Contextual-RTM : un cadre général pour la modélisation de thématiques dans les réseaux de documents

Jean Dupuy, Adrien Guille, Julien Jacques

Contextual-RTM : un cadre général pour la modélisation de thématiques dans les réseaux de documents

Jean Dupuy[*,**], Adrien Guille[**], Julien Jacques[**]

[*]MeetSYS
[**]Université de Lyon, Lyon 2, ERIC EA3083

Résumé. Les longs parcours de bases de connaissance ou de grands corpus de documents en réseau peuvent parfois perdre le lecteur et lui faire manquer certains liens entre les documents qu'il consulte. Nous proposons ici un cadre de modélisation thématique pour les réseaux de documents, Contextual-RTM, qui généralise *Relational Topic Model* (RTM). Alors que RTM est agnostique à la position des liens dans les documents, Contextual-RTM les prend en compte, permettant une meilleure contextualisation. Nous définissons le contexte comme une fenêtre de taille ajustable centrée autour du lien et proposons 3 méthodes d'agrégation du contexte : uniforme, positionnelle et sémantique. Contextual-RTM se montre compétitif sur des tâches d'identification de mots à l'origine de liens entre documents. Nous intégrons ces méthodes dans un système d'aide à la lecture capable d'inférer localement des liens latents entre documents. Ainsi le lecteur garde une trace de ses précédentes lectures, et s'en voit recommandé de nouvelles.

1 Introduction

Nous nous intéressons à la prédiction de liens entre documents, avec comme objectif la mise au point d'un système de recommandation et d'aide à la lecture. Nous nous attachons plus particulièrement aux réseaux de documents où les liens entre les pages sont localisés dans le texte. C'est le cas par exemple des liens hypertextes dans les pages Wikipédia. C'est également le cas des citations dans les articles scientifiques.

Nous portons une attention particulière à la localisation des liens dans le texte, et aux mots qui les entourent (que nous appelons leur contexte) car nous pensons que cette information est porteuse de sens et permet de mieux expliquer l'existence d'un lien entre deux documents. Par exemple les liens qui se trouvent dans l'introduction d'un article scientifique sont généralement vers d'autres articles relativement généraux, ceux dans le corps de l'article plus spécialisés sur une tâche similaire à celle qui est traitée dans l'article, et ceux en conclusion vers des tâches spécifiques mais différentes.

Tandis que les système existants recommandent un ensemble de documents à lire, étant donné le dernier document consulté, nous souhaitons mettre au point un système

guidant encore plus la lecture. Notre objectif est de positionner directement les recommandations à l'intérieur du document en cours de lecture. Un tel système ne peut-être mis en œuvre sur la base des méthodes de prédiction de liens existantes, qui modélisent un lien entre un document A et un document B.

Par conséquent, nous proposons dans cet article une nouvelle méthode de prédiction de lien, modélisant les liens entre une portion du document A (un mot ou groupe de quelques mots) et un document B. Celle-ci repose sur une généralisation d'un modèle thématique probabiliste : Relational Topic Model Chang et Blei (2009).

RTM modélise les liens par une fonction de score, dont le résultat est d'autant plus grand que les proportions de thématiques des deux documents liés sont plus similaires. Cependant RTM ne prend pas en compte le contexte d'apparition des liens car le score de lien est fonction de tous les mots de chacun des deux documents. Nous proposons avec CRTM trois variantes de cette fonction score de lien, qui diffèrent selon l'importance qui est donnée aux mots du contexte. Ces trois variantes (uniforme, positionnelle et sémantique) nous semblent couvrir les approches les plus générales permettant de prendre en compte le contexte d'apparition des liens. D'autres méthodes plus spécifiques peuvent bien entendu être envisagées.

Nous montrons d'abord que CRTM surpasse RTM sur une tâche de prédiction des mots à l'origine des liens, tant en terme de précision @k qu'en terme d'aire sous la courbe ROC. Pour ces évaluations RTM est entraîné avec la fonction de lien originale. Nous substituons celle-ci à une fonction de lien de CRTM lors de la prédiction, car la fonction de lien originale ne permet pas de répondre à la tâche évaluée. Nous montrons ainsi que la prise en compte du contexte lors de l'entraînement des modèles donne de meilleurs résultats que son exploitation uniquement lors de la prédiction. Nous illustrons ensuite sur des exemples comment elle permet de mieux positionner les documents recommandés, en les associant à des termes pertinents dans les documents en cours de lecture.

2 État de l'art

Les corpus de documents sont étudiés depuis longtemps. En particulier, beaucoup d'attention a été donnée à la représentations de ces documents, soit par des méthodes probabilistes, par exemple pour inférer une structure thématique latente comme dans *Latent Dirichlet Allocation* (LDA) de Blei et al. (2003), soit par des méthodes visant à apprendre des représentations denses en faible dimension comme Doc2Vec de Le et Mikolov (2014).

Très souvent ces corpus ne sont pas une simple collection de textes et s'avèrent être structurés en réseaux. On pense notamment à la littérature scientifique, les contenus encyclopédiques comme Wikipedia, les bases de connaissances ou plus généralement le Web. De nombreuses méthodes s'appuient sur cette information supplémentaire pour déterminer de meilleures représentations. Certaines méthodes se basent sur des modèles probabilistes pour représenter les thématiques latentes, comme NetPLSA (Zhang et al., 2016) ou RTM (Chang et Blei, 2009) et d'autres sur des méthodes d'apprentissage de plongements de document comme TADW (Yang et al., 2015), CANE (Tu et al., 2017),

GVNR (Brochier et al., 2019), RLE (Gourru et al., 2020) ou IDNE (Brochier et al., 2020).

Nous retenons l'approche fondée sur la modélisation de thématiques car elle fournit un cadre formel pour apprendre les représentations des documents et autorise une prise en compte simple du contexte. Pour toute ces raisons nous nous basons sur *Relational Topic Model*. Puisque RTM se base sur LDA, nous commençons par détailler ce dernier avant de présenter RTM.

2.1 LDA

Latent Dirichlet Allocation (Blei et al., 2003) est un modèle génératif probabiliste. En supposant qu'un document répond d'un mélange d'un petit nombre K de thématiques, il vise à en attribuer une à chaque mots observés.

Modèle : Soit un corpus de documents, dont chaque document $d \in [1, D]$ comporte N_d mots appartenant à un vocabulaire fixe V. Soit $\mathbf{w}$ tous les mots observés de touts les documents, on note $w_{d,n} \in V$ la valeur prise par le n-*ième* mot du document d. On considère plusieurs variable latentes pour expliquer $\mathbf{w}$: $\mathbf{z}$ et $\boldsymbol{\theta}$, avec $z_{d,n} \in \{1, \cdots, K\}$ l'affectation thématique de $w_{d,n}$ et θ_d le mélange de thématiques du document d. On introduit enfin $\alpha \in \mathbb{R}^K$ et $\beta \in \mathbb{R}^{K \times V}$, respectivement le paramètre de la distribution des thématiques des documents et l'a priori sur la distribution des mots par thématiques.

Processus génératif et estimation : Pour chaque document d on tire le mélange de K thématiques θ_d selon une distribution de Dirichlet de paramètre α. Ensuite, pour chaque mot $w_{d,n}$ de d on tire sa thématique latente $z_{d,n}$ à partir d'une distribution multinomiale de paramètre θ_d. Enfin on tire $w_{d,n}$ selon une distribution multinomiale de paramètre $\beta_{z_{d,n}}$. L'estimation des paramètres se fait en trouvant le maximum de vraisemblance. La vraisemblance étant impossible à calculer en pratique on utilise un algorithme Variational EM (Jordan et al., 1999) afin d'optimiser la ELBO (Bishop, 2006) alternativement vis à vis des paramètres variationnels et des paramètres du modèle.

2.2 RTM

Comme dit précédemment, *Relational Topic Model* (Chang et Blei, 2009) étend LDA en y incorporant une information sur les liens entre les documents.

Soit $\mathbf{y}$ la variable aléatoire binaire, indiquant l'observation ou non d'un lien entre deux document. On définit la fonction ψ comme étant une fonction de score de lien :

$$\psi_e(y = 1) = \exp\left(\boldsymbol{\eta}^{\mathrm{T}} \left(\overline{\mathbf{z}}_d \circ \overline{\mathbf{z}}_{d'}\right) + \nu\right) \tag{1}$$

où $\overline{\mathbf{z}}_d = \frac{1}{N_d} \sum_n z_{d,n}$ représente la répartition des thématiques latentes du document d, et ν et η des paramètres.

On remarque que le produit de Hadamard $\overline{\mathbf{z}}_d \circ \overline{\mathbf{z}}_{d'}$ amplifie les thématiques communes et tend à gommer les thématiques distinctes entre les deux documents. De plus, de part la nature linéaire de ces fonctions de score, on constate que le paramètre η,

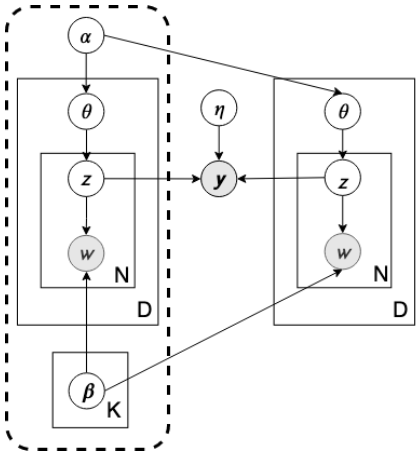


FIG. 1 : Modèle graphique de *Relational Topic Model* (LDA entouré en pointillés)

qui est un vecteur de taille K, ne permet pas de modéliser un éventuel lien entre deux thématiques différentes. On peut donc raisonnablement penser qu'il permet principalement d'amoindrir l'importance de certaines thématiques ne présentant pas d'intérêts pour la modélisation des liens. L'estimation des paramètres β, η et ν s'effectue de la même façon que pour LDA. A noter également que les auteurs ne prennent en compte que les cas où $y = 1$, considérant qu'un lien non observé n'induit pas qu'il n'existe pas. Le processus génératif de RTM est résumé dans l'Algorithme 1 et la figure 1.

pour tous *document d de D* **faire**
  tirer la proportion de thématiques de d : $\theta_d|\alpha \sim Dirichlet\,(\alpha)$
  pour tous *mot* $w_{d,n}$ *de d* **faire**
    tirer l'assignation thématique : $z_{d,n}|\theta_d \sim Multinomiale\,(\theta_d)$
    tirer le mot : $w_{d,n}|z_{d,n} \sim Multinomiale\left(\beta_{z_{d,n}}\right)$
pour tous *paire de documents (d,d′)* **faire**
  tirer l'indicateur de lien : $y_{d,n}\,|\boldsymbol{z}_d, \boldsymbol{z}_{d'} \sim \psi\,(\cdot|\boldsymbol{z}_d, \boldsymbol{z}_{d'})$

Algorithme 1 : Processus génératif de RTM

3 Généralisation de RTM : Contextual-RTM

Contextual-RTM généralise RTM, en proposant maintenant de prendre en compte la localisation et le contexte d'apparition d'un lien au sein du texte, permettant ainsi des usages différents et pouvant être adapté à des corpus de natures variées.

Là où RTM suppose qu'un lien entre deux documents d et d' est marqué par la similitude de $\overline{\mathbf{z}}_d$ et $\overline{\mathbf{z}}_{d'}$, nous modifions la manière de calculer $\overline{\mathbf{z}}_d$ afin de prendre en compte les thématiques des mots du contexte d'apparition des liens. Nous proposons

ici 3 variations de la fonction de lien présentée à l'Equation (1), permettant de prendre en compte ce contexte, de façon uniforme, positionnelle ou sémantique. Nous choisissons d'utiliser la version exponentielle de la fonction de lien car selon Chang et Blei (2009) c'est celle qui donne de meilleurs résultats lors de leurs évaluations.

3.1 Approche Uniforme

On peut considérer qu'un lien est déterminé à égale importance par tous les mots de son contexte immédiat, et les thématiques qui s'y rapportent. Ainsi dans l'Equation 2, nous remplaçons la moyenne $\overline{\mathbf{z}}_d$ des assignations des thématiques dans l'ensemble du document de l'Equation 1 par celle des mots contenus dans une fenêtre de taille l autour du lien :

$$\psi_{uni}(y=1) = \exp\left(\boldsymbol{\eta}^{\mathrm{T}}\left(\overline{\mathbf{z}}_{d,n}^{(u)} \circ \overline{\mathbf{z}}_{d'}\right) + \nu\right), \tag{2}$$

avec

$$\overline{\mathbf{z}}_{d,n}^{(u)} = \frac{1}{2l+1}\sum_{i=n-l}^{n+l} z_{d,i}.$$

Une fenêtre couvrant tout le document permettra de retrouver l'expression de RTM, et une fenêtre de taille nulle ne prendra en compte que le mot à l'origine du lien. On peut également ajuster la taille de cette fenêtre pour la faire coïncider avec une phrase ou un paragraphe.

3.2 Approche positionnelle

On peut également supposer que les mots influencent d'autant plus l'apparition d'un lien qu'ils en sont plus proches dans le texte. Pour modéliser cela nous proposons de lisser les assignations thématiques des mots présents dans une fenêtre de taille l au moyen d'une pondération de type gaussienne, centrée sur l'origine du lien, comme présenté dans l'Equation 3 :

$$\psi_{pos}(y=1) = \exp\left(\boldsymbol{\eta}^{\mathrm{T}}\left(\overline{\mathbf{z}}_{d,n}^{(p)} \circ \overline{\mathbf{z}}_{d'}\right) + \nu\right), \tag{3}$$

avec :

$$\overline{\mathbf{z}}_{d,n}^{(p)} = \frac{1}{2l+1}\sum_{i=n-l}^{n+l} e^{-\frac{1}{2}\left(\frac{i-n}{\sigma}\right)^2} z_{d,i},$$

où σ est l'écart-type choisi.

3.3 Approche sémantique

La dernière approche suppose que tous les mots ne sont pas pertinents pour induire un lien, et que seuls les plus proches sémantiquement ont une véritable influence. Cette approche requiert d'incorporer dans notre modèle de la connaissance externe sur cette proximité sémantique, ce que nous proposons de faire en utilisant des plongements de

mots pré-appris. Nous utilisons dans l'équation 4 le *Scaled-Dot Product* décrit dans Vaswani et al. (2017), variante normalisée de *Dot-Product Attention* des mêmes auteurs permettant que le softmax ne prenne des valeurs trop extrêmes, pour donner plus de poids aux thématiques latentes des mots qui sont les plus proches sémantiquement de ceux à l'origine du lien. On gommera ainsi facilement les mots vides ou non lexicaux présents dans la phrase tout en renforçant l'importance des mots similaires :

$$\psi_{sem}(y=1) = \exp\left(\boldsymbol{\eta}^{\mathrm{T}}\left(\overline{\mathbf{z}}_{d,n}^{(s)} \circ \overline{\mathbf{z}}_{d'}\right) + \nu\right), \tag{4}$$

avec :

$$\overline{\mathbf{z}}_{d,n}^{(s)} = \mathrm{softmax}\left(\frac{e_n.E_l^T}{\sqrt{dim}}\right) z_{d,l}.$$

où :

— e_n est la représentation du mot à la source du lien,
— E_l est la matrice des représentations de tous les mots de la fenêtre,
— dim est la dimension de l'espace de représentation des mots,
— z_l est la matrice des assignations de thématique de tous les mots de la fenêtre.

Cette méthode permet tout autant d'utiliser des plongement de mots entraînés sur les données étudiées que des versions pré-entraînées sur un autre corpus.

4 Expériences

Nous évaluons l'intérêt de la contextualisation des liens en comparant Contextual-RTM à RTM sur leur capacité à retrouver les mots à l'origine d'un lien entre deux documents. Dans un premier temps nous présentons les corpus utilisés et comment ils ont été construits, puis les tâches et les métriques utilisées. Enfin nous commenterons les résultats quantitatifs et qualitatifs de CRTM par rapport à RTM.

4.1 Corpus

Les modèles sont évalués sur des jeux de données constitués des paragraphes introductifs d'articles Wikipedia appartenant à diverses catégories. En effet les corpus habituellement utilisés en prédiction de liens ne donnent pas la localisation des liens dans le texte. Ces corpus sont soit vectorisés (i.e. sac de mots) soit pré-traités (sans marquage des liens dans le texte).

Ils sont construits en parcourant récursivement la chaîne de citations en partant de la page principale d'une catégorie Wikipedia. A chaque itération on ajoute les documents non encore vus, jusqu'à ce que la profondeur désirée soit atteinte.

Deux raisons nous ont poussé à utiliser les paragraphes introductifs de Wikipedia. Premièrement ils sont relativement complets et comportent déjà un nombre important de liens, les rendant comparable à une fiche de base de connaissance en entreprise. Ensuite le format Wikitext, dérivé de Markdown, permet un pré-traitement simple des documents, en particulier la représentation des liens. Un lien dans une page Wikipedia prendra la forme suivante : `[[United_Kingdom|UK]]`. On retrouve à gauche de la barre verticale le titre de la page vers lequel pointe le lien, et à droite le mots qui portera l'hyperlien.

Le tableau 1 résume les propriétés de jeux de données utilisés dans nos évaluations. Celles-ci se feront en deux langues : Français et Anglais.

Catégories	*Physics*	*Arts*	*Society*	*Physique*	*Société*
Langue	Anglais	Anglais	Anglais	Français	Français
Nb de pages	12769	15007	16716	8281	14770
Nb de liens	119671	97774	108275	53738	88643
Nb moyen de mots	149	153	146	149	151
Nb moyen de phrases	7.84	7.74	7.70	4.71	4.65
Connectivité moyenne	9.37	6.52	6.47	6.49	6.00

TAB. 1 : Description des jeux de données.

4.2 Tâches et métriques d'évaluation

Nous évaluons RTM et CRTM sur deux tâches : leur capacité à retrouver le mot à l'origine d'un lien donné entre deux documents en le plaçant parmi les mots les plus probables, et une évaluation qualitative pour juger de la cohérence des mots suggérés. Pour chaque corpus, nous cachons certains liens avant l'entraînement des modèles.

Précision à K et courbe ROC : Pour tous les liens retirés avant l'entraînement des modèles on cherche à prédire quels mots en sont à l'origine et dans quelle proportion ils apparaissent dans le top-K des mots les plus probables. Nous présentons pour chaque corpus et chaque modèle la moyenne des précisions à K pour tous les liens cachés ainsi que l'aire sous la courbe ROC.

Évaluation qualitative : Nous cherchons ici à montrer la capacité de Contextual-RTM à proposer un ensemble de mots cohérent liants des pages et mettant en lumière des liens non triviaux pouvant échapper au lecteur. Nous proposons pour cela de comparer pour deux documents les mots suggérés par RTM et Contextual-RTM les liants à un troisième.

4.3 Méthodes

Pour la première évaluation nous retirons aléatoirement 2000 liens de chaque corpus. Nous comparons RTM aux quatre variantes de Contextual-RTM que nous proposons :

- **RTM** : entraîné comme décrit par Chang et Blei (2009), c'est à dire avec pour fonction de lien Ψ_e (cf. Eq 1). Il est impossible avec cette fonction de lien de mesurer la probabilité que les mots soient à l'origine des liens (tous les mots d'un document ont la même probabilité d'être à l'origine du lien). En conséquence, lors de l'évaluation, nous substituons ψ_e par les fonctions de lien des variantes de CRTM que nous proposons.
- **Variantes de CRTM** :
  - CRTM-Uni-0 : une variante uniforme où le contexte se limite aux mots à l'origine du lien (taille de fenêtre nulle).

— CRTM-Uni-S : une variante uniforme où le contexte correspond à tous les mots de la phrase.
— CRTM-Pos-S : une variante positionnelle où le contexte correspond à tous les mots de la phrase.
— CRTM-Sem-S : une variante sémantique où le contexte corresponds à tous les mots de la phrase. Les représentations des mots utilisés ont été obtenues à partir de *Skip-gram* avec échantillonage négatif (Mikolov et al., 2013), entraîné sur les corpus étudiés, avec une fenêtre de 10 mots.

Tous les modèles sont entraînés avec 50 thématiques, qui conduit aux meilleurs résultats pour RTM et Contextuel-RTM.

Nous fixons l'hyperparamètre α à 5.0 pour tous les modèles (identique à la valeur choisie par Chang et Blei (2009) dans leurs évaluations), celui-ci ayant donné les meilleurs résultats individuellement pour chacun d'entre eux.

4.4 Résultats

4.4.1 Quantitatif

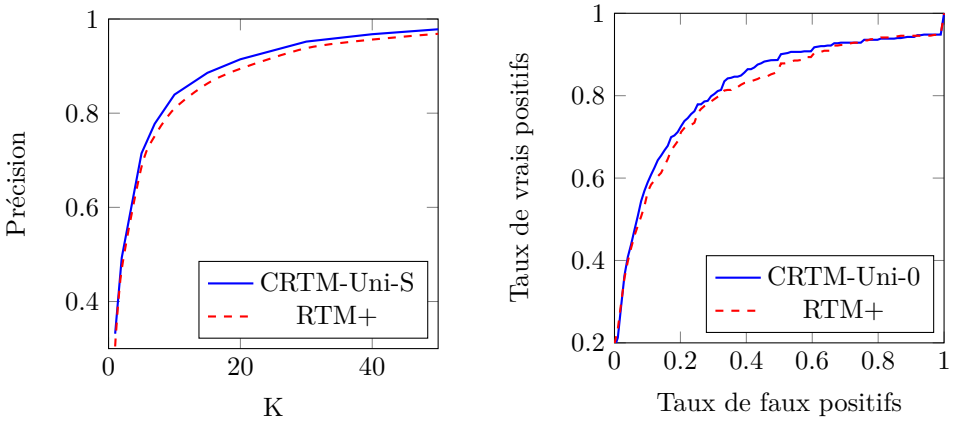
Nous étudions ici la précision moyenne à K pour différents corpus, reportée dans la Table 2. Nous présentons pour RTM le meilleur (RTM+) et le pire (RTM-) résultat obtenus avec les différentes fonctions de score utilisées. On constate que la prise en compte du contexte lors de l'entraînement des modèles offre de meilleures performances que RTM lorsqu'il s'agit de retrouver le mot à l'origine d'un lien dans les 10 mots jugés les plus probables. On note également que la version de Contextual-RTM uniforme au niveau de la phrase donne de meilleurs résultats en précision à 10 sur trois des cinq réseaux étudiés et la variante sémantique au niveau de la phrase (CRTM-Sem-S) sur deux, tout en étant la seule à se placer toujours devant RTM+. À noter que quelque soit le corpus, RTM obtient de meilleurs résultats en utilisant la fonction de lien ψ_{uni} avec une fenêtre de 0.

Corpus	**RTM**		**CRTM**			
	-	+	**Uni-0**	**Uni-S**	**Pos-S**	**Sem-S**
Physics	0.738	0.813	0.840	**0.842**	0.724	0.829
Arts	0.782	0.819	0.815	**0.833**	0.830	**0.833**
Society	0.787	0.818	0.858	**0.862**	0.844	0.822
Physique	0.796	0.820	0.833	0.826	0.837	**0.844**
Société	0.781	0.820	**0.833**	0.817	0.808	0.830

TAB. 2 : Précision moyenne sur les 10 mots les plus probables inférés par les modèles (en gras le meilleur résultat par corpus et souligné les résultats supérieurs à RTM).

Ne pouvant pas reporter les résultats pour toutes les valeurs de K dans cet article nous présentons à la Figure 2 (a) l'évolution de la précision moyenne de CRTM-Uni-S et RTM+, qui utilise pour l'évaluation la fonction de lien ψ_{uni} avec une fenêtre de 0. On remarque que les modèles arrivent tous les deux à retrouver la très grande majorité

de ces mots bien avant que le cinquantième mots soit atteint. Contextual-RTM semble néanmoins retrouver ces liens plus rapidement.



(a) Précision à K (b) Courbe ROC

FIG. 2 : (a) : Évolution des précisions moyennes à K en fonction de K entre CRTM-Uni-S et RTM sur le corpus Société. (b) : Courbes ROC moyennées entre CRTM-Uni-0 et RTM sur le corpus Society.

On présente dans le tableau 3 les aires sous la courbe ROC moyennée de chaque modèles pour les jeux de données étudiés, et à la Figure 2 (b), à titre d'exemple, les courbes ROC moyennes de RTM+ et CRTM-Uni-S.

Corpus	**RTM+**	**CRTM**			
		Uni-0	**Uni-S**	**Pos-S**	**Sem-S**
Physics	0.865	**0.874**	0.868	0.862	0.873
Arts	0.829	0.829	0.826	0.830	**0.832**
Society	0.802	**0.820**	0.811	0.812	0.817
Physique	0.832	**0.852**	0.844	0.842	0.850
Société	0.811	0.814	0.812	0.812	**0.822**

TAB. 3 : Aires sous les courbe ROC (en gras le meilleur résultat et souligné les résultats supérieurs à RTM).

On remarque que la prise en compte du contexte d'apparition des liens améliore systématiquement les performances sur la tâche. La version ne prenant en compte que les mots à l'origine des liens donne les meilleurs résultats sur trois des cinq corpus étudiés, et la version sémantique utilisant le produit scalaire normalisé sur les deux autres, tout en se plaçant seconde le reste du temps.

4.4.2 Influence de la taille de la fenêtre

Puisque Contextual-RTM comprend le contexte comme étant une fenêtre de l mots de part et d'autre du mot à l'origine d'un lien, nous étudions l'évolution de la précision à 10 pour les trois variantes de CRTM suivant neuf tailles de fenêtre, allant de zéro à l'intégralité du contenu d'un document. La Figure 3 présente l'évolution de la précision à 10 moyennée pour toutes les variantes de CRTM pour tous les corpus anglophones en fonction de la taille de la fenêtre. On remarque que les petites tailles de fenêtres tendent à donner de meilleurs résultats contrairement à un contexte de plus de 100 mots ou équivalent à un document entier. De plus on note que les meilleurs précision moyenne proviennent d'un contexte inférieur ou égale à une vingtaine de mots (une fenêtre de 10), ce qui correspond environ au nombre moyen de mots dans une phrase des corpus.

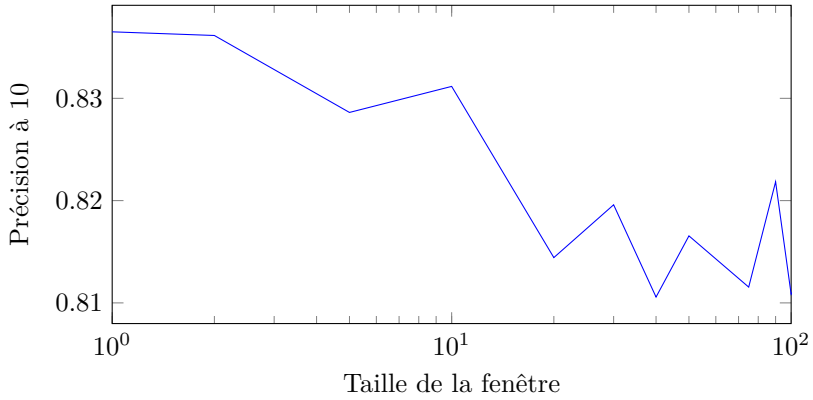


FIG. 3 : Précision à 10 moyennée pour tous les corpus anglophones pour les trois variantes de CRTM en fonction de la taille de la fenêtre (de 0, à 100 mots)

4.4.3 Qualitatif

Au delà de l'évaluation permettant de juger de la performance de ces modèles à retrouver l'origine d'un lien, nous désirons également juger de la cohérence des autres mots suggérés par les modèles.

Pour illustrer cela nous présentons à la Table 4 un exemple de parcours de pages Wikipedia. Le premier document est le paragraphe d'introduction du philosophe et mathématicien britannique Bertrand Russel. Le deuxième est celui du philosophe français Jean-Paul Sartre et le troisième celui de la page décrivant le concept de rationalité. Pour chacun des deux derniers document nous inférons avec RTM (qui utilise la fonction de lien ψ_{uni} avec une fenêtre de 0, qui donne le meilleurs résultats) et CRTM-Sem-S, la variante sémantique de CRTM au niveau des phrases, quels mots contribuent le plus aux liens entre ces documents et le premier document.

Concernant l'introduction de la page de Sartre on remarque que les deux modèles les identifient comme des *philosophes* ayant marqué la vie *intellectuelle*. Cependant

notre modèle met également en avant des liens plus ténus, comme des points de vues politiques similaires (*humanisme*) ou encore le fait qu'ils se soient vus tout deux proposer le prix Nobel de littérature, quand RTM propose des liens moins pertinents comme *prolifique* ou *idées*.

Dans le dernier document, parlant du concept de rationalité, on note une fois de plus que les deux modèles mettent en valeur la *philosophie* et le *rationalisme* comme lien entre les deux documents. RTM propose cependant d'autres termes moins pertinents comme *fonder* quand notre modèle retrouve des termes liés à Russel comme *science*, *logique*, ou *Descartes*.

Bertrand Arthur William Russell, né le 18 mai 1872 à Trellech (Monmouthshire, pays de Galles), et mort le 2 février 1970 près de Penrhyndeudraeth, au pays de Galles, est un mathématicien, logicien, philosophe, épistémologue, homme politique et moraliste britannique. [...]

Jean-Paul Sartre		**Rationalité**	
CRTM-pos- S	**RTM**	**CRTM-pos-S**	**RTM**
philosophe	philosophe	philosophie	philosophie
intellectuelle	Sartre	rationalisme	sociologie
existentialisme	intellectuelle	Descartes	humain
humanisme	philosophique	science	rationalisme
littéraire	prolifique	logique	fonder
Nobel	idées	connaissance	économique

Tab. 4 : Mots proposés par CRTM-pos-S et RTM pour les pages wikipedia relatives à J-P Sartre et à la Rationalité, après lecture de celle sur Bertrand Russell.

5 Conclusion et perspectives

Nous avons présenté dans cet article Contextual-RTM, un cadre de modélisation probabiliste de thématiques latentes pour les réseaux de documents, généralisant *Relational Topic Model*, et permettant de prendre en compte le contexte d'apparition des liens de différentes façons. Nous avons montré que les trois méthodes introduites étaient compétitives dans une tache d'identification de mots à l'origine de liens. Nous avons également montré que les mots suggérés avaient tendance à être plus plausibles et cohérents. Nous prévoyons dans nos travaux futurs d'étudier l'intérêt d'adopter une forme linéaire pour la fonction ψ afin de capturer des motifs thématiques plus complexes susceptibles de mieux expliquer les liens entre les documents.

Références

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Blei, D. M., A. Y. Ng, et M. I. Jordan (2003). Latent dirichlet allocation. *Journal of machine Learning research 3*(Jan), 993–1022.

Brochier, R., A. Guille, et J. Velcin (2019). Global vectors for node representations. In *The World Wide Web Conference*, pp. 2587–2593.

Brochier, R., A. Guille, et J. Velcin (2020). Inductive document network embedding with topic-word attention. In *European Conference on Information Retrieval*, pp. 326–340. Springer.

Chang, J. et D. Blei (2009). Relational topic models for document networks. In *Artificial Intelligence and Statistics*, pp. 81–88.

Gourru, A., A. Guille, J. Velcin, et J. Jacques (2020). Document network projection in pretrained word embedding space. In *European Conference on Information Retrieval*, pp. 150–157. Springer.

Jordan, M. I., Z. Ghahramani, T. S. Jaakkola, et L. K. Saul (1999). An introduction to variational methods for graphical models. *Machine learning 37*(2), 183–233.

Le, Q. et T. Mikolov (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pp. 1188–1196.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, et J. Dean (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119.

Tu, C., H. Liu, Z. Liu, et M. Sun (2017). Cane : Context-aware network embedding for relation modeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pp. 1722–1731.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, et I. Polosukhin (2017). Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008.

Yang, C., Z. Liu, D. Zhao, M. Sun, et E. Y. Chang (2015). Network representation learning with rich text information. In *IJCAI*, Volume 2015, pp. 2111–2117.

Zhang, F., N. J. Yuan, D. Lian, X. Xie, et W.-Y. Ma (2016). Collaborative knowledge base embedding for recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 353–362.

Summary

In the context of documents networks, we propose Contextual-RTM for documents contents and links between documents, by generalizing the Relation Topic Model. The originality of our approach is to model the location of a link into a document, i.e. taking in account the context where a link to another document occurs. Contextual-RTM is competitive with state of the art in retrieving link location on large networks of web documents. Consequently Contextual-RTM could assist readers making connections between documents.