, pages 122–135.
HAL : hal-02784760.

ces modèles génériques ne sont pas directement utilisables (El Boukkouri *et al.*, 2019). Or, construire des word embeddings à partir de corpus spécialisés pose des problèmes spécifiques : ces corpus sont généralement de taille plus modeste que les grands corpus qui servent à entraîner les modèles de référence –même s'il faut nuancer cette affirmation pour certains domaines comme la médecine, dans le cas de l'anglais. En outre, les unités de sens sont souvent des termes complexes, qui, du fait de leur spécificité, réduisent encore le volume des contextes exploitables. Enfin, les protocoles d'évaluation sont plus difficiles à établir, sauf à disposer de ressources termino-ontologiques. En revanche, ces corpus spécialisés peuvent posséder des caractéristiques potentiellement intéressantes : le lexique est réduit, moins ambigu, et les documents sont généralement très structurés. C'est cette dernière caractéristique dont nous cherchons à tirer parti dans ce travail.

L'expérience que nous décrivons dans cet article cherche à évaluer la possibilité de prendre en compte la structure des documents dans la construction de modèles sémantiques distributionnels à partir de corpus spécialisés. Ce niveau d'information structurel a été exploité dans différentes tâches, comme le résumé ou l'extraction de connaissances (Hofmann *et al.*, 2009; Teufel & Moens, 2002), en partant du principe que certaines zones textuelles facilitent la détection de contenus plus saillants et donc plus utiles. Dans le même ordre d'idées, notre objectif ici est de déterminer si certaines zones de texte ont un impact particulier sur le modèle sémantique et pourraient donc être privilégiées. Notre hypothèse est en effet que certaines parties des textes comme le résumé ou l'introduction sont particulièrement denses en information et susceptibles de fournir une information distributionnelle de meilleure qualité.

Dans cet objectif, nous avons choisi de privilégier un corpus structuré homogène, dont les parties peuvent être identifiées de façon systématique. Le corpus sur lequel est fondé notre expérience est constitué d'articles scientifiques dans le domaine du traitement automatique des langues. Le caractère standardisé des articles scientifiques a été largement étudié. La spécificité des différentes sections ou étapes qui jalonnent ces textes a permis de dégager des modèles argumentatifs génériques valables à travers la diversité des disciplines (Swales, 1990; Teufel *et al.*, 2009). Dans les disciplines expérimentales, le format de type IMRaD (Introduction, Method, Results and Discussion) s'est imposé (Sollaci & Pereira, 2004). Cette codification facilite l'exploitation de la structure des textes, accessible par les titres de section, qui constituent des indices de présentation externe relativement stables. La spécificité argumentative de chaque section se traduit également sur le plan lexical, ce qui permet de faire l'hypothèse de différences de contribution dans la construction de modèles distributionnels. Ainsi, (Bertin & Atanassova, 2014), s'intéressant au vocabulaire verbal lié aux contextes de citation, font état de différences lexicales importantes d'une section à l'autre. Plus près des objectifs qui sont les nôtres dans cet article, (Badenes-Olmedo *et al.*, 2017) ont étudié la capacité de différents fragments spécifiques d'un article à fournir un équivalent informationnel représentatif de l'ensemble. Leur étude compare l'apport du résumé et celui d'autres zones du texte présentant le contexte, l'approche, ou les résultats de l'article. Ils montrent des différences nettes de représentativité selon les sections considérées, mettant en cause l'utilisation par défaut du résumé comme source privilégiée d'accès au texte.

Nous examinons le rôle spécifique de certaines zones des documents, à savoir l'introduction, le résumé, les titres de section et les conclusions, qui constituent les sections les plus systématiquement représentées dans les articles du corpus TALN, dont le degré de codification est limité. Nous mesurons leur impact en utilisant comme critère le score de variation entre des modèles construits en intégrant ou en supprimant les zones concernées. Nous présentons tout d'abord le dispositif déployé : le corpus constitué pour l'occasion (section 2), les modèles distributionnels construits et la façon de les comparer (section 3). La section 4 présente les analyses quantitatives (mesures globales de la variation entre les modèles) et quantitatives (étude des éléments du lexique plus ou moins impactés).

2 Le corpus TALN

Le corpus TALN utilisé dans cette expérience a été construit à partir des archives qui rassemblent les articles des conférences TALN et RÉCITAL [1] des années 1997 à 2019. Ce corpus a été constitué en plusieurs étapes :

— constitution des archives PDF et récupération des métadonnées des articles de 1997 à 2015, *cf.* (Boudin, 2013) [2] ;
— collecte additionnelle des actes de 2016 à 2019 sur le site des éditions de la conférence ;
— sélection des articles rédigés en français ;
— extraction du contenu textuel en balisant les éléments de la structure logique du document présentée dans la figure 1.

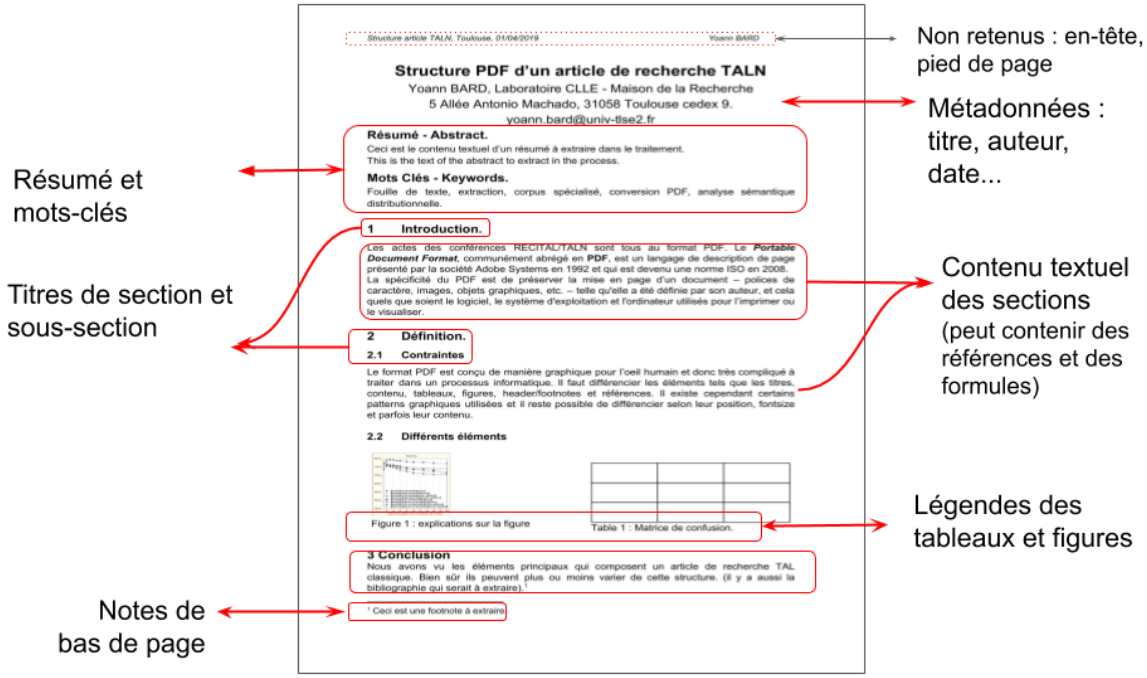

FIGURE 1 – Éléments structurels extraits et annotés dans un article TALN

Cette dernière étape a été réalisée en convertissant les fichiers PDF au format XML grâce à des outils de conversion comme la bibliothèque *pdfminer* de python [3], puis en récupérant le contenu textuel pertinent et en segmentant chaque partie des articles grâce à l'outil ParsCit (Councill *et al.*, 2008).

Un premier travail de restructuration a été effectué sur la sortie de ParsCit pour conserver la structure interne des articles en la représentant dans un format XML *ad hoc*. Nous avons ensuite procédé à un nettoyage automatique pour corriger certains problèmes liés aux caractères spéciaux, traiter les césures et éliminer des éléments résiduels comme les numéros de page et les noms d'auteur figurant dans les en-têtes et les pieds de page, ainsi que différents symboles et idéogrammes. Plusieurs légendes de figures et de tableaux, titres de section et notes de bas de page n'avaient pas été segmentés correctement ou se trouvaient affectés à d'autres éléments du document. La plupart de ces erreurs de segmentation suivant un schéma bien précis, il a été possible de les corriger automatiquement. Les erreurs qui subsistaient après cette phase correspondaient à des ambiguïtés d'interprétation de la

1. Disponibles sur https://www.atala.org/-Conference-TALN-RECITAL
2. Disponibles sur https://github.com/boudinfl/taln-archives
3. https://pypi.org/project/pdfminer/

forme des documents, parmi lesquelles des titres de section identifiés comme des notes de bas de page ou encore des parties de légende sectionnés qui se confondaient avec le corps d'un paragraphe. Nous avons opté pour une approche semi-automatique consistant à développer des règles pour identifier les cas ambigus et à faire ensuite appel à une vérification manuelle des zones correspondantes.

Le marquage obtenu indique donc les méta-données (auteurs, date, type d'article), les informations de la première page (titre, résumés et mots-clés en français et en anglais), les titres des sections et sous-sections, les paragraphes, les légendes de figures et de tableaux, les notes de bas de page ainsi que la bibliographie (mais cette dernière reste non analysée à ce stade). Les sections ont été dotées d'un attribut supplémentaire qui indique leur fonction sur la base de règles ad hoc, pour identifier les introductions et les conclusions. Nous n'avons considéré comme introductions que les sections dont le titre est "Introduction". Pour identifier les conclusions nous avons intégré un ensemble de variantes ("Conclusion", "Perspectives", "Conclusion et Perspectives", etc.). Notons que certains articles ne disposent pas de telles sections, soit parce que les auteurs ont fait un autre choix dans la structuration ou le titrage, soit parce qu'ils relèvent d'un type de communication spécifique (démonstration de logiciel, prise de position, charte, etc.).

Le corpus final est constitué de 1602 articles pour un total de 5,8 millions de mots (tokens). Parmi eux, 1321 articles contiennent un corps de texte. Pour les autres, seuls les métadonnées, résumés et mots-clés ont pu être identifiés. Ce cas de figure s'explique par le codage PDF spécifique de certains fichiers empêchant leur conversion en texte, et par la présence d'articles en anglais dont les métadonnées ont été conservées lorsqu'on disposait d'au moins un contenu exploitable en français (résumé, traduction du titre, mots-clés). L'ensemble du corpus (métadonnées et contenu) a été mis sous format XML en respectant la norme TEI P5. Le corpus TALN est utilisable librement pour les besoins de la recherche et à des fins non commerciales grâce à une licence spécifique accordée par l'ATALA. Il est disponible sur la plateforme de diffusion *Ortolang* [4].

3 Construction et comparaison des modèles distributionnels

Nous présentons ici les opérations qui mènent du corpus TALN aux modèles sémantiques et à leur comparaison.

3.1 Construction des versions de corpus

Nous avons créé pour notre expérience plusieurs corpus dérivés du corpus TALN décrit dans la section précédente. Nous avons commencé par en supprimer les éléments suivants : métadonnées, titre de l'article, auteurs et bibliographie. Nous avons ensuite construit un corpus de référence et 4 sous-corpus correspondant chacun à la suppression d'une zone particulière des documents :

- un corpus de référence comportant l'ensemble du contenu des zones de document suivantes : titre, résumé en français, titres des sections et sous-sections, paragraphes, notes de bas de page et légendes de figure ou de tableau (pour un total de 4 915 365 mots).
- un corpus sans titres de section et sous-section (4 865 324 mots, 99% du corpus de référence).
- un corpus sans les résumés (4 734 393 mots, 96 % du corpus de référence).
- un corpus sans les introductions (4 420 989 mots, 90 % du corpus de référence).
- un corpus sans les conclusions (4 640 218 mots, 94 % du corpus de référence).

La taille et le nombre des éléments supprimés du corpus de référence peuvent varier selon les articles.

4. https://www.ortolang.fr/market/corpora/corpus-taln

La table 1 fournit les caractéristiques des segments pris en compte.

Zone	Nombre	Longueur moy.	Longueur min.	Longueur max.	Total (mots)
Titres de section	14 027	4 mots	1 mot	28 mots	50 041
Résumés	1534	118 mots	16 mots	907 mots	180 972
Introductions	1202	422 mots	23 mots	1798 mots	494 376
Conclusions	1165	241 mots	21 mots	1171 mots	275 147

TABLE 1 – Nombre et taille des zones de documents étudiées dans le corpus

3.2 Modèles distributionnels

Ces corpus ont été utilisés pour construire des modèles distributionnels en utilisant une technique fréquentielle classique, basée sur l'outil DISSECT (Dinu *et al.*, 2013). Ces modèles sont plus précisément construits en extrayant les cooccurrences entre mots dans une fenêtre symétrique de trois mots (en respectant les limites des phrases et des éléments de structure). Les mots sont identifiés par leur lemme et leur catégorie grammaticale. Nous retenons ceux qui appartiennent à une classe ouverte (nom, adjectif, verbe ou adverbe) et qui ont une fréquence minimale de 50 occurrences. La valeur de la cooccurrence entre deux mots est mesurée par le score d'information mutuelle positive (PPMI) et la similarité entre les mots par la mesure de cosinus entre les vecteurs correspondants, sans étape préalable de réduction du nombre de dimensions de la matrice. La simplicité de ces modèles est justifiée par notre volonté de préserver leur interprétabilité et notamment de pouvoir identifier le rôle des contextes individuels sur les similarités calculées.

Nous avons ainsi construit cinq modèles distributionnels à partir des cinq corpus présentés en 3.1, que nous appelons : (1) modèle complet, (2) modèle sans titres, (3) modèle sans résumés, (4) modèle sans introductions, (5) modèle sans conclusions. La comparaison entre le modèle complet et un modèle calculé sur un sous-corpus nous permet d'observer les variations induites par le retrait d'une des quatre zones de document et d'identifier les mots dont la représentation est modifiée de façon importante, ou est au contraire inchangée.

3.3 Modèles de contrôle

Afin de vérifier si ces éléments structurels jouent un rôle spécifique, nous avons également construit des modèles basés sur des corpus construits en retirant de manière aléatoire une quantité de mots égale à celle des structures étudiées, sans scinder les phrases. Pour chacun des quatre sous-modèles nous avons construit dix modèles aléatoires équivalents. L'analyse de ces derniers par rapport aux sous-modèles originaux nous aide à vérifier si la variation est seulement causée par la diminution de la taille du corpus ou si le contenu des zones ciblées a un impact spécifique.

Nous avons également testé des versions plus récentes des modèles distributionnels, en construisant des modèles prédictifs. Nous nous sommes limités à l'utilisation de la version de base de *Word2vec* (Mikolov *et al.*, 2013) et construit pour chacun des cinq corpus un modèle *SGNS* (*skip-gram with negative sampling*) en choisissant pour les hyperparamètres les valeurs par défaut de l'outil (plus précisément de son implémentation dans la bibliothèque GenSim (Řehůřek & Sojka, 2010)) sauf pour ceux qui font écho aux caractéristiques des modèles fréquentiels précédents : fréquence minimale de 50 occurrences et fenêtre de taille 3 pour les contextes.

3.4 Mesure de la variation entre deux modèles

Pour évaluer la variation nous utilisons le *Ranked Biased Overlap* (ou RBO) (Webber *et al.*, 2010) qui calcule la similarité entre deux listes ordonnées L_1 etL_2 jusqu'au rang n suivant la formule suivante :

$$RBO(L_1, L_2, n) = (1-p)\sum_{k=1}^{n} p^{k-1}\frac{|L_1(1:k) \cap L_2(1:k)|}{k}$$

Où $L(1:k)$ représente les k premiers éléments de la liste L. Cette mesure considère donc le recouvrement partiel des deux listes comparées à chaque rang, en accordant plus d'importance aux débuts des listes. Le paramètre p (fixé ici à 0,9) est le coefficient d'atténuation de la prise en compte des différences lorsque l'on avance dans les listes. Les listes comparées ici sont les $n = 20$ plus proches voisins d'un mot donné dans deux modèles distributionnels. Le score RBO est normalisé pour varier de 0 (aucun voisin commun) à 1 (listes identiques).

La table 2 illustre cette mesure dans le cas du nom *significativité* dont les 5 premiers voisins sont donnés pour 3 modèles. On peut voir que le RBO est très élevé (0,89) dans le cas d'une suppression des titres de section, avec une différence très faible entre les deux listes (identiques pour les 5 premiers voisins). Le voisinage est de plus en plus perturbé par les autres suppressions, avec un score RBO qui tombe à 0,1 lorsqu'on enlève les conclusions.

Rang	**Modèle complet**	**Modèle sans titres**	**Modèle sans résumés**	**Modèle sans introductions**	**Modèle sans conclusions**
1	corrélation	corrélation	corrélation	corrélation	performance
2	statistiquement	statistiquement	performance	statistiquement	proximité
3	performance	performance	statistiquement	proximité	statistiquement
4	proximité	proximité	proximité	significatif	empiriquement
5	cohésion	cohésion	empiriquement	performance	confiance
RBO	-	0,89	0,46	0,34	0,10

TABLE 2 – Comparaison du score RBO et des voisins du mot *significativité* entre le modèle de référence et chacun des quatre modèles obtenus par suppression d'une zone de texte dans le corpus

Nous avons également considéré deux autres mesures qui permettent de comparer le voisinage distributionnel d'un mot entre deux modèles. La première est le coefficient de Jaccard, autrement dit le ratio de voisins en commun qu'a un mot entre les deux modèles. Cette mesure a été couramment utilisée pour comparer des modèles distributionnels, notamment dans (Pierrejean & Tanguy, 2018b). Nous l'avons calculé en prenant en compte, comme pour le RBO, les 20 premiers voisins de chaque mot dans l'ordre de similarité. La différence principale avec le RBO est que l'ordre des voisins n'est pas pris en compte.

La dernière mesure que nous avons calculée se concentre exclusivement sur le plus proche voisin du mot dans chacun des deux modèles comparés, et consiste en la moyenne des différences de rangs de ces plus proches voisins, comme indiqué dans la formule ci-dessous :

$$diffrang_{M_1,M_2}(m) = \frac{|rang^m_{M_2}(m_1)| + |rang^m_{M_1}(m_2)|}{2}$$

où m_i est le plus proche voisin du mot-cible m suivant le modèle M_i et $rang^m_M(n)$ est le rang du mot n dans la liste des voisins du mot m suivant le modèle M, ordonnés par similarité décroissante

(en comptant à partir de 0). Suivant cette formule, une différence de zéro indique que les deux plus proches voisins sont identiques et une différence importante indique que l'un ou l'autre de ces premiers voisins (ou les deux) se trouve relégué plus loin dans la liste.

3.5 Mesure de la spécificité des mots

Dans un corpus spécialisé, la question de la qualité de la représentation des termes du domaine est cruciale. Afin de distinguer l'impact des modèles sur les mots spécifiques au domaine du TAL, nous avons utilisé une mesure de spécificité en comparant les fréquences dans notre corpus avec celles du corpus FrWac telles que fournies par le lexique GLAFF (Sajous *et al.*, 2013). Nous avons utilisé le score de χ^2 pour identifier les mots dont la fréquence relative est significativement plus élevée dans le corpus TALN. Les mots relatifs au domaine de spécialité comme *corpus*, *sémantique*, *annotation*, *supervisé* ou encore *syntaxique* se démarquent ainsi avec un χ^2 élevé. On y retrouve aussi des mots du discours scientifique tels que *afin* et *ci-dessous*. Certains mots sont, à l'inverse, sous-représentés (*avoir*, *numéro* ou *national*). Pour les calculs statistiques utilisant cette mesure, nous avons attribué un signe négatif au χ^2 des mots dont la fréquence relative est inférieure dans le corpus TALN.

Nous avons repris le même procédé pour identifier le vocabulaire spécifique aux zones de texte étudiées par rapport au corpus TALN entier. La table 3 donne les cinq mots les plus spécifiques de chaque partie envisagée (par rapport au reste du corpus).

Titres de section	**Résumés**	**Introductions**	**Conclusions**
introduction	article	introduction	conclusion
conclusion	présenter	section	perspective
référence	automatique	automatique	envisager
perspective	montrer	travail	améliorer
discussion	proposer	langue	montrer

TABLE 3 – Mots avec le score de spécificité (χ^2) le plus élevé pour chaque partie d'article

On voit que les mots les plus spécifiques aux titres sont ceux des sections génériques (*introduction*, *conclusion*, *références* etc.), que l'on retrouve également dans les parties correspondantes. Les mots des autres colonnes sont aisément interprétables comme des éléments de formulations canoniques pour les résumés ("Dans cet article nous présentons/montrons/proposons [...]") ou les conclusions ("Nous avons montré que [...] ; dans la suite nous envisageons de [...]"). Les introductions contiennent quant à elles des termes très génériques du domaine (*langue*, *automatique*) qu'on retrouvera moins dans le développement du texte, conformément au rôle de cette section de créer une niche au sein d'un espace de recherche englobant (Swales, 1990). Quant à *section*, il correspond à la présentation du plan de l'article qui est une des fonctions de l'introduction d'un article scientifique.

4 Analyse

Toutes les analyses présentées ici portent sur les mots dont la fréquence est de 50 ou plus dans chacun des quatre sous-corpus considérés, soit 3107 mots représentés par leur catégorie et leur lemme (uniquement les classes ouvertes).

4.1 Variation entre les modèles distributionnels

Dans un premier temps, nous avons calculé le score RBO de chaque lemme en comparant chaque modèle au modèle de référence. La figure 2 montre la distribution de ce score dans les quatre cas.

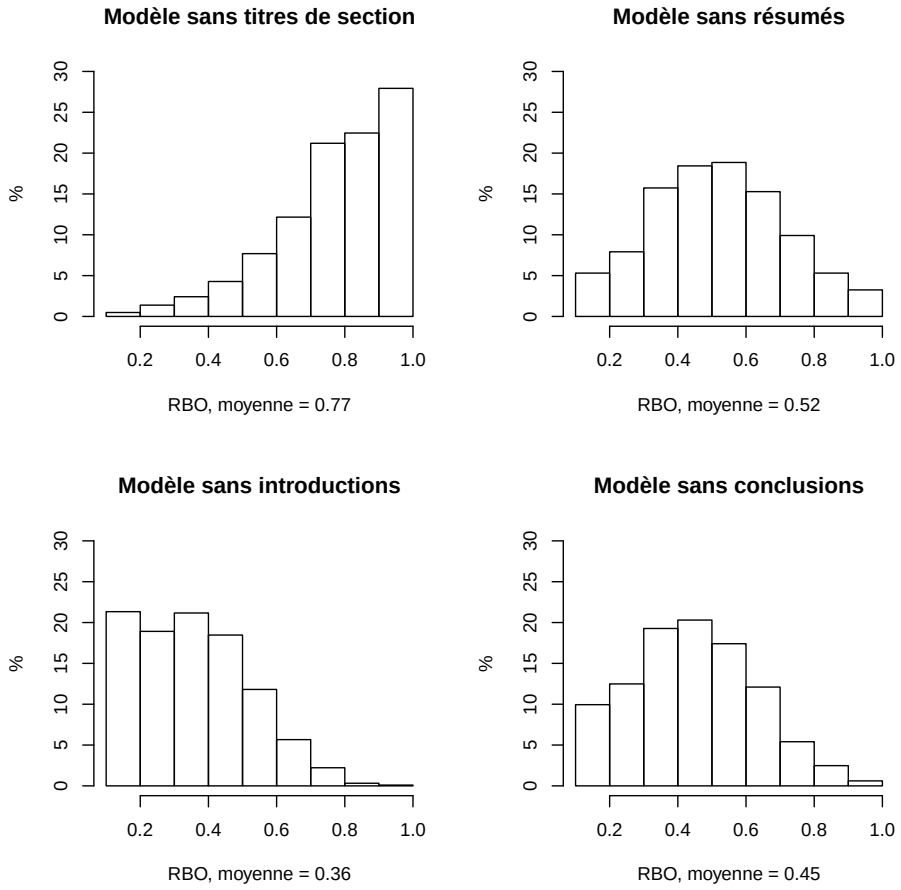

FIGURE 2 – Distribution du score de RBO pour les quatre modèles comparés à celui du corpus complet

Dans le modèle sans titres de section, le RBO moyen est à 0,77, avec une importante proportion de valeurs très élevées, correspondant à des voisinages distributionnels inchangés (RBO de 1). Ce score est nettement plus bas pour les autres modèles : il tombe à 0,36 lorsqu'on supprime les introductions, ce qui correspond au modèle le plus éloigné du corpus entier, avec un nombre important de voisinages radicalement différents (RBO proche de 0). Les modèles correspondant à la suppression des résumés et des conclusions présentent des profils intermédiaires (RBO moyen de 0,52 et 0,45 respectivement) avec une distribution plus symétrique et plus homogène, indiquant que la plupart des mots y voient leur voisins sémantiques partiellement modifiés, mais que les perturbations majeures sont minoritaires.

Cette différence de comportement s'explique en grande partie par le fait que l'amplitude de la variation est proportionnelle à la taille de la zone supprimée. Néanmoins, il apparaît au vu des distributions de la figure 2 que les scores de RBO varient de façon importante au sein du lexique, l'ensemble du spectre étant occupé dans les quatre cas. Ces observations soulèvent deux questions : l'instabilité d'un modèle dépend-elle uniquement de la quantité de texte supprimé ou bien la nature des zones de document ciblées joue-t-elle aussi un rôle déterminant ? Peut-on identifier les principaux facteurs expliquant les variations internes au sein du lexique ?

La première question nous amène à l'analyse des modèles aléatoires. Nous avons calculé la moyenne des RBO (par rapport au modèle complet) pour chacune des quatre séries de 10 modèles aléatoires puis nous l'avons comparée au RBO moyen de chacun des quatre modèles correspondant aux suppressions de parties spécifiques, les résultats étant présentés en table 4.

Zones supprimées	RBO du modèle	RBO moyen des 10 modèles aléatoires (IC 95%)
Titres de section	$\mathbf{0,772}$	$0,757 \pm 0,002$
Résumés	$\mathbf{0,519}$	$0,497 \pm 0,003$
Introductions	$0,360$	$0,358 \pm 0,002$
Conclusions	$0,447$	$0,443 \pm 0,007$

TABLE 4 – Comparaison des RBO moyens pour les quatre modèles et de la moyenne sur les dix modèles aléatoires équivalents en taille, par rapport au modèle complet.

On peut y voir une différence entre le modèle sans titres de section et les dix modèles aléatoires équivalents : le RBO est plus élevé (*i.e.* la variation est moindre) quand on enlève ces titres que lorsqu'on retire une quantité de texte identique choisie aléatoirement (0,772 contre 0,757), la différence étant significative pour les 10 modèles ($p < 0,05$) suivant le test des rangs signés de Wilcoxon sur l'ensemble du vocabulaire. L'analyse du modèle sans les résumés donne des résultats identiques avec un RBO moyen de 0,519 contre 0,497 sur les modèles aléatoires (différence significative également). Par contre, le RBO moyen d'un modèle sans introductions est identique à celui des modèles aléatoires (0,360 et 0,358, aucune différence significative pour les 10). La situation des conclusions est plus mitigée mais la différence avec les modèles aléatoires n'est significative que dans 5 cas sur 10, ce qui nous conduit à conclure que la différence n'est pas suffisamment marquée pour être prise en considération.

Les résumés et les titres de section ont donc bien un effet singulier, mais qui va à l'encontre de notre hypothèse de départ : ces zones influencent moins les modèles distributionnels que d'autres parties de texte de taille équivalente. Quant aux introductions et aux conclusions, elles n'apportent aucune différence nette : elles se comportent sur ce plan comme le reste du texte.

Nous avons enfin calculé le coefficient de corrélation (ρ de Spearman) entre le RBO et le χ^2 du vocabulaire distinctif de chaque zone. La relation linéaire est négative pour les quatre modèles au même niveau : titres de section ρ = -0,39, résumés ρ = -0,37, introductions ρ = -0,38 et conclusions ρ =-0,35. En toute logique, plus un mot est spécifique à une zone du texte (quelle qu'elle soit), moins sa représentation a de chances d'être stable lorsqu'on la retire. Ce n'est cependant pas systématique car la relation linéaire négative n'est pas si forte. La variation dépend donc également d'autres facteurs.

4.2 Variation sur le vocabulaire spécifique

Nous avons cherché à mesurer si la représentation du vocabulaire spécifique du corpus TALN était plus affectée que le vocabulaire courant par le retrait d'une des différentes zones de document. Nous avons mesuré son impact sur le vocabulaire spécialisé du corpus (3.5) en calculant le coefficient de corrélation de Spearman entre le χ^2 (obtenu en comparant les fréquences dans notre corpus et dans FrWac) et le RBO dans la table 5.

Pour les modèles sans résumés, sans introductions et sans conclusions le ρ de Spearman est positif (de 0,15 à 0,24). Les mots spécifiques du corpus TALN (avec un χ^2 élevé) tendent donc à être moins perturbés dans leur représentation (RBO plus haut) que les autres. La corrélation est pratiquement

Modèles	ρ de Spearman
Sans titres de section	-0,06
Aléatoires	0,10
Sans résumés	0,15
Aléatoires	0,18
Sans introductions	0,24
Aléatoires	0,19
Sans conclusions	0,15
Aléatoires	0,18

TABLE 5 – Corrélation de Spearman entre spécificité du mot dans le corpus (χ^2) et variation (RBO) par rapport au modèle complet

nulle (-0,06) pour les titres de section qui n'influent pas plus sur le vocabulaire du domaine que sur les autres mots. Si l'on regarde les modèles aléatoires utilisés comme point de comparaison, la corrélation entre le χ^2 et la moyenne du RBO de ceux-ci est plus élevée, sauf pour les introductions. L'effet de perturbation des mots du domaine semble donc être amoindri à mesure que l'on ôte des portions de texte plus larges.

4.3 Autres mesures

Comme indiqué en section 3.4, nous avons mesuré la variation du voisinage distributionnel d'un mot en utilisant deux autres techniques : le recouvrement des 20 premiers voisins avec le coefficient de Jaccard et la différence de rang des premiers voisins.

Concernant le Jaccard, il est fortement corrélé avec le RBO (ρ de 0,6 à 0,8 suivant les modèles) si bien que l'ensemble des conclusions obtenues précédemment avec le RBO sont confirmées par cette mesure.

La variation du rang des premiers voisins, quant à elle, présente un profil de distribution différent : pour les quatre paires de modèles ce score est nul dans 87% des cas (i.e. les deux modèles comparés ont le même premier voisin pour un mot donné) ce qui rend son usage très délicat pour les analyses statistiques. Sa corrélation est de ce fait faible avec les deux autres coefficients (ρ de 0,1 à 0,4 en valeur absolue [5]). Par contre, ce score est très utile pour isoler les cas de perturbation importante du voisinage distributionnel et donc effectuer une analyse qualitative permettant de mieux comprendre les mécanismes à la base de ces variations.

Nous avons enfin voulu comparer ces résultats à ceux que l'on obtient en utilisant des méthodes prédictives de construction des représentations distributionnelles. En mesurant les scores RBO pour comparer les modèles construits sur les sous-corpus à celui du corpus complet, nous n'observons que très peu de différence entre les modèles. Le RBO moyen va en effet de 0,62 à 0,68 en suivant le même ordre que celui observé en 4.1, mais avec un resserrement qui ne permet pas de distinguer clairement le rôle des différentes parties de corpus supprimées. On observe d'ailleurs une très forte corrélation entre les 4 séries de score RBO (0,7 en moyenne, contre seulement 0,4 sur les modèles fréquentiels). Il nous apparaît donc clairement que ces modèles prédictifs sont des instruments trop grossiers pour mettre au jour des variations fines, et que les différences entre les corpus d'apprentissage ne sont pas suffisantes au vu des techniques qui tendent à lisser le matériau (notamment les processus aléatoires qui ont une influence très importante sur les modèles, *cf* (Pierrejean & Tanguy, 2018a)).

5. La corrélation est négative, deux listes identiques entraînent un RBO de 1 et une variation de rang de 0.

4.4 Analyse qualitative

Dans cette dernière partie nous effectuons des observations plus fines en nous basant sur les scores (RBO et différence de rang), les listes des plus proches voisins, et les contextes d'apparition. Nous nous concentrons tout d'abord sur les variations des représentations des termes dans le modèle **sans résumés**, en observant les mots présentant des variations importantes. On y trouve deux cas de figure distincts :

— Des mots spécifiques aux résumés (fort χ^2 par rapport aux autres parties du corpus), qui subissent donc en toute logique une forte variation de voisins relative à la diminution de leur fréquence. C'est le cas de *démonstration*, *assistance*, *arboré* et *informatisé*. Ces termes sont employés régulièrement dans le domaine.

— Des mots non spécifiques aux résumés mais présentant contre toute attente une forte variation. Ces mots sont pour la plupart polysémiques, tels que *synthétique*, *couper*, *interrompre* et ont des emplois multiples dans le corpus. Cela se voit notamment dans la grande variété de leurs voisins, que l'on peut rattacher à plusieurs sens. Leurs contextes sont donc très variés. Logiquement, une légère modification des données d'entrée du calcul distributionnel suffit à provoquer leur instabilité, sans qu'on puisse pour autant identifier clairement une variation de sens.

À l'inverse, si l'on regarde du côté des mots très stables (RBO élevé) lorsqu'on supprime les résumés, on trouve deux cas de figure :

— Des mots spécifiques aux résumés dont les voisins sont stables malgré la baisse importante de leur fréquence. Parmi eux on retrouve les verbes d'exposition tels que *présenter*, *finaliser* ou encore des évaluatifs comme *vaste*, *important*, *faiblement*. Leur stabilité viendrait de leur ancrage dans des contextes très réguliers, ce qui les rendrait moins vulnérable à la suppression d'un passage spécifique. La suppression de ces zones n'affecte donc pas leur représentation.

— Des mots non spécifiques aux résumés et dont le voisinage n'est pas affecté tels que *nœud*, *pattern* ou *possessif*. Ils sont aussi inscrits dans des zones de texte stables indépendantes de la structure du document.

Cette étude qualitative confirme certains résultats des analyses précédentes. Tout d'abord on retrouve peu de mots du domaine du TAL affectés par l'absence de résumés. La plupart du vocabulaire instable est en fait spécifique à cette zone du texte et pas au corpus. Cependant on remarque que certaines classes (ou clusters) de mots sont très stables et leur représentation le restera quels que soient la quantité ou le type de texte supprimé. L'instabilité d'un mot n'est donc pas seulement provoquée par la suppression de ses occurrences mais dépend surtout des contextes dont il est entouré.

L'examen des variations entraînées par la suppression d'une autre zone nous permet de dégager des phénomènes sémantiques intéressants. C'est le cas lorsqu'on observe les voisinages fortement impactés par la suppression des **conclusions**. Parmi les mots spécifiques à cette partie des articles, on trouve le champ lexical habituel à la présentation des perspectives d'un travail de recherche : *futur*, *prévoir*, *affiner*, *approfondir*, *poursuivre*. Les perturbations des voisinages sont importantes mais limitées à un réordonnancement local des mots les plus proches sans tendance globale précise, comme on l'a vu pour les résumés. En revanche, pour certains de ces mots spécifiques la baisse de fréquence s'accompagne plus clairement d'un changement de sens. C'est le cas notamment de l'adjectif *idéal* qui voit ses voisins passer d'un sens technique (*optimal*, *final*, *contrôlé*) à un évaluatif (*trivial*, *véritable*, *suffisant*). Une autre distinction polysémique est observable pour *engager* dont les emplois correspondent au sens de /participer/, /initier une action/ (avec des voisins comme *participant*, *organisateur*, *entreprendre*), ou au sens plus abstrait et dialectique : *défendre*, *exprimer*, *confronter*, *inciter*.

Ces observations nous conduisent à formuler l'hypothèse que la stabilité d'un mot quand on compare deux modèles distributionnels est liée à l'existence d'un groupe de "bons" voisins, dont l'identification résiste à des modifications partielles des contextes. L'absence de tels points d'ancrage semble entraîner une forte instabilité, quelle que soit la modification subie par le corpus. Les cas intéressants mais rares (qu'illustrent *idéal* et *engager*) que nous recherchons correspondraient à des situations où la représentation d'un mot passe d'un cluster de proches voisins à un autre.

5 Conclusion et perspectives

Les études de variation entre un modèle distributionnel de référence et d'autres modèles obtenus en retirant une partie du document des mêmes données d'apprentissage nous permettent de tirer certaines conclusions concernant l'impact de la structure des articles scientifiques sur la représentation distributionnelle d'un corpus spécialisé. De manière globale les résumés et titres de section impactent moins le voisinage des mots que les autres parties de texte. Ce résultat, qui semble contre-intuitif, peut s'expliquer en faisant l'hypothèse que ces passages sont redondants par rapport au reste du texte, ou qu'y figurent des mots qui s'y comportent différemment, comme nous le suggère l'analyse qualitative. Quant aux introductions et aux conclusions, elles n'ont pas d'influence significative, les enlever a le même effet que retirer une part aléatoire équivalente du corpus.

La représentation des mots spécifiques au domaine et au genre n'est pas affectée par le retrait des titres et se trouve même plus stable que les autres lorsque les résumés et introductions sont retirés. Nous pouvons en déduire que le cœur du vocabulaire technique n'est pas particulièrement employé à l'intérieur de ces zones. Seul le vocabulaire propre à la zone est logiquement touché. Ce dispositif nous pousse à penser que ces parties de document scientifique ne jouent pas de rôle prépondérant vis-à-vis de l'analyse distributionnelle. Une optimisation/pondération des contextes au regard de la zone dans laquelle ils se trouvent serait donc négligeable voir déconseillée selon l'objectif de modélisation recherché.

Pour mettre au jour de véritables variations entre les emplois des mots dans les différentes parties d'un document, l'idéal aurait été de comparer directement des espaces sémantiques construits sur des parties différentes du corpus, à la manière des travaux initiés par (Hamilton *et al.*, 2016) pour la diachronie. Cependant, notre corpus et les sous-parties visées représentent clairement un trop petit volume pour ce type de méthode. L'approche indirecte que nous avons suivie ici pourrait néanmoins être poursuivie à plus large échelle, mais devrait alors se concentrer sur le discours scientifique général, puisque la restriction à un domaine particulier ne permettrait pas d'atteindre la masse critique nécessaire.

Remerciements

Le travail présenté dans cet article a été réalisé dans le cadre du projet ADDICTE [6] (Analyse distributionnelle en domaine spécialisé), financé par l'Agence Nationale de la Recherche (ANR-17-CE23-0001). Il a également reçu le soutien du consortium CORLI pour la finalisation du corpus TALN. Les auteurs tiennent à remercier l'ATALA et son président Patrick Paroubek pour avoir autorisé l'utilisation et la diffusion du corpus constitué des actes des conférences TALN et RECITAL. Ils remercient enfin Alice Adnot-Albinet, Charline Fabre et Clémentine Mailly pour leur aide dans la correction du corpus.

6. https://anr-addicte.ls2n.fr/

Références

BADENES-OLMEDO C., GARCÍA J. L. R. & CORCHO O. (2017). An initial analysis of topic-based similarity among scientific documents based on their rhetorical discourse parts. In *Workshop on Enabling Open Semantic Science co-located with the 16th International Semantic Web Conference (ISWC)*, p. 15–22, Vienna, Austria.

BERNIER-COLBORNE G. (2014). Identifying semantic relations in a specialized corpus through distributional analysis of a cooccurrence tensor. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (* SEM 2014)*, p. 57–62.

BERTIN M. & ATANASSOVA I. (2014). A study of lexical distribution in citation contexts through the IMRaD standard. In *Proceedings of the First Workshop on Bibliometric-enhanced Information Retrieval co-located with the 36th European Conference on Information Retrieval (ECIR 2014)*, Amsterdam, Netherlands.

BOUDIN F. (2013). TALN Archives : une archive numérique francophone des articles de recherche en Traitement Automatique de la Langue. In *Actes de TALN'2013*, p. 507–514, Les Sables d'Olonne.

COHEN T. & WIDDOWS D. (2009). Empirical distributional semantics : methods and biomedical applications. *Journal of biomedical informatics*, **42**(2), 390–405.

COUNCILL I. G., LEE GILES C. & KAN M.-Y. (2008). An open-source CRF reference string and logical document structure parsing package. In *Proceedings of the Language Resources and Evaluation Conference (LREC 08), Marrakesh, Morrocco, May.*

DINU G., THE PHAM N. & BARONI M. (2013). Dissect - distributional semantics composition toolkit. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, p. 31–36.

EL BOUKKOURI H., FERRET O., LAVERGNE T. & ZWEIGENBAUM P. (2019). Embedding strategies for specialized domains : Application to clinical entity recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics : Student Research Workshop*, p. 295–301.

HAMILTON W. L., LESKOVEC J. & JURAFSKY D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, p. 1489–1501, Berlin.

HOFMANN K., TSAGKIAS M., MEIJ E. & DE RIJKE M. (2009). The impact of document structure on keyphrase extraction. In *Proceedings of the 18th ACM conference on Information and knowledge management*, p. 1725–1728.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. In *ICLR 2013, workshop track*.

PIERREJEAN B. & TANGUY L. (2018a). Predicting word embeddings variability. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM)*, New Orleans.

PIERREJEAN B. & TANGUY L. (2018b). Towards qualitative word embeddings evaluation : Measuring neighbors variation. In *Conference of the North American Chapter of the Association for Computational Linguistics : Student Research Workshop*, p. 32–39.

ŘEHŮŘEK R. & SOJKA P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, p. 45–50, Valletta, Malta : ELRA.

SAHLGREN M. & LENCI A. (2016). The effects of data size and frequency range on distributional semantic models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 975–980, Austin, Texas : Association for Computational Linguistics.

SAJOUS F., HATHOUT N. & CALDERONE B. (2013). GLÁFF, un Gros Lexique Á tout Faire du Français. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, p. 285–298, Les Sables d'Olonne, France.

SOLLACI L. B. & PEREIRA M. G. (2004). The introduction, methods, results, and discussion (imrad) structure : a fifty-year survey. *Journal of the medical library association*, **92**(3), 364.

SWALES J. (1990). *Genre analysis : English in academic and research settings*. Cambridge University Press.

TEUFEL S. & MOENS M. (2002). Summarizing scientific articles : experiments with relevance and rhetorical status. *Computational linguistics*, **28**(4), 409–445.

TEUFEL S., SIDDHARTHAN A. & BATCHELOR C. (2009). Towards discipline-independent argumentative zoning : evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 3-Volume 3*, p. 1493–1502 : Association for Computational Linguistics.

WEBBER W., MOFFAT A. & ZOBEL J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, **20**.