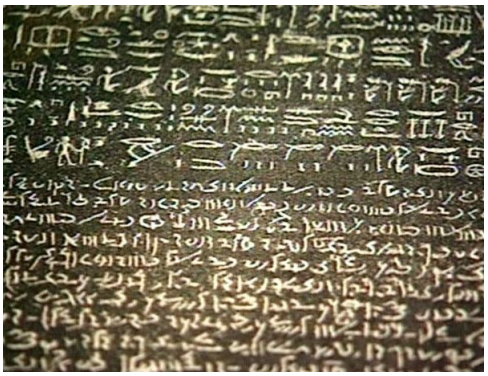
TextMine

Atelier sur la Fouille de Textes



Organisateurs :

Pascal Cuxac (INIST - CNRS),
Vincent Lemaire (Orange Labs)

Organisé conjointement à la conférence EGC
(Extraction et Gestion des Connaissances)
le 23 janvier 2018 à Paris

Editeurs :

Pascal Cuxac - INIST - CNRS
2 allée du Parc de Brabois, CS 10310, 54519 Vandoeuvre les Nancy Cedex
Email : pascal.cuxac@inist.fr

Vincent Lemaire - Orange Labs
2 avenue Pierre Marzin, 2300 Lannion
Email : vincent.lemaire@orange.com

Publisher:

Vincent Lemaire, Pascal Cuxac
2 avenue Pierre Marzin
22300 Lannion

Lannion, France, 2017

PRÉFACE

C'est une évidence que de dire que nous sommes entrés dans une ère ou la donnée textuelle sous toute ses formes submerge chacun de nous que ce soit dans son environnement personnel ou professionnel : l'augmentation croissante de documents nécessaires aux entreprises ou aux administrations, la profusion de données textuelles disponibles via Internet, le développement des données en libre accès (OpenData), les bibliothèques et archives en lignes, les medias sociaux ne sont que quelques exemples illustrant l'évolution de la notion de texte, sa diversité et sa prolifération.

Face à cela les méthodes automatiques de fouille de données (data mining), et plus spécifiquement celles de fouille de textes (text mining) sont devenues incontournables. Récemment, les méthodes de deep learning ont créées de nouvelles possibilités de recherche pour traiter des données massives et de grandes dimensions. Cependant, de nombreuses questions restent en suspens, par exemple en ce qui concerne la gestion de gros corpus textuels multi-thématiques. Pouvoir disposer d'outils d'analyse textuelle efficaces, capables de s'adapter à de gros volumes de données, souvent de nature hétérogène, rarement structurés, dans des langues variées, des domaines très spécialisés ou au contraire de l'ordre du langage naturel reste un challenge.

La fouille de textes couvre de multiples domaines comme, le traitement automatique des langues, l'intelligence artificielle, la linguistique, les statistiques, l'informatique et les applications sont très diversifiées, que ce soit la recherche d'information, le filtrage de spam, le marketing, la veille scientifique ou économique, la lutte antiterroriste...

Le but de cet atelier est de réunir des chercheurs sur la thématique large de la fouille de textes. Cet atelier vise à offrir une occasion de rencontres pour les universitaires et les industriels, appartenant aux différentes communautés de l'intelligence artificielle, l'apprentissage automatique, le traitement automatique des langues, pour discuter des méthodes de fouille de texte au sens large et de leurs applications.

P. Cuxac — INIST-CNRS
V. Lemaire — Orange Labs

Membres du comité de lecture

Le Comité de Lecture est constitué de:

Guillaume Cabanac (IRIT, Toulouse)

Martine Cadot (Loria, Nancy)

Mariane Clausel (Université de Lorraine, Nancy)

Vincent Claveau (IRISA, Rennes)

Guillaume Cleuziou (LIFO, Orléans)

Gael Dias (Université de Normandie, Caen)

Dominique Gay (U. Réunion, Saint Denis de la Réunion)

Natalia Grabar (STL - Lille3, Lille)

Mustapha Lebbah (LIPN, Paris)

Denis Maurel (Université F. Rabelais, Tours)

Patrick Paroubeck (LIMSI, Orsay)

David Reymond (Université du Sud, Toulon - Nice)

Isabelle Tellier (Sorbonne, Paris)

Julien Velcin (Université de Lyon, Lyon)

PROGRAMME

- 14h00 : Ouverture de l'atelier
- 14h05 : Exposé Invité : "Analyse exploratoire de données textuelles à l'aide de modèles thématiques probabilistes" - Julien VELCIN
- 15h00 : "Réseau neuronal convolutif sur des séquences de caractères pour la classification de textes" - Idriss El Asry, Wissam Siblini, Frank Meyer
- 15h30 : "Fouille de publications scientifiques pour une analyse bibliométrique de l'activité de recherche sur la déforestation" - Nathalie Neptune, Josiane Mothe, Julius Akinyemi
- 16h00 : pause-café + posters
- 16h30 : "Constitution d'un corpus d'articles scientifiques avec représentation sémantique" - Jean-Claude Moissinac
- 17h00 : "L'évaluation des représentations vectorielles de mots en utilisant WordNet" - Nourredine Aliane, Jean-Jacques Mariage, Gilles Bernard
- 17h30 : "Graph2Bots: Assistance automatisée à la conception d?agents dialoguants" - Jean Leon Bouraoui, Vincent Lemaire
- 18h00 : Clôture de l'atelier

TABLE DES MATIÈRES

Exposé Invité

Session Exposés

Analyse exploratoire de données textuelles à l'aide de modèles thématiques probabilistes

Julien Velcin

Laboratoire ERIC
Université de Lyon
julien.velcin@univ-lyon2.fr

Résumé. L'analyse exploratoire de vastes corpus textuels nécessite le recours à des techniques d'apprentissage non ou peu supervisés : techniques de projection en faible dimension, (co-)clustering de documents, modélisation thématique. Dans cet exposé, après un bref panorama des différentes techniques à la disposition du scientifique des données, je détaillerai davantage les modèles thématiques probabilistes en cherchant à illustrer l'étendue des utilisations possibles à l'aide de cas concrets. Ces modèles ont par exemple été utilisés pour améliorer la recherche d'information, pour analyser l'opinion ou pour détecter la nouveauté dans les flux de données. Je terminerai en donnant quelques travaux récents qui cherchent à combiner ces modèles avec des techniques de plongement de mots.

Réseau neuronal convolutif sur des séquences de caractères pour la classification de textes

Idriss El Asry*, Wissam Siblini*,**
Frank Meyer*

*Orange Labs
2 Avenue Pierre Marzin, 22300 Lannion
prénom.nom@orange.fr
**Laboratoire des Sciences du Numérique de Nantes (LS2N)
Rue Christian Pauc, 44300 Nantes
prénom.nom@univ-nantes.fr

Résumé. Avec la croissance des volumes de données textuelles et l'explosion des applications associées, la classification de textes est devenue un enjeu majeur dans le domaine du traitement automatique des langues, de l'apprentissage automatique et de nombreuses methodes ont été proposées. D'une part, les stratégies standard et éprouvées consistent à construire un dictionnaire pour encoder les textes comme des sac-de-mots, à y appliquer des pondérations, et enfin à apprendre un classifieur efficace sur les textes encodés. D'autre part, les approches proposées récemment utilisent des réseaux de neurones convolutifs directement sur les textes bruts représentés comme des séquences de caractères. Le but de cet article est de comparer ces deux stratégies. Dans cette étude prémilinaire sur des données de *chats* internes (Orange Labs), il apparait que leurs performances de classification sont équivalentes. En revanche, l'approche réseau neuronal convolutif se distingue en permettant une meilleure automatisation de la tache par suppression de la phase de construction du dictionnaire de mots.

1 Introduction

Avec l'avènement du numérique et l'évolution des technologies d'acquisition, d'archivage et de partage de documents, les volumes des données textuelles mises à disposition ne cessent de croitre (Partalas et al., 2015). L'un des enjeux majeurs est la classification automatique de ces données par analyse de leur contenu. Avec de nombreuses applications potentielles, de plus en plus de recherches se concentrent sur la discipline de la classification (en particulier multi-label) de textes (Tsoumakas et Katakis, 2006). Il s'agit d'un paradigme d'apprentissage dans lequel un modèle est ajusté pour associer des textes à une ou plusieurs classes prédéfinies (par exemple *drôle*, *traite de sport*, *exprime un avis positif*).

Au cours des dernières décennies, de nombreux modèles de classification multi-label de textes ont été proposés et jusqu'à récemment, la plupart d'entre eux consistaient à optionnellement effectuer un ensemble de prétraitements linguistiques (*lemmatisation*, *stematisation*,

normalisation) sur le texte puis à l'encoder (construction d'un dictionnaire et codage *bag-of-words*) et enfin à utiliser un classifieur efficace (par exemple *SVM* (Vapnik, 1998)) pour l'associer aux classes (Aggarwal et Zhai, 2012). Ces dernières années, avec l'explosion des méthodes basées sur des réseaux de neurones et l'amélioration des moyens de calculs (en particulier la parallélisation sur GPU), la recherche s'oriente vers encore plus d'automatisation dans la classification de textes (Zhang et al., 2015). En effet, des méthodes proposées récemment consistent à considérer le texte brut comme une séquence de caractères et à effectuer la classification sur cette séquence avec des technologies neuronales adéquates comme les réseaux de neurones convolutifs (Zhang et al., 2015) ou les Long Short Term Memory (LSTM) (Hochreiter et Schmidhuber, 1997). Dans un tel scénario, le modèle doit capturer automatiquement toute la syntaxe, liée à des suites de caractères (les radicaux, les mots, les suites de mots), qui lui sera utile et en extraire du sens. Pour estimer les capacités de ces nouvelles approches très automatisées, on compare, dans cet article, l'une d'entre elles (réseau neuronal convolutif sur suite de caractères) avec une méthode standard reconnue pour son efficacité (encodage TF et SVM). La comparaison des performances est effectuée sur des données internes (Orange Labs) de conversation *chat* [1]. Les résultats de cette étude préliminaire montrent que les deux approches ont des performances équivalentes. Dans la suite, une étude plus extensive (avec plus de jeux de données) sera nécessaire pour confirmer ces conclusions.

Cet article est organisé de la façon suivante. Nous rappelons d'abord les familles de méthodes standard en classification de textes et décrivons les deux méthodes sélectionnées pour notre étude. Puis, nous présentons le protocole des expériences réalisées (paramétrage, mesures utilisées, données). Enfin, nous présentons et discutons les résultats avant de conclure.

2 Approches pour la classification de textes

La classification de textes a toujours été un point d'intérêt important dans le domaine du traitement automatique des langues, de l'extraction d'information et de l'apprentissage automatique. Identiquement aux autres problèmes de Machine Learning, cette tâche consiste en une phase de préparation des données et une phase d'apprentissage.

La phase de préparation, pour l'encodage des textes, a été largement étudiée ces dernières années. Aujourd'hui, il existe de nombreuses représentations possibles, systématiques ou apprises (Le et Mikolov, 2014). Parmi les plus répandues, on peut citer par exemple les représentations systématiques bag-of-words (voir partie 2.1), séquence de caractères (voir partie 2.2), ou des représentations apprises comme word2vec (Mikolov et al., 2013) et GloVe (Pennington et al., 2014).

Pour la phase de classification, depuis les premiers travaux datant de la deuxième moitié du 20ème siècle, de nombreux classifieurs appartenant à plusieurs familles (modèles de mélanges, arbres de classification, réseaux de neurones, SVM, classifieurs bayésiens) ont été proposés (Aggarwal et Zhai, 2012). Parmi ces familles d'approches, deux se distinguent particulièrement. La première est la classification SVM appliquée sur des représentations de type sac-de-mots. Elle est très utilisée, présente des propriétés théoriques intéressantes et obtient souvent des performances "état de l'art". La seconde utilise l'apprentissage profond et émerge depuis quelques années grâce au développement de nouvelles architectures et de nouveaux

1. conversation en ligne

moyens de calculs (GPU). Elle présente de très bonnes performances. Dans cette partie, on se concentre sur deux approches (TF et SVM et réseau neuronal convolutif au niveau caractère) qui appartiennent respectivement aux deux familles.

2.1 Approche TF et SVM

Le classifieur SVM appliqué sur un encodage TF des documents est une méthode très répandue pour la classification de textes. Dans cette partie, on présente d'abord l'encodage TF puis le classifieur SVM.

BOW et TF. L'une des méthodes de base pour l'encodage vectoriel de textes consiste en la construction d'un dictionnaire et en la représentation des documents selon la présence/absence, dans leur contenu, des mots du dictionnaire. Cette représentation est appelée sac-de-mots (en anglais : bag-of-words). Un document est donc représenté par un vecteur de la taille du dictionnaire, dont la i-ème composante indique par 1 (resp. 0) la présence (resp. l'absence) dans le document de l'i-ème mot du dictionnaire. Pour un grand corpus, le texte encodé a généralement une très grande dimension, mais il est aussi souvent creux car il ne contient qu'une petite portion des termes possibles.

Cette représentation vectorielle de base a connu de nombreuses évolutions, parmi lesquelles nous pouvons mentionner la pondération de type TF (Term Frequency) telle que la i-ème composante du vecteur représente la fréquence de l'i-ème mots dans le texte (Aggarwal et Zhai, 2012).

SVM. Pour chaque texte encodé, l'objectif est de déterminer les classes qui lui sont associées. Ici, on va plutôt considérer le problème suivant : Pour chaque classe, on souhaite déterminer les textes qui lui sont associés. On passe alors d'un problème de classification multi-label à plusieurs problèmes de classification mono-label que l'on peut traiter avec le classifieur SVM. Cette transformation du problème est standard (Tsoumakas et Katakis, 2006).

Pour une classe donnée, le classifieur SVM, introduit par Vladimir Vapnik dans les années 1990, permet de déterminer l'hyperplan séparateur à marge optimale, dans l'espace des textes encodés (optionnellement transformé avec des noyaux [2]), entre les instances qui sont associées à la classe et celles qui ne le sont pas. Pour trouver cet hyperplan, le classifieur SVM résout le problème d'optimisation suivant sur les données d'apprentissage :

$$\begin{aligned} &\min_{w,b} \|w\|_2 \\ &\text{avec } \forall i \in \{1, ..., N\}, \quad y_i(w^T x_i + b) > 1 \end{aligned} \tag{1}$$

où y_i vaut 1 (resp. 0) si l'i-ème texte appartient (resp. n'appartient pas) à la classe donnée, x_i est le vecteur TF de l'i-ème texte, (w, b) sont les caractéristiques de l'hyperplan et N est le nombre de textes dans la base d'apprentissage.

2. Dans notre étude, nous utilisons un noyau linéaire.

2.2 Approche réseau neuronal convolutif

Les réseaux de neurones convolutifs permettent d'effectuer la classification directement sur les suites de caractères qui constituent les textes.

Séquence de caractères. Pour avoir une séquence de caractères interprétable par le réseau neuronal convolutif, on commence par indexer les caractères possibles. Dans cet article, on considère l'alphabet (a, b, c, ...), le point et l'espace. Nous avons donc un ensemble de caractères de taille $d = 28$. Chaque caractère j de cet ensemble a un encodage one-hot qui correspond à un vecteur de taille d pour lequel la j-ème composante vaut 1 et les autres sont nulles. Les textes sont représentés comme des matrices $X \in \mathbb{R}^{d \times l}$ telles que pour tout $i \in \{1, ..., l\}$, la i-ème ligne de X est l'encodage one-hot de l'i-ème caractère du texte. l est la taille de la séquence qui forme le plus grand texte de la base utilisée ($l = 712$ dans notre cas). Si le texte contient $l' < l$ caractères, les caractères l', $l' + 1$, ..., l du texte sont considérés comme des espaces.

Réseau neuronal convolutif (ConvNet). Le réseau neuronal utilisé est un empilement de couches de transformations de plusieurs natures appliquées à la matrice X (voir figure 1) :

- Couches de convolutions (CONV) qui appliquent des filtres sur les données ;
- Couches de pooling (POOL), qui permettent de compresser l'information par sous-échantillonnage (par exemple sélection du maximum) ou par agrégation (par exemple calcul de la moyenne) ;
- Fonctions d'activation (σ) qui effectuent des transformations de "seuillage" (sigmoid, ReLU) ;
- Couches "entièrement connectées" (Fully Connected) qui sont souvent des perceptrons et qui effectuent finalement la classification.

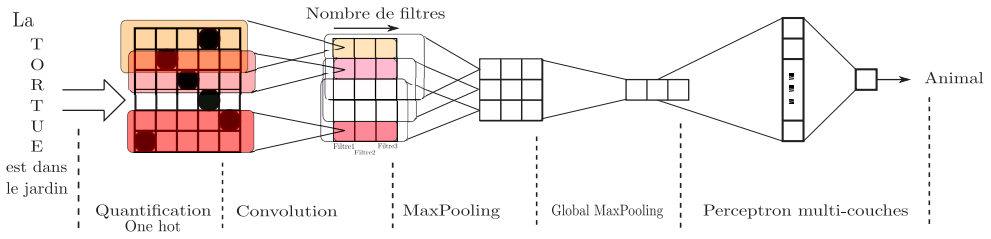


FIG. 1 – *Principe de fonctionnement du réseau neuronal convolutif sur une séquence de caractères.*

Sur les données d'apprentissage, le réseau apprend automatiquement les valeurs des filtres et les poids du perceptron de sorte à optimiser les performances en classification.

Les rôles des différentes couches n'est pas le même dans la classification globale. Les couches de convolutions et de pooling permettent d'apprendre des représentations intéressantes des séquences (identifications de suites pertinentes comme la présence de radicaux, de mots ou d'expressions utiles à la classification) et le perceptron utilise cette représentation pour déterminer les classes associées au texte.

3 Expériences et résultats

Pour comparer les deux approches candidates, des tests sont réalisés sur un corpus réel de textes catégorisés. Dans cette partie, on décrit d'abord précisément le protocole des expériences (données, prétraitements, paramètres des algorithmes), les mesures utilisées pour la comparaison et les résultats obtenus.

3.1 Protocole

Données Des données internes (Orange Labs) ont été utilisées pour les expériences. Il s'agit d'une base comportant 2306 *chats* clients sur les problèmes rencontrés en utilisant le matériel d'Orange. Chaque *chat* est associé à une ou plusieurs catégories parmi dix possibilités qui définissent la raison de la discussion *chat* ("demande d'échange","problème écran noir","problème droit tv", "problème identifiant", etc...). Par exemple, la réclamation "bonjour je ne recois plus la tv orange depuis une coupure d'éléctricité" est associée au label "problème écran noir".

Les textes ont été mis en minuscule, les signes diacritiques (accents) et les mots d'une lettre ont été supprimés.

Mesures et validation croisée Pour mieux estimer le comportement du modèle nous utilisons la validation croisée 5-folds (80% train, 20% test). Cette technique consiste à subdiviser la base de données en cinq parties, puis à effectuer cinq tests en utilisant à chaque fois quatre parties pour l'apprentissage et une partie pour le test.

Les performances des algorithmes ont été évaluées par accuracy sur l'ensemble de test. Cette mesure correspond à la proportion d'échantillons correctement classés (vrais positifs et vrais négatifs).

Paramétrage des modèles L'algorithme de SVM utilisé est celui de la librairie scikit learn (python). Le réseau neuronal convolutif est implémenté avec Keras et Tensorflow-gpu (python). Son architecture, inspirée par la proposition de Zhang et al. (2015), est présentée en détail dans le tableau 1.

Couches	Taille de la sortie
Couche d'entrée	712
Couche Lambda (quantification)	712×28
Couche de convolution	706×4096
Couche de pooling	235×4096
Couche de MaxPooling	4096
Couche cachée	128
Couche de sortie	10

TAB. 1 – *Architecture du réseau neuronal convolutif utilisé.*

Le modèle applique une convolution avec 4096 filtres de taille 7×28 (706 correspond au nombre de positions que peut prendre une fenêtre de taille 7 dans une séquence de taille 712), puis un pooling-maxpooling pour avoir une seule valeur pour chaque filtre. Le vecteur de taille

4096 obtenu est traité par un perceptron avec 128 neurones cachés qui réalise la classification sur les 10 catégories. La couche "lambda" correspond à l'encodage des textes comme des matrices (voir détails dans la section précédente). Le modèle minimise la fonction de perte "binary crossentropy" et apprend avec la méthode numérique "ADAM" (Kingma et Ba, 2014) et avec une procédure de "early stopping" pour éviter le sur-apprentissage. Des mini-batchs de taille 32 sont utilisés pour éviter les variations brusques des poids du modèle et diminuer le temps d'apprentissage.

3.2 Résultats

Les résultats des deux approches (TF-SVM et ConvNet), comparées à un classifieur bayésien naïf utilisé comme baseline, sont présentés dans la figure 2.

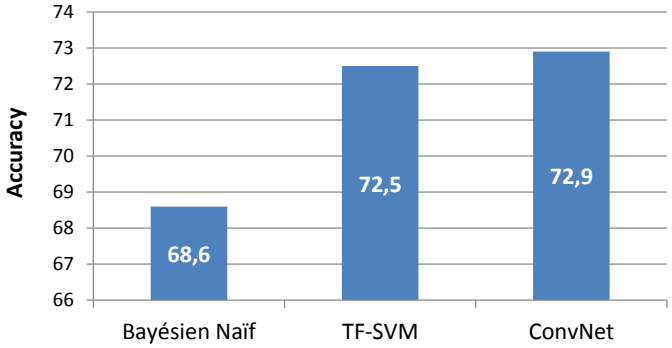


FIG. 2 – *Accuracy globale des différents modèles sur les données de chat.*

Les performances du SVM et du réseau neuronal convolutif sont proches et meilleures que celle de la baseline.

4 Conclusion

Dans cet article nous avons effectué une comparaison préliminaire entre une approche standard de classification de textes (TF et SVM) avec une approche récente automatisée (ConvNet sur séquences de caractères). Sur des données internes de *chat* (Orange), les performances des deux approches sont équivalentes. Pour le confirmer, il serait néanmoins nécessaire d'approfondir l'analyse sur plus de jeux de données (publics). De plus il serait intéressant d'ajouter des traitements linguistiques (par exemple la lemmatisation ou l'utilisation d'attributs latents) pour rendre plus efficace le SVM. Si l'équivalence est confirmée sur une étude plus extensive, la seconde stratégie est préconisée car elle a des avantages pratiques considérables. D'abord, comme il est aujourd'hui le cas pour l'image, le réseau de neurones convolutif devrait permettre de s'affranchir de la nécessité de prétraitements « expert » sur les textes et d'être robuste à la spécificité d'un langage. Ensuite, ce modèle a les moyens d'être résistant aux fautes

de frappes/orthographes car, dans une représentation séquences de caractères, un mot mal orthographié résulte seulement sur une légère modification de la séquence tandis que dans une représentation sac-de-mots, un mot mal orthographié est un mot différent. Enfin, il est très rapide car le nombre de variables est petit (un ensemble de caractères a généralement une dimension beaucoup plus faible qu'un dictionnaire de mots). Pour toutes ces raisons, les approches d'apprentissage profond sur les textes encodés comme des séquences de caractères ont des perspectives très prometteuses pour l'avenir de la classification de textes. Dans les travaux futurs, nous prévoyons d'approfondir notre comparaison pour expliciter les atouts des méthodes standards et des méthodes à base de réseaux de neurones convolutifs.

Références

Aggarwal, C. C. et C. Zhai (2012). A survey of text classification algorithms. *Mining text data*, 163–222.

Hochreiter, S. et J. Schmidhuber (1997). Long short-term memory. *Neural computation 9*(8), 1735–1780.

Kingma, D. et J. Ba (2014). Adam : A method for stochastic optimization. *arXiv preprint arXiv :1412.6980*.

Le, Q. et T. Mikolov (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 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.

Partalas, I., A. Kosmopoulos, N. Baskiotis, T. Artieres, G. Paliouras, E. Gaussier, I. Androutsopoulos, M.-R. Amini, et P. Galinari (2015). Lshtc : A benchmark for large-scale text classification. *arXiv preprint arXiv :1503.08581*.

Pennington, J., R. Socher, et C. Manning (2014). Glove : Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Tsoumakas, G. et I. Katakis (2006). Multi-label classification : An overview. *International Journal of Data Warehousing and Mining 3*(3).

Vapnik, V. N. (1998). *Statistical learning theory*, Volume 1. Wiley New York.

Zhang, X., J. Zhao, et Y. LeCun (2015). Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pp. 649–657.

Summary

With the increasing amount of textual data and the explosion of associated applications, text classification has become a major issue in the field of machine learning and many methods have been proposed. On the one hand, standard and proven strategies consist in building a dictionary to encode texts as bag-of-words and applying a standard classifier on the encoded

texts. On the other hand, recently proposed approaches use convolutional neural networks directly on raw texts represented as character sequences. The purpose of this article is to compare these two strategies. On this preliminary study on *chat* data (Orange Company), it appears that their classification performances are equivalent. However, the convolutional neural network is distinguished because it allows a better automation of the task by removing the dictionary construction phase.

Fouille de publications scientifiques pour une analyse bibliométrique de l'activité de recherche sur la déforestation

Julius Akinyemi*, Josiane Mothe**, Nathalie Neptune**

*Entrepreneur-In-Residence, MIT Media Lab/Founder-CEO UWINCorp Inc.
akinyemi@media.mit.edu
**Institut de Recherche en Informatique de Toulouse
Université de Toulouse
118 Route de Narbonne, F-31062 Toulouse Cedex 9
Josiane.Mothe@irit.fr
Nathalie.Neptune@irit.fr

Résumé. La déforestation est un phénomène très répandu qui touche des portions de territoires assez importantes surtout dans les régions tropicales. La télédétection permet aux chercheurs de suivre et d'analyser l'évolution spatio-temporelle de ce phénomène.

En utilisant la fouille de texte et de méta-données sur les publications scientifiques sur le thème de la déforestation, nous visons à identifier les lieux de la production scientifique sur la déforestation et les collaborations entre chercheurs. L'analyse de ces collaborations nous permet de voir les tendances de la distribution de la production parmi les auteurs, à savoir si elle est concentrée au niveau des auteurs particuliers des pays développés ou bien si elle tend à être répartie de manière équilibrée entre plusieurs pays développés et émergents.

Nous nous appuyons pour cela sur des analyses de réseaux. Par ailleurs, grâce à l'analyse des mots-clés nous identifions les sites touchés par la déforestation auxquels les chercheurs s'intéressent, les forêts tropicales et l'Amazonie, de même que des sujets connexes ayant rapport à l'environnement et à la santé.

1 Introduction

La déforestation est un phénomène environnemental qui peut selon Foley et al. (2005) avoir un impact négatif sur l'écosystème de la terre. Dès 1992, Diegues et al. (1992) passant en revue les liens entre les processus qui conduisent à la déforestation ainsi que ces conséquences dans le basin Amazonien du Brésil, avait estimé que le taux de déforestation était élevé et augmentait rapidement et dangereusement.

La fouille de texte sur les publications scientifiques produites en rapport à la déforestation permet de quantifier les informations sur ces publications ainsi que leur

contenu, Pritchard (1969). Nous avons réalisé une analyse sur ces textes pour voir l'évolution géographique et temporelle de la recherche scientifique sur la déforestation. Cette étude permet d'identifier les principaux acteurs et leur localisation. Par ailleurs, les sujets sur lesquels ils effectuent leurs travaux sont aussi mis en lumière.

Les collaborations entre chercheurs est un aspect important de la recherche. De plus, les travaux de recherche sur la déforestation, de part la nature du phénomène, font souvent appel à des expertises dans des disciplines diverses. Il en découle qu'une analyse des collaborations dans les publications sur la déforestation peut faire ressortir le caractère multidisciplinaire de ces recherches.

Les résultats des analyses effectuées pourront guider des travaux futurs explorant le lien spatio-temporel entre la déforestation et les autres phénomènes liés.

Dans cet article, la méthodologie utilisée pour effectuer les analyses sera d'abord présentée puis les résultats obtenus seront présentés en détails et interprétés. Finalement, des perspectives pour la suite des travaux seront proposées suivies de la conclusion de l'article.

2 Méthodologie

Les données sur les publications ont été collectées à partir du Web of Science Core Collection(`https://webofknowledge.com`). La recherche a été effectuée par sujet en utilisant le terme "deforest*'. Toutes les publications datées de 1975 à 2016 ont été collectées le 23 octobre 2017. Cette approche permet de faire un premier travail de prise de connaissance du domaine. Des requêtes plus sophistiquées auraient pu être utilisées notamment pour retrouver les publications qui ne mentionnent pas explicitement la déforestation tout en y étant liées. Les données ont été analysées avec Tetralogie [1], une plateforme de veille scientifique et technologique qui a été développée à l'Institut de Recherche en Informatique de Toulouse. Pour une description détaillée voir Dousset (2009).

Les fouilles de textes et d'information que nous avons réalisées ont eu pour objectif de répondre aux questions suivantes :

1. Comment le nombre de publications a-t-il évolué avec le temps ?
2. Quels sont les pays au centre de la recherche sur la déforestation ?
3. Quels sont les auteurs qui publient le plus sur le sujet ?
4. Quelle est l'importance et la nature des collaborations entre les chercheurs des différents pays ?
5. À quelles disciplines scientifiques appartiennent les auteurs ?

Des analyses mono et bi-variées ont été réalisées ; les résultats obtenus ont été utilisés pour produire des diagrammes de dispersions, des réseaux de co-auteurs et de catégories de recherche. Finalement, nous avons croisé des données de l'Organisation des Nations Unis pour l'alimentation et l'agriculture et de la Banque Mondiale afin de voir la productivité des pays qui publient le plus par rapport au nombre d'habitants

1. Tetralogie est un logiciel de veille technologique `http://atlas.irit.fr/` utilisée en recherche et en enseignement et pouvant être utilisée à distance contractuellement.

et au produit intérieur brut, pour l'année 2016. Ce croisement met en perspective la production scientifique sur le thème de la déforestation pour un pays par rapport à la taille de sa population et par rapport à la taille de son économie, pour l'année 2016. Ce croisement permet d'avoir une idée de l'effort humain et financier que représente, pour chaque pays leur contribution à la production scientifique sur la déforestation. Les pays des auteurs ont été extraits à partir de l'adresse présente dans les données fournies par le Web of Science. Les différentes orthographes des noms des pays sont prises en compte. Pour certaines publications, l'adresse de l'auteur peut être manquante.

Un total de 16136 publications ont été collectées avec 31772 auteurs dans 149 pays et territoires.

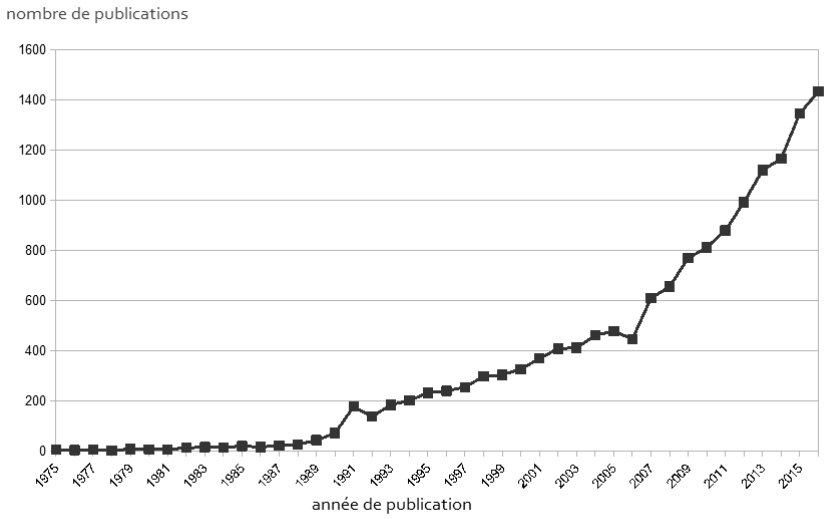


FIG. 1 – *Évolution du nombre de publications avec le temps.*

Le nombre de publications présentes dans la collection est très faible pour les premières années avec moins d'une dizaine de publications par année jusqu'en 1982 qui compte 12 publications. À partir du début des années 90 une augmentation remarquable est constatée dans la production annuelle. Cette tendance à la hausse continue jusqu'en 2016 (voir la figure 1). Cette tendance est similaire à l'évolution constatée sur d'autres domaines notamment les sciences naturelles et les sciences de la santé, selon Bornmann et Mutz (2015). Sur la période des 10 dernières années, de 2006 à 2016, le nombre de publications a augmenté de 220%.

3 Étude relative aux pays

Dans cette section, une analyse des contributions par pays est effectuée puis le réseau de collaboration formé par les pays est présenté. Enfin, le nombre de publications pour chaque pays par rapport à leur produite intérieur brut et à leur population pour l'année 2016 est présenté.

3.1 Contributions par pays et collaborations inter-pays

En regardant les données par pays il en ressort que parmi les dix pays comptant le plus de publications sur la déforestation, sur la période 1996-2016, trois sont des pays émergents et deux sont de l'Amérique latine : le Brésil, la Chine et le Mexique. Voir la figure 2. Il s'agit d'une tendance atypique qui n'est pas constatée sur d'autres domaines. En effet, pour les publications sur les géosciences, uniquement deux pays émergents (Chine et Inde), dont aucun de l'Amérique latine, figurent dans les dix premiers en nombre de publications, voir la figure 3.

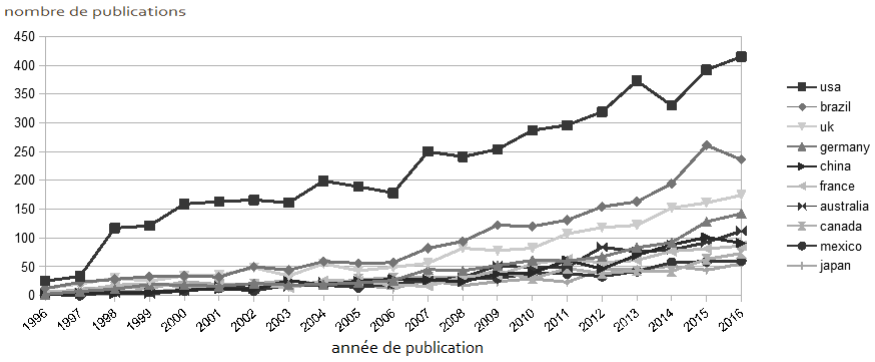


FIG. 2 – *Évolution du nombre de publications par pays avec le temps, pour les dix pays ayant le plus de publications pour toute la période 1996-2016. Les années 1975-1995 ne sont pas représentées compte tenu du faible nombre de publications pour ces années.*

L'évolution de la production de chaque pays peut aussi être évaluée par année par rapport à la production totale du pays. C'est ce que montre la figure 4 pour les 30 dernières années, de 1996 à 2016. Chaque point représente un pourcentage qui est calculé en divisant le nombre de publications du pays pour l'année par la somme des publications du pays pour toutes les années de 1975 à 2016. À partir de 1998, les États-Unis et le Canada cheminent en tête et maintiennent une augmentation de production par rapport à chaque année précédente. Toutefois, cette tendance change à partir de 2008. De 2012 à 2016, les pays ayant le plus augmenté leur production sont l'Allemagne, l'Australie, la Chine et le Brésil.

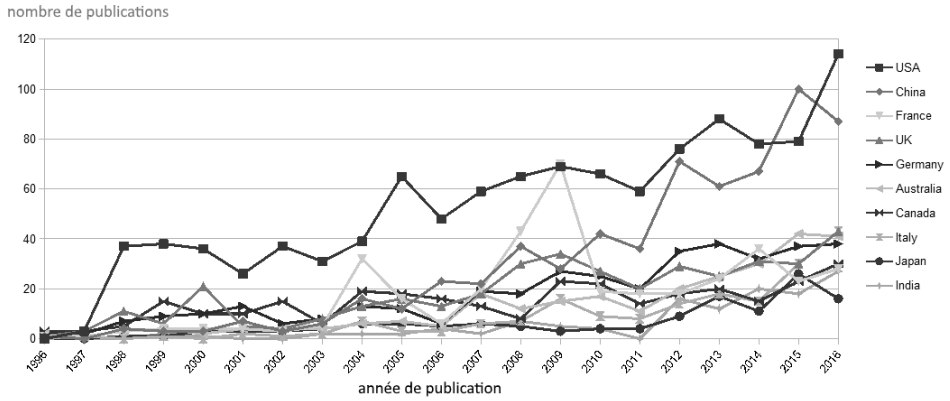


FIG. 3 – *Évolution du nombre de publications contenant le mot clé "geosciences" par pays, pour les dix pays ayant le plus de publications pour toute la période 1996-2016. Un total de* 4641 *publications ont été collectées avec* 13699 *auteurs dans* 87 *pays et territoires.*

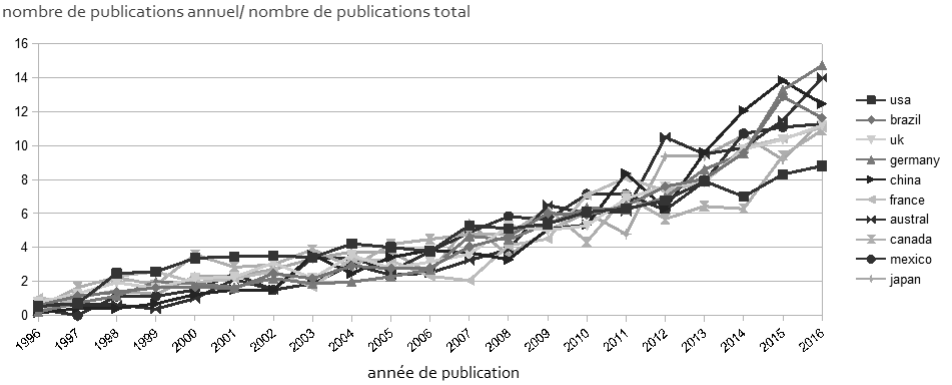


FIG. 4 – *Rapport entre le nombre de publications pour chaque pays sur le nombre total de publications pour le pays, par année, de 1996 à 2016 en pourcentage.*

3.2 Réseau de collaboration entre les pays

Le réseau de collaboration entre les pays fait ressortir un fort niveau de collaboration entre les auteurs issus des différents pays (voir figure 5). Chaque nœud représente un pays et l'intensité des liens entre eux représente le nombre de publications co-écrites par des auteurs de différents pays. On peut observer que tous les pays se retrouvent dans un grand groupe centré principalement autour des États-Unis.

La figure 6 permet de voir les pays ayant le plus de collaborations avec les États-Unis. La taille de chaque nœud représente le nombre de publications du pays. Le Brésil

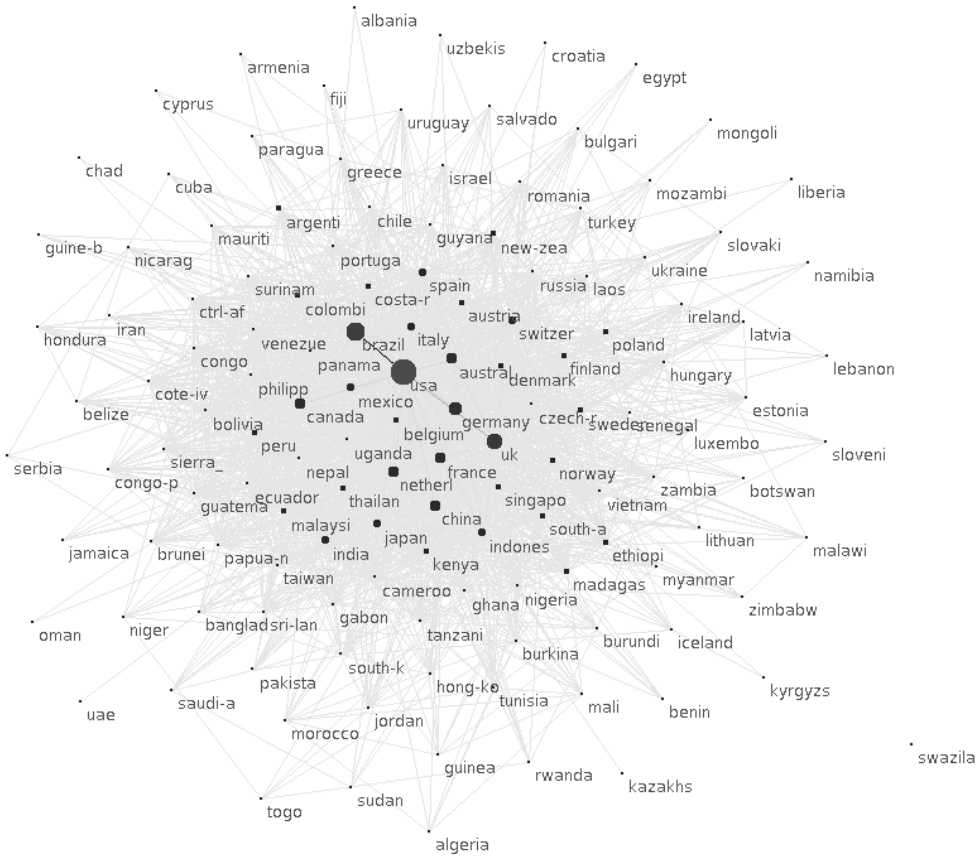


FIG. 5 – *Réseau formé par les pays publiant sur la déforestation pour les années 1975 à 2016.*

se révèle le pays ayant le plus souvent collaboré avec les États-Unis. Suivent ensuite le Royaume-Uni et l'Allemagne. Il est possible que la position très élevée de ces derniers dans le classement des pays publiant le plus sur la déforestation soit en partie due au fait de leur très forte collaboration avec des chercheurs des États-Unis. En effet Malhado et al. (2014) a démontré que la proportion d'articles par des auteurs de la région amazonienne sur l'Amazonie, particulièrement Brésiliens, a augmenté avec le temps mais que par ailleurs la proportion d'articles sur l'Amazonie n'impliquant pas d'auteurs de la région a également augmentée. Il peut s'avérer que même en augmentant considérablement leur participation à la production scientifique sur la déforestation, les Brésiliens n'arrivent pas forcément à en prendre le leadership.

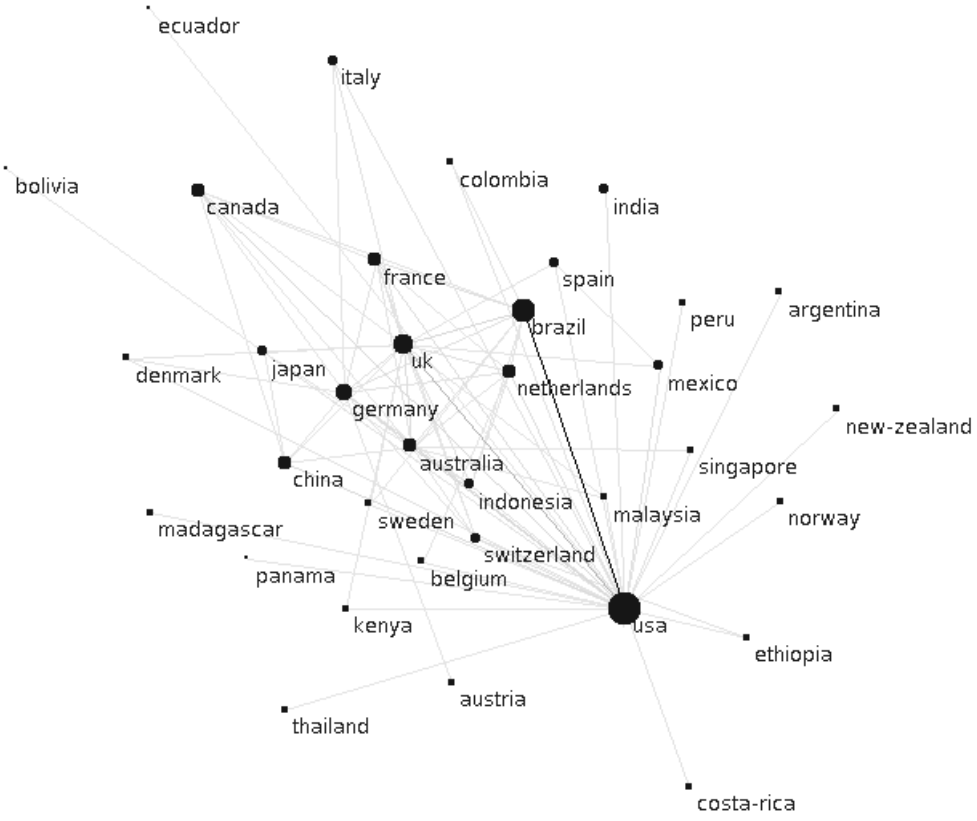


FIG. 6 – *Réseau de collaboration autour des États-Unis. Pour la période allant de 1975 à 2016.*

3.3 Nombre de publications par rapport au produit intérieur brut et la population en 2016

Le nombre de publications par habitant et par le produit intérieur brut exprimé en milliards de dollars américains a été calculé en utilisant les données fournies par la Banque Mondiale [2], pour l'année 2016.

	publi.	pop.	publi./pop	PIB'	publi./PIB'
brazil	236	207,65	1,14	1800	0,13
australia	112	24,12	4,64	1200	0,09
netherlands	72	17,01	4,23	771	0,09
uk	174	65,63	2,65	2620	0,07
canada	73	36,28	2,01	1530	0,05
germany	142	82,66	1,72	3470	0,04
france	86	66,89	1,29	2470	0,03
india	65	1324,17	0,05	2260	0, 03
usa	415	323,12	1,28	18600	0,02
china	91	1378,66	0,07	11200	0,01

TAB. 1 – *Nombre de publications pour chacun des 10 pays ayant le plus de publications par rapport à la population et au PIB (en milliards de dollars US) pour 2016. La première colonne "publi." représente le nombre de publications pour l'année 2016. La deuxième colonne "pop." représente la population en millions d'habitants. La troisième colonne "publi./pop" représente le ratio entre le nombre de publications et la population exprimée en million. La quatrième colonne "PIB'" représente le produit intérieur brut en milliards de dollars américains. La cinquième colonne "publi./PIB'" représente le nombre de publications divisé par le produit intérieur brut en milliards.*

Dans le tableau 1, nous voyons que l'Australie et les Pays-Bas ressortent comme étant les pays produisant le plus de publications par habitant. Dans le classement du nombre de publications en fonction du PIB, le Brésil arrive en tête suivi du duo Australie et Pays-Bas. Ces derniers voient peut-être leur leadership dans la recherche sur la déforestation limité par la taille de leur économie à l'échelle mondiale.

4 Sujets d'étude des publications

Les résumés des articles fournissent également des informations sur les pays, régions ou territoires auxquels les auteurs se sont les plus intéressés. Le tableau 2 fait ressortir ceux qui ont été le plus souvent mentionnés. L'Amazonie et le Brésil arrivent en tête, ce qui est un résultat attendu vu le nombre important de contributions Brésiliennes et le fait que la forêt Amazonienne est la plus importante au Brésil. Bien que les autres pays et régions soient moins souvent mentionnés, il est intéressant de constater que la quasi totalité des continents figure dans cette liste à l'exception de l'Océanie.

2. `https://data.worldbank.org/`

	publications
amazon	3863
brazil	2806
indonesia	956
africa	955
china	920
america	810
europe	785
mexico	635
costa rica	330
malaysia	248

TAB. 2 – *Les pays et régions les plus souvent mentionnés dans les résumés des publications pour la période de 1975 à 2016.*

4.1 Production par auteur et réseau de co-auteurs

Le graphe des co-auteurs dans la figure 7 fournit un aperçu sur les tendances de collaboration. Il montre les nombreux groupes formés par les auteurs qui collaborent sur le sujet.

Les auteurs ayant publié le plus figurent dans le tableau 3 avec un auteur Brésilien en tête, Fearnside, suivi d'un auteur d'Australie (Laurance) et d'un auteur Américain (Houghton) en deuxième et troisième positions respectivement. Deux auteurs sont donc issus des deux pays avec la plus importante production globale tandis que le troisième est issu d'un des pays avec la plus importante production par personne pour l'année 2016.

	publications
Fearnside, PM	97
Laurance, WF	73
Houghton, WF	63
Lambin, EF	58
Shimabukuro, YE	57
Koh, LP	56
Herold, M	49
Asner, GP	49
Achard, F	48
Peres, CA	44

TAB. 3 – *Publications des 10 auteurs ayant publié le plus.*

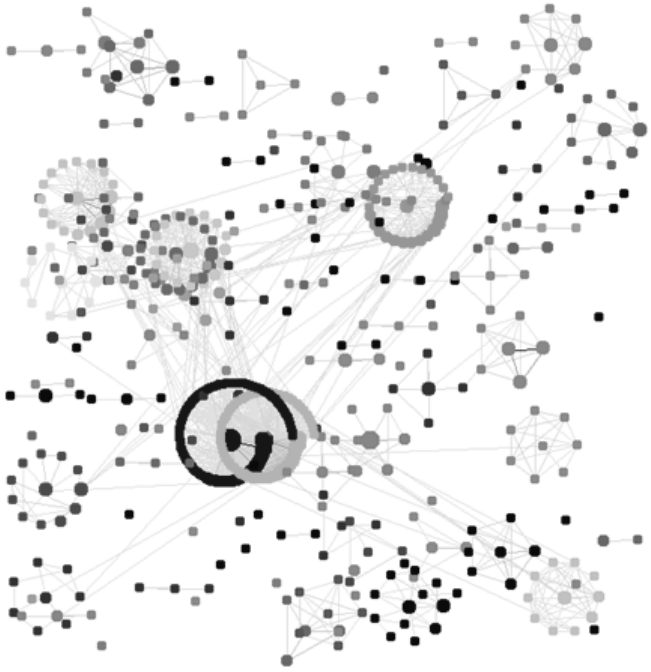


FIG. 7 – *Regroupement de co-auteurs.*

4.2 Les disciplines représentées

Le réseau des disciplines scientifiques les plus représentées est construit avec les catégories définies par le Web of Science (champ WC). Une publication pouvant se retrouver dans plusieurs catégories, il est possible de calculer les co-occurences des catégories et d'utiliser le résultat pour construire un réseau. Ainsi, il en ressort que la plupart des disciplines se regroupe autour de l'environnement et de l'écologie. Il se forme également un deuxième groupe ayant rapport à la médecine comprenant notamment la médecine tropicale et la parasitologie.

Le tableau 4 permet de voir les disciplines dans lesquelles le plus de publications ont été classées (champ WC du WoS). Ainsi, la majorité des publications est classée sous la catégorie des sciences environnementales. En regardant les autres disciplines il est possible de voir quelles sous-disciplines des sciences environnementales sont les plus représentées. L'écologie et les géosciences comptent le plus de publications. Il est possible de conclure que les publications sur la déforestation ont tendance à être interdisciplinaires, les sciences environnementales étant elles-mêmes interdisciplinaires et regroupant entre autre l'écologie et les géosciences.

	publications
Environmental Sciences	3890
Ecology	2812
Geosciences, Multidisciplinary	1571
Environmental Studies	1491
Biodiversity Conservation	1298
Forestry	1262
Meteorology & Atmospheric Sciences	1104
Geography, Physical	907
Remote Sensing	901
Multidisciplinary Sciences	807

TAB. 4 – *Les 10 catégories (champ WC du WoS) avec le plus de publications.*

5 Travaux reliés

Les analyses que nous avons présentées portent sur un corpus dans lequel se retrouvent les publications issues de diverses universités, laboratoires, unités de recherches et autres institutions. Les publications sont toutes liées au même thème de la déforestation.

Neptune (2014) a déjà utilisé des analyses bibliométriques de publications scientifiques pour analyser les activités de recherche au sein d'un unité scientifique spécifique, l'Institut de Recherche en Informatique de Toulouse. Ce travail portait sur toutes les publications présentes dans la base de donnée de l'unité de recherche, tous thèmes confondus. En utilisant les données sur l'organisation et le personnel du laboratoire, les analyses sur la production par équipe ainsi que sur les collaborations inter-équipe et la collaboration avec des auteurs extérieurs à l'unité ont pu être effectuées. L'auteur a montré comment l'analyse bibliométrique peut être utilisée pour certains aspects de l'évaluation d'une unité scientifique tels que la production scientifique, le rayonnement, l'implication dans la formation par la recherche et les perspectives scientifiques. Mothe et al. (2006) a démontré comment la plate-forme Tétralogie permet de combiner la fouille de données avec les fonctionnalités des systèmes d'information géographique pour découvrir la structure géographique d'un domaine. Les auteurs ont présenté une étude de cas en utilisant les actes de la conférence SIGIR (Special Interest Group on Information Retrieval) de l'ACM (Association for Computing Machinery). Les auteurs ont utilisé des cartes géographiques pour représenter visuellement la dimension géographique révélée par la fouille des données.

Les analyses présentées ici font suite à ces travaux. Nous utilisons la fouille de données et de méta-données de publications scientifiques pour analyser les activités de recherche sur le sujet de la déforestation. Nous nous sommes intéressés à la dimension géographique présente dans les données non seulement par rapport à la localisation des auteurs mais aussi par rapports aux zones et régions sur lesquelles portent leurs travaux de recherche.

Skupin (2014) a utilisé conjointement la bibliométrie et la visualisation des réseaux

pour faire ressortir la structure d'un domaine ainsi que les communautés basées sur les co-citations avec les publications de l'auteur David Mark. Cette approche a permit de réaliser une analyse visuelle de la dimension de l'influence de David Mark et sa persistance avec le temps dans le domaine des systèmes d'information géographiques.

Kang et al. (1990) a proposé une étude de faisabilité sur la méthode FODA (Feature-Oriented Domain Analysis) pour l'analyse d'un domaine. Cette méthode permet de créer un modèle du domaine en effectuant notamment une analyse de l'étendue du domaine. Les analyses présentées ici permettent d'élucider le domaine de la déforestation en vue de guider des travaux futurs sur les données liés à ce thème.

6 Perspectives

Les données collectées augmentées avec des données externes telles que celles de l'Organisation des Nations Unis pour l'alimentation et l'Agriculture et de la Banque Mondiale peuvent permettre de répondre à de nombreuses autres questions telles que celles liées à l'évolution des thèmes de recherche et à l'apport des institutions à chaque thème. Par exemple, le lien entre l'expérience de la déforestation dans les pays et la publication sur le sujet peut être examiné à partir de ces données. De plus, les collaborations entre institutions peuvent aussi être explorées avec des analyses de réseaux sociaux.

Ce travail pourrait être complété par une analyse plus poussée des résumés ou des articles complets notamment avec l'extraction des entités nommées qui faciliterait la mise en évidence des sujets spécifiques d'intérêts liés à la déforestation telles que des maladies, des plantes et des animaux spécifiques.

7 Conclusion

La fouille de texte des données et méta-données sur les publications scientifiques en rapport à la déforestation a permis d'avoir un aperçu de l'évolution de l'activité de recherche sur ce sujet au fil des années ainsi que la collaboration quasi-généralisée entre les pays et les nombreux réseaux de collaboration entre auteurs. Une augmentation quasi-régulière du nombre de publications est constatée d'une année à l'autre.

Le Brésil ressort comme un acteur important tant par la contribution de ses auteurs que par les mentions qui sont faites du pays dans les publications analysées.

Références

Bornmann, L. et R. Mutz (2015). Growth rates of modern science : A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology 66*(11), 2215–2222.

Diegues, A. C. S., P. Kageyama, et V. Viana (1992). *The social dynamics of deforestation in the Brazilian Amazon : an overview*, Volume 36. United Nations Research Institute for Social Development.

Dousset, B. (2009). Tetralogie : Software for monitoring science and technology.

Foley, J. A., R. DeFries, G. P. Asner, C. Barford, G. Bonan, S. R. Carpenter, F. S. Chapin, M. T. Coe, G. C. Daily, H. K. Gibbs, J. H. Helkowski, T. Holloway, E. A. Howard, C. J. Kucharik, C. Monfreda, J. A. Patz, I. C. Prentice, N. Ramankutty, et P. K. Snyder (2005). Global consequences of land us. *Science 309*, 570–574.

Kang, K. C., S. G. Cohen, J. A. Hess, W. E. Novak, et A. S. Peterson (1990). Feature-oriented domain analysis (foda) feasibility study. Technical report, Carnegie-Mellon Univ Pittsburgh Pa Software Engineering Inst.

Malhado, A. C. M., R. S. D. de Azevedo, P. A. Todd, A. M. C. Santos, N. N. Fabré, V. S. Batista, L. J. G. Aguiar, et R. J. Ladle (2014). Geographic and temporal trends in amazonian knowledge production. *Computers, Environment and Urban Systems 46*, 6–13.

Mothe, J., C. Chrisment, T. Dkaki, B. Dousset, et S. Karouach (2006). Combining mining and visualization tools to discover the geographic structure of a domain. *Computers, Environment and Urban Systems 30*, 460–484.

Neptune, N. (2014). ánalyses bibliométriques des publications de l'irit.

Pritchard, A. (1969). Statistical bibliography or bibliometrics. *Journal of Documentation 25*, 348–349.

Skupin, A. (2014). Making a mark: a computational and visual analysis of one researcher's intellectual domain. *International Journal of Geographical Information Science 28*(6), 1209–1232.

Summary

Deforestation is a widespread phenomenon that affects fairly large portions of land, especially in tropical regions. Remote sensing allows researchers to track and analyze the spatio-temporal evolution of this phenomenon.

Using text and metadata mining on scientific publications, on the theme of deforestation, we aim to identify the location of scientific production on deforestation and find out how researchers are connected to each other. Through network analysis, it is possible to highlight trends in terms of collaboration between authors. This network analysis reveals trends in the distribution of production among authors, whether it is concentrated at the level of particular authors in developed countries or whether it tends to be distributed in a balanced way between several developed and developing countries.

For this we rely on network analyses. Moreover, thanks to the analysis of the keywords we identify deforestation-affected sites that researchers are interested in, tropical forests and the Amazon, as well as related subjects related to the environment and the environment and health.

Constitution d'un corpus d'articles scientifiques avec représentation sémantique

Jean-Claude Moissinac*

*LTCI, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France
jean-claude moissinac@telecom-paristech.fr,
https://moissinac.wp.imt.fr/

Résumé. Dans le cadre du projet SemBib, nous avons entrepris une représentation sémantique de la production scientifique de Télécom Paristech. Au delà des objectifs internes, ce corpus enrichi est une source d'expérimentation et une ressource pédagogique. ce travail repose sur l'utilisation de méthodes de fouilles de texte pour construire des graphes de connaissances, puis sur la production d'analyses à partir de ces graphes. La proposition principale exposée est la méthodologie de production de graphes disjoints, aux rôles bien identifiés, afin de permettre des utilisations différenciées, et en particulier la comparaison entre méthodes de production et d'exploitation des graphes. Cet article est avant tout une proposition méthodologique pour l'organisation de représentation sémantique de publications, en s'appuyant sur des méthodes de fouille de texte. La méthode proposée facilite des approches d'enrichissement progressifs des représentations avec possibilités d'évaluation à chaque étape.

1 Introduction

Le projet SemBib est une initiative au sein de Télécom ParisTech pour constituer et exploiter un graphe de connaissances sur nos publications scientifiques. Face à de grands entrepôts de références bibliographiques, nous considérons qu'une fédération de projets analogues à SemBib a du sens. En particulier, une institution est mieux à même de travailler sur la qualité de son propre corpus et de tendre vers une représentation significative de sa propre production, ce que les grands entrepôts ne peuvent pas assurer. Nous plaidons donc pour une fédération de dépôts locaux d'articles scientifiques sémantiquement reliés [Moissinac (2017)].

SemBib nous sert aussi de base d'expérimentation pour des travaux sur les représentations sémantiques et l'exploration de graphes. Nous pensons utile de nous appuyer sur ce type de représentation pour réaliser et étendre les nombreuses approches bibliométriques développées au cours des décennies écoulées. Certains situent en 1950, d'autres en 1926, d'autres encore avant les bases de la bibliométrie et de la scientométrie. Bien sûr, nous tenons compte de cette longue histoire ;, mais nous nous intéressons surtout aux services que peuvent rendre aux chercheurs les méthodes de représentation sémantique appliquées à la production scientifique, en croissance rapide, dépassant souvent les capacités individuelles d'exploration. Enfin, SemBib s'avère un bon cas d'étude pour découvrir l'usage des graphes de connaissances, motivant pour nos étudiants, qui choisissent régulièrement de mener à bien des projets sur ces données.

Dans cet article, nous présentons la constitution du corpus dans la section 2 et les problèmes spécifiques posés. Dans la section 3, nous illustrons des utilisations de cette représentation et présentons des méthodes d'enrichissement du corpus. Dans la section 4, nous présentons des réalisations basées sur SemBib et ébauchons des méthodes croisant graphe de connaissances et méthodes d'apprentissage. Enfin, en conclusion, nous présentons nos perspectives concernant la publication de ces données et de conseils méthodologiques pour interconnecter des graphes similaires.

2 Constitution du corpus

2.1 Contexte

De nombreuses initiatives visent à améliorer les parcours dans la masse de connaissances que constituent les publications scientifiques. Certaines s'appliquent à donner une vision analytique d'un ensemble de citations. Citons par exemple le travail de Sateli et al. (2016) pour associer des compétences à des personnes en analysant leurs publications. D'autres, par exemple, aident à trouver des documents pertinents sur un sujet donné comme Rizzo et al. (2015). Le travail le plus abouti que nous avons identifié à ce sujet est celui de Tang et al. (2008).

L'accès massif à des données bibliographiques est aussi offert par quelques grands systèmes de référencement (Google Scholar [1], Microsoft Academic Graph [2] (consulté le 1/12/2017), par exemple). Voir [Moissinac (2017)] pour une analyse plus complète de sources de données bibliographiques.

Cependant, toutes ces solutions se révèlent donner une vision très partielle de notre production scientifique et ne sont pas ou peu ouvertes à la production d'analyses et d'exploitation étendant celles qu'ils proposent déjà. Comme montré par Larsen et von Ins (2010), le nombre de publications scientifiques croit rapidement ; il devient donc de plus en plus important d'outiller les utilisateurs de ces publications pour qu'ils puissent les exploiter efficacement.

2.2 Données de référence

SemBib s'appuie sur l'existence d'une base de données déjà constituée qui recense les principales méta-données sur les publications de Télécom ParisTech [3]. Il existe des bases analogues dans de nombreuses institutions. Les approches présentées ici permettent d'enrichir et interconnecter de telles bases, dans une approche décentralisée de l'information bibliographique.

Cette base identifie 11311 documents [4] publiés depuis 1969. Cette base constitue une référence essentielle, au moins pour les titres, l'année et les auteurs de chaque publication, ainsi que l'attribution à Télécom ParisTech. D'autres méta-données sont inégalement renseignées : identifiant unique DOI, URL vers le document complet, mots-clés...

Sur les 11311 documents référencés :

— 1394 associent des mots-clés proposés par un auteur, soit 12%,

1. https ://scholar.google.fr/ (consulté le 1/12/2017)
2. https ://academic.microsoft.com
3. http ://biblio.telecom-paristech.fr/cgi-bin/selectform.cgi (consulté le 1/12/2017)
4. au 20/11/2017

— 1048 associent une URL, supposée désigner une source du document complet, soit 9% ; cependant, dans beaucoup de cas cette URL est obsolète, ou pointe vers une page qui offre une redirection vers le document souvent payante ou difficile d'accès par programme,
— 1178 associent un identifiants DOI, soit 10% ; les notations utilisées ici, renseignées par les auteurs, ont une grande variabilité, par exemple ils se présentent avec ou sans préfixe doi :.

De plus, nous avons noté une grande variabilité dans les chaînes de caractères utilisées pour désigner les canaux de publication (conférences ou journaux).

Ces caractéristiques rendent cette base peu propre à une exploitation approfondie, mais elle permet de connaitre les entités -auteurs et publications- sur lesquelles nous devons récolter de l'information et d'estimer si des données récoltées doivent bien être associées avec une de nos publications. De plus, à partir de cette base, notre connaissance d'une bonne partie des entités qu'elle mentionne nous permet de créer des ensembles de données de référence, constituant des 'vérités terrain'.

De grands entrepôts de données bibliographiques existent par ailleurs. Malheureusement, ils donnent une vision très tronquée de notre production, notamment parce qu'ils ne sont pas en mesure de résoudre les changements de dénomination de notre institution et leurs variantes usuelles. De plus ces bases ne disposent pas d'informations sur des structures internes de recherche : projets, départements et groupes de recherche...

Les documents connus des grands entrepôts utilisent de nombreux intitulés différents au niveau de l'affiliation des auteurs, ce qui rend incertaine la recherche sur ces entrepôts de toutes les publications attribuables à Télécom ParisTech. Par exemple, HAL ne recense que 3001 publications [5] sur les 11311 connues de notre base. Ils utilisent souvent aussi des variantes multiples pour les noms d'auteurs.

2.3 Fonctions recherchées et problèmes posés

L'objectif est d'abord d'enrichir les descriptions sémantiques de chaque auteur de chacune de nos publications et une description de chaque publication et canal de publication.

Pour cela, nous devons faire des choix de modèle de représentation. Notre hypothèse est que les avancées du web sémantique sont à même de nous donner de nouveaux moyens pour exploiter efficacement les données que nous rassemblons, en vue d'assurer des fonctions de recherche et d'analyse.

Compte tenu des faiblesses des données de base dont nous disposons, nous devons d'abord trouver d'autres sources de données et consolider l'ensemble par des croisements entre ensembles de données. Une fois les données récoltées, et en particulier les documents publiés, nous devons analyser ceux-ci pour en produire des représentations sémantiques (dans l'hypothèse où ces représentations assurent une bonne base pour des analyses très diverses)

Cela suppose notamment :
— de valider la pertinence des documents et données collectées
— de lever les ambiguïtés sur les affiliations, les noms d'auteur, les titres, les canaux de publication

5. au 8/1/2018

— de modéliser les différents types de données pour leur intégration dans un ou plusieurs graphes sémantiques dans l'esprit du LOD (Linked Open Data)

Ces problèmes -et d'autres- montrent pourquoi la constitution d'un corpus qualifié peut être complexe, l'intérêt qu'il y a à outiller sa constitution et enfin, l'intérêt que les données ainsi réunies peuvent avoir pour des travaux thématiques sur la fouille de texte s'appuyant sur la constitution de graphes de connaissances.

2.4 Choix de représentation

Nos choix sont portés par le choix principal : s'appuyer sur les technologies du web sémantiques. Pour les données issues de notre base de référence, il nous a semblé inutile de créer des liens forts avec le reste du Linked Open Data. Nous avons donc créé notre propre vocabulaire qui constitue une projection directe des données de la base de référence, sauf pour quelques données pour lequel un choix évident était disponible : foaf pour les noms de personnes, dc (Dublin Core) pour certaines informations sur les documents.

Dans un deuxième temps, après étude de plusieurs ontologies spécialisées, nous avons retenu la famille d'ontologies SPAR pour sa couverture d'un large ensemble de concepts liés à la représentation bibliographique. Nous avons représenté avec des concepts et des propriétés de SPAR des valeurs associées à nos publications (par exemple avec l'URI fabio :ResearchPaper). Enfin pour les concepts qui constituent nos vocabulaires du domaine (voir plus loin) nous avons cherché les liens avec schema.org, DBPedia, Wikidata et nous allons intégrer Wordnet.

2.5 Réunir les documents et opérer des traitements élémentaires

Un crawler Python a permis de chercher de façon systématique des documents référencés dans notre base, lorsqu'ils n'étaient pas disponibles sur notre site. Nous avons pu récolter environ 5000 documents sur les 11000 référencés comme publication de Télécom Paristech. Étant producteurs de ces documents, nous avons le droit d'en détenir une copie pour nos traitements ; une analyse est en cours pour déterminer ceux qui vont pouvoir être rendus publiquement disponibles dans le respect des droits des éditeurs. La plupart des documents sont au format PDF.

Le crawler récupère des documents cherchés sur le Web à partir du titre. Ensuite, il faut vérifier que le document obtenu correspond bien au document cherché. Nous nous sommes inspiré de la stratégie de Tang et al. (2008) : nous vérifions que le titre est très similaire , que le nombre d'auteurs est identique, que les noms d'auteur sont très similaires et enfin que la date de publication est la même (si elle est disponible) ; les documents qui ne remplissent pas ces conditions sont conservés pour alimenter notre base de documents, mais pas associés aux auteurs, ni à Télécom ParisTech. Cette approche simple nous a permis une grande précision (100

Pour chaque document :

— une représentation sémantique des metadonnées a été réalisée en s'appuyant sur les ontologies SPAR (bibo, fabio) et les ontologies couramment utilisées pour des documents (Dublin Core, schema, foaf...) ;

— l'extraction d'une représentation structurée a été réalisée à l'aide de l'outil GROBID [Lopez (2009)] qui produit une représentation TEI[6] de l'article ; elle nous permet

6. http ://www.tei-c.org/

d'avoir accès aux différentes parties du texte, en conservant des informations structurelles et rhétoriques sur chaque portion de texte ; un graphe pour chaque article est en cours de conception en s'appuyant sur l'ontologie DoCO [7][Constantin et al. (2016)] ;
— les mots-clés indiqués par les auteurs dans notre base ou extraits des documents ont été ajoutés dans un graphe constituant un vocabulaire de base du domaine ; nous reviendrons sur les graphes de concepts construits à partir de mots ; plusieurs graphes ont été construits ;
— le résumé a été tiré de cette dernière représentation pour être ajouté dans le graphe sémantique ;
— la fréquence des termes de chaque document a été calculée et stockée dans un fichier associé ; nous verrons plus tard s'il y a lieu de mettre ces informations dans un graphe de connaissances ;
— les termes les plus fréquents de chaque document, hors termes vides ou mis en liste noire, ont été ajoutés à un autre graphe.

2.5.1 Graphes

Pour la représentation sémantique, notre approche consiste à répartir les données dans plusieurs graphes qui se référencent les uns les autres ; il y a par exemple un graphe pour les personnes dont les URIs sont utilisées dans le graphe sur les publications, lequel référence aussi le graphe sur les canaux de publication. Cette approche transpose les principes de design SoC -Separation of Concerns- appliquées dans le design logiciel ; elle contribue à faciliter l'évolution disjointe de graphes aux fonctions différentes. Les références utilisent les principes du web sémantique : chaque référence est désignée par une URI.

Les principaux graphes sont :

— un graphe pour les personnes (auteurs), comportant 7478 personnes (il inclue les co-auteurs externes à Télécom ParisTech) décrits par 50411 faits ;
— un graphe pour les publications, comportant 11311 publications décrits par 195857 faits (triplets RDF) ;
— un graphe pour les canaux de publication, comportant 4407 canaux décrits par 6599 faits ;
— un graphe pour les concepts du domaine, comportant 15964 entités décrits par 43878 faits

Les tailles des graphes sont en constante évolution au fil des enrichissements successifs apportés aux différents graphes par intégration de données issues de nouvelles sources. On voit par exemple que le graphe sur les canaux de publications contient à peine plus d'un fait par canal. C'est le prochain graphe que nous allons significativement enrichir avec WikiCFP et les pages des sites des conférences référencées dans notre base de référence.

Le choix est fait d'ajouter des entités au graphe de concepts du domaine chaque fois qu'un nouveau concept parait utile -par exemple sur des critères liés au TfIdf sur le corpus- mais de limiter ce graphe à des concepts isolés les uns des autres, éventuellement reliés à des entités externes, comme des entités de DBPedia. Les relations entre concepts sont décrites dans des graphes séparés afin de pouvoir tester différentes stratégies de mise en relation et de comparer

7. http ://www.sparontologies.net/ontologies/doco/source.html (consulté le 24/11/2017)

leurs résultats. Nous verrons en section 3 des conséquences de l'approche générale qui conduit à travailler avec un ensemble de graphes disjoints.

La répartition en graphes disjoints facilite aussi la constitution de graphes qui vont servir de "vérité terrain" pour diverses opérations, en particulier concernant les méthodes de collecte de données et les méthodes d'association de thématiques à d'autres entités (auteurs, publications, conférences).

2.5.2 Les graphes de concepts

Nous avons choisi de créer plusieurs graphes de concepts, qui vont principalement différer par leur méthode de construction, mais éventuellement aussi par leurs méthodes d'exploitation. Comme vu plus haut, les mots-clés indiqués par les auteurs ont été mis dans un premier graphe. Cependant, nous avons vu la faible utilisation de ces mots dans notre base : un dixième des publications ont des mots-clés associés, moins de la moitié des auteurs renseignent les mots-clés toujours ou quelques fois. Aussi, nous avons ajouté à ce graphe les mots-clés tirés des articles par notre extraction TEI avec GROBID. Pour chaque mot retenu, après normalisation, une URI dans notre espace de nommage a été créée. Ce graphe constitue notre premier vocabulaire de domaine.

Un deuxième graphe a été alimenté par une sélection de mots obtenues en calculant les coefficients TfIdf pour l'ensemble des mots rencontrés d'une sous-partie du corpus portant sur 5 années récentes (environ 4000 références, mais seulement environ 1200 textes complets au moment de la réalisation) et en gardant les mots apparaissant comme les plus significatifs. Cela constitue un deuxième vocabulaire du domaine.

Un troisième graphe a été construit en éliminant les mots creux et des mots mis en liste noire du texte intégral obtenu grâce à l'extraction avec GROBID des textes complets d'une sous-partie du corpus portant sur 5 années récentes.

Nous pensons qu'il est utile de construire ces graphes, et peut-être d'autres par la suite, afin de comparer facilement différentes approches. Une comparaison systématique de ces trois graphes reste à établir. Les exploitations des données de SemBib actuelles ont été réalisées avec le premier graphe. Nous verrons plus loin que les concepts retenus peuvent constituer une base pour une analyse des textes basés sur des vecteurs de concepts au lieu de vecteurs de mots.

La constitution de ces graphes liés à des graphes sémantiques externes constitue une base de travail pour la fouille de textes en s'appuyant sur des représentations sémantiques et pas seulement sur des ensembles de mots ou des structures linguistiques. Cela prolonge les démarches que nous avons présentées dans [Vincent et al. (2014)] pour le couplage de représentation structurelles et sémantiques des textes. Des graphes compagnons sont en cours pour associer leurs concepts avec des concepts de graphes de référence.

2.5.3 Le graphe des canaux de diffusion et des affiliations : traitements associés

Que ce soit dans notre base de référence ou dans les sources externes, de nombreuses variantes de chaînes de caractères sont utilisées pour désigner une même série de conférences - acronymes, acronymes avec année, nom développé, ...- ou une même conférence. Par exemple, les conférences de la série où nous publions le plus, ICASSP, sont référencées dans nos sources

sous 31 dénominations différentes. Il en est de même pour la désignation de Télécom ParisTech, pour lequel nous avons rencontré 53 désignations différentes.

Nous avons travaillé à la construction d'un graphe dont les noeuds principaux sont des URIs uniques associées à chaque canal de publication, conférence par exemple, à chaque série de conférence, à chaque affiliation. Chaque conférence est associée à un ensemble de dénominations ainsi, le cas échéant, à une série de conférences. Dans le graphe des publications, chaque publication est associée à un canal de publication et à une ou plusieurs affiliations. Pour ramener à une URI unique, chaque désignation se rapportant à une seule et même organisation, nous avons utilisé une méthode semi-interactive, qui trouve sa place lorsque le nombre d'éléments à traiter est limité.

Par exemple, pour Télécom ParisTech, nous avons :

— établi manuellement une première liste de représentations connues
— ajouté une version normalisée de ces représentations : tout mis en minuscule, caractères accentués remplacés par leur équivalent non accentué...
— cherché dans toutes les affiliations rencontrées dans nos sources celles qui contiennent une des représentations de base connues ou une représentation similaire
— vérifié dans la liste trouvée les nouvelles désignations acceptables qui sont alors ajoutées dans la liste des désignations connues
— ré-itéré ce processus jusqu'à ce qu'aucune nouvelle désignation acceptable ne soit trouvée

Ce traitement a deux conséquences :

— toutes les publications concernées par ces désignations ont pu être rattachées à une seule organisation désignée par son URI ;
— dans de futures publications, les affiliations pourront être comparées à cette liste de référence pour multiplier les chances de bien attacher les publications des auteurs de Télécom ParisTech à cette institution

Le même processus a été appliqué à des conférences pour lesquelles la liste initiale était réduite à l'acronyme de la conférence, éventuellement complété par une forme du nom développé. Les méthodes ainsi mise en oeuvre pour l'identification, la désambiguïsation et la cohérence sont applicables à d'autres organismes.

3 Explorer un graphe bibliographique

3.1 Point d'accès SPARQL

Un mode d'accès habituel sur des graphes de faits consiste à établir des requêtes sur ce graphe via un point d'accès SPARQL. Les graphes présentés précédemment sont exploitables via un point d'accès SPARQL. SPARQL est un langage de requêtes sur des ensembles de faits décrits par des triplets RDF et rassemblés dans des graphes (voir exemple ci-après). Plusieurs mises en oeuvre ont été assurées concernant le base de données et le point d'accès SPARQL (Triple Store) : avec Virtuoso, avec Jena-Fuseki, avec ARC2.

L'avantage d'avoir réparti les données dans plusieurs graphes est lié à la limitation de la taille de chaque graphe, à la fois en nombre de noeuds et en nombre de relations, mais, surtout, de séparer les faits pour les organiser. Avec ARC2, cela nécessite de rédiger les requêtes en sachant quand différencier les graphes où chercher certaines données. Cela complexifie la ré-

daction de requêtes (voir exemple ci-après). C'est la seule possibilité avec ARC2. Des options de Virtuoso et Jena-Fuseki permettent d'établir des requêtes qui considèrent comme graphe par défaut la fusion de l'ensemble des graphes, ce qui simplifie grandement leur écriture. Duan et al. (2011) montrent que les performances de différents points d'accès varient en fonction de caractéristiques complexes des graphes réellement ciblés ; nous ne nous sommes donc pas appuyés sur des considérations de performances mesurées par des benchmarks pour nos choix de point d'accès, mais sur des considérations techniques liées à l'hébergement.

Des tests sur des requêtes portant simultanément sur plusieurs graphes ont été effectués avec Jena-Fuseki. De façon surprenante, ils révèlent de meilleures performances sur les graphe fusionnés qu'avec une requête explicitant les graphes où chercher les données.

Par exemple, la requête [8] :

```
select distinct ?title { ?si ieee:title ?title .
                         ?s dcterms:title ?title }
```

qui explore le pseudo-graphe par défaut pour donner 824 titres s'exécute 100 fois en 19s, tandis que la requête

```
select distinct ?title {
graph sb:ieee { ?si ieee:title ?title } .
graph tpt:biblio { ?s dcterms:title ?title }}
```

qui distingue les sources de données de chaque graphe, donne le même résultat en 24s.

3.1.1 Exemple : les auteurs partageant des mots-clés

La requête suivante :

```
select distinct ?c1 ?c2 {
graph tpt:biblio {
    ?s1 a fabio:ResearchPaper;
        dcterms:creator ?c1;
       schema:keyword  ?k1;
       schema:keyword  ?k2 .
    filter(?k1!=?k2) .
    ?s2 a fabio:ResearchPaper;
            dcterms:creator ?c2;
       schema:keyword  ?k1;
       schema:keyword  ?k2 .
    filter(?k1!=?k2) .
    filter(?s1!=?s2) .
    filter(?c1!=?c2) .
  }
}
}
```

8. pour simplifier, ici et dans la suite, les préfixes ont été omis

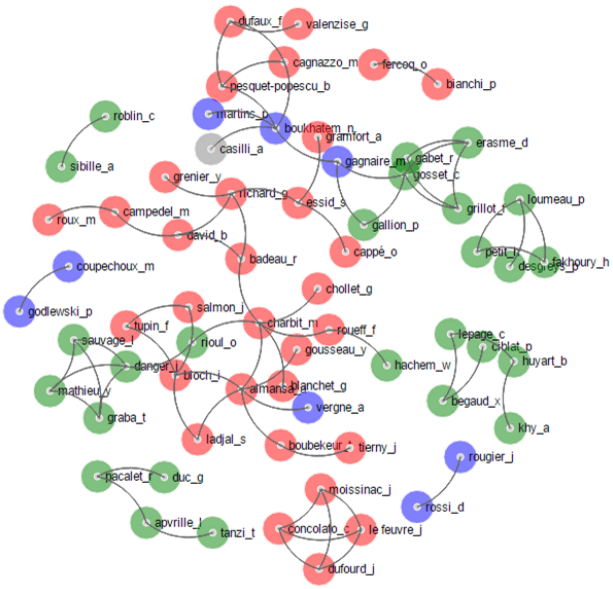


FIG. 1 – *Graphe des permanents partageant des mots-clés*

permet de sélectionner deux auteurs (?c1 et ?c2) de deux publications différentes (?s1 et ?s2) partageant deux mots-clés.

Une requête similaire, portant seulement sur les permanents de Télécom ParisTech, a permis d'établir la représentation graphique de la Figure 1. Elle révèle des communautés qui n'apparaissent pas directement dans l'organisation structurelle de Télécom ParisTech.

Les points d'accès SPARQL permettent d'obtenir leur résultats au format JSON, qui est le format privilégié de la représentation de données sur le Web. Ce format est très adapté à la consommation des données par la librairie D3, avec laquelle la Figure 1 a été produite. D3 permet de produire des graphiques pour le Web avec la technologie SVG.

Des progrès importants restent à faire pour améliorer l'exploitation des graphes sémantiques. leur adoption, en particulier sur le Web, est encore très limité par rapport au potentiel de ces représentations, surtout porté par des exploitations thématiques -musique, évènements...- portées par les grands moteurs de recherche (cf schema.org). Les utilisateurs, qu'ils soient développeurs, avec des APIs bien pensées, ou utilisateurs finaux souhaitant explorer les données, doivent être aidés pour tirer parti de ces représentations. Des facilités doivent être offertes aux développeurs pour produire des interfaces intégrant en profondeur les enrichissements possibles grâce au Web sémantique.

Nous avons exploré l'utilisation de :

— SemanticForms [9] pour créer des formulaires permettant de créer, de modifier et d'enrichir des faits dans un graphe sémantique,
— Uduvudu qui propose un modèle de développement d'interfaces Web à base de requêtes sur des graphes sémantiques ; le modèle permet de faire une bonne séparation des compétences entre le spécialiste des données sémantiques, le concepteur d'application et le développeur d'interfaces ; le portail dont il est question à la section suivante s'appuie sur Uduvudu [Luggen et al. (2015)].

3.2 Accès aux données liées au portail SemBib

Un portail d'accès aux données de SemBib a été mis en oeuvre [10]. Il est en évolution rapide pour intégrer les enrichissements possibles grâce aux graphes générés.

Pour un auteur, nous aurons des pages HTML comme :

`http://givingsense.eu/sembib/onto/persons/David_Bertrand`

ce même accès permet d'obtenir des informations au format JSON facilement exploitable par des logiciels en tous langages, avec l'adresse similaire suivante :

`http://givingsense.eu/sembib/onto/persons/David_Bertrand.json`

Les mêmes principes sont appliqués aux publications. Ainsi, sans avoir besoin de requêtes SPARQL, il est aussi possible d'accéder par programme à une partie des données. De plus, les pages générées embarquent les données correspondantes en RDFa à l'intérieur de la page. Cela rend ces données directement exploitables par les moteurs de recherche et autres référenceurs, ce qui contribue à améliorer la visibilité de nos travaux.

3.2.1 Exploration et enrichissement du graphe de concepts

Une approche courante consiste en l'association de concepts aux auteurs, aux publications et aux canaux de publications. Cette association est une base pour une mise en correspondance entre des (groupes d')auteurs, ou entre un (groupe d')auteur et une publication ou un canal de publication. Dans cette section, nous abordons l'association de concepts aux différentes entités.

Seulement un dixième des publications ont des mots-clés associés par les auteurs lorsqu'ils enregistrent leurs publications dans notre base. Moins de la moitié des auteurs renseignent toujours ou quelques fois des mots-clés. Seulement 39 mots-clés sont utilisés plus de 5 fois dans la base. Cette relative faiblesse de notre base nous a incité à collecter beaucoup plus de valeurs -mots-clés, concepts, thématiques- directement à partir du contenu des articles (voir plus haut les graphes de concepts).

Notre principal axe de travail actuel porte sur l'enrichissement et l'exploitation du graphe de concepts générés à partir des vocabulaires rencontrés dans les articles (mots-clés, mots distinctifs obtenus avec Tf-Idf...). D'autres méthodes classiques de fouille de textes pourront à l'avenir être intégrées pour les coupler et les enrichir par les méthodes des graphes sémantiques. L'idée est d'utiliser des graphes externes -DBpedia, Wikidata, Wordnet- permettant d'établir des relations entre les concepts que nous utilisons et d'exploiter ces relations pour mieux interpréter les données.

9. https ://github.com/jmvanel/semantic_forms (consulté le 8/1/2018)
10. http ://givingsense.eu/sembib/

Des graphes compagnons de ceux précédemment évoqués sont en cours de constitution. Il contiennent des liens owl :sameAs, dc :subject et skos :broader entre entités des différents graphes ou vers des graphes externes tels que DBPedia, Wikidata,... owl :sameAs permet d'indiquer que deux URIs sont considérées comme désignant la même entité- et skos :broader permet d'indiquer une relation hiérarchique entre des entités. Nous isolons ces associations du graphe principal de concepts afin de pouvoir travailler sur différentes stratégies d'association.

La première étape consiste à tenter d'associer chaque concept de notre graphe à au moins un concept d'un des graphes externes. Plusieurs approches ont été mises en oeuvre pour cette association. Par exemple, en utilisant le service DBPedia Spotlight ou le service DBPedia Lookup.

Ensuite, nous devons établir des liens entre concepts de notre graphe. L'approche qui semble la plus prometteuse est celle proposée par Mirizzi et al. (2012). Le principe est d'établir une représentation de chaque concept de notre graphe à partir de relations existants dans DBPedia, puis d'établir une évaluation de la similarité entre ces représentations et, enfin, d'établir une relation entre les concepts les plus similaires.

Nous avons aussi entrepris d'évaluer des méthodes d'apprentissage se basant nous plu sur des vecteurs de mots, mais sur des vecteurs de concepts, en exploitant la possibilité de regrouper des concepts à partir de similarités pré-établies [Mirizzi et al. (2012)]. Enfin, l'association entre des entités de SemBib s'appuyant sur les concepts associés à chacune de ces entités permet de répondre à de nombreux besoins, voir Tang et al. (2008).

Une évaluation rigoureuse de ces différentes possibilités n'a pas encore été réalisée. La structuration des données que nous avons adoptées se prête bien à la mise en oeuvre d'approches très diverses. C'est notamment ce qui permet de proposer des projets d'étudiants abordant à la fois la fouille de textes et les représentations sémantiques RDF.

3.2.2 Réalisations s'appuyant sur SemBib

La disponibilité de ces données structurées en graphe de connaissances a déjà permis de nombreuses réalisations. Nous avons déjà évoqué la recherche d'auteurs partageant des mots-clés, nous pouvons aussi citer à titre d'exemple :

- la recherche des canaux de publication où les chercheurs de Télécom ParisTech sont les plus actifs ;
- les chercheurs de Télécom ParisTech qui ont déjà publié dans un même canal,
- le graphe des co-auteurs
- la recherche d'articles similaires à un article donné (projet d'étudiants)

4 Conclusion

Nous avons vu une approche méthodologique générale permettant de tester différentes approches concernant la description d'entités bibliographiques par association avec des concepts. Notre approche consiste en la production d'ensembles de données sous forme de graphes disjoints dans leur sujets et leur implémentation, mais inter-connectés par des relations. Nous avons aussi abordé des méthodes de levée d'ambiguïté sur les caractéristiques des entités bibliographiques. Une prochaine étape importante est la documentation RDF de cet ensemble de données avec DCAT, puis la publication de ces données. En effet, nous sommes persuadés

qu'un tel ensemble de données présentant des données qualifiées peut constituer un matériau précieux pour des travaux en scientométrie et en fouille de documents scientifiques.

Références

Constantin, A., S. Peroni, S. Pettifer, D. Shotton, et F. Vitali (2016). The document components ontology (DoCO). *Semantic Web 7*(2), 167–181.

Duan, S., A. Kementsietsidis, K. Srinivas, et O. Udrea (2011). Apples and oranges : a comparison of RDF benchmarks and real RDF datasets. In *SIGMOD Conference*, pp. 145–156. ACM.

Larsen, P. O. et M. von Ins (2010). The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics 84*(3), 575–603.

Lopez, P. (2009). Grobid : Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL'09, Berlin, Heidelberg, pp. 473–474. Springer-Verlag.

Luggen, M., A. Gschwend, B. Anrig, et P. Cudré-Mauroux (2015). Uduvudu : a graph-aware and adaptive ui engine for linked data. In *LDOW@WWW*.

Mirizzi, R., T. D. Noia, E. D. Sciascio, et A. Ragone (2012). Using DBpedia for searching related terms in the IT domain. Technical report, Politecnico di Bari, Via Orabona, 4, 70125 Bari, Italy.

Moissinac, J.-C. (2017). Pour une fédération de dépôts locaux d'articles scientifiques sémantiquement reliés. In *ToTh*.

Rizzo, G., Tomassetti Federico, A. Vetrò, L. Ardito, M. Torchiano, Morisio Maurizio, et R. Troncy (2015). Semantic enrichment for recommendation of primary studies in a systematic literature review. *Digital Scholarship in the Humanities, Oxford University Press, 13 August 2015*.

Sateli, B., F. Löffler, B. König-Ries, et R. Witte (2016). Semantic user profiles : Learning scholars' competences by analyzing their publications. In *Semantics, Analytics, Visualisation : Enhancing Scholarly Data (SAVE-SD 2016)*. Springer : Springer.

Tang, J., J. Zhang, L. Yao, J. Li, L. Zhang, et Z. Su (2008). Arnetminer : extraction and mining of academic social networks. *ACM Knowledge Discovery and Data Mining*, 990–998.

Vincent, G., J.-C. Moissinac, et A. Luc (2014). Automated generation of a "lossless semantic" eBook. In *17ème Colloque International sur le Document Numérique (CIDE 17), le livre post-numérique : historique, mutations et perspectives.*, Fès, Morocco.

Summary

As part of the SemBib project, we build a semantic representation of the scientific production of Telecom Paristech. Beyond the internal objectives of analysis and exploration, we think that this enriched corpus is a source of experimentation and a teaching resource. Finally, we advocate for a multiplication of inter-connected local repositories of bibliographic archives.

L'évaluation des représentations vectorielles de mots en utilisant WordNet

Nourredine Aliane*, Jean-Jacques Mariage*, Gilles Bernard*

*Laboratoire LIASD Université Paris 8, 2 rue de la Liberté 93200
nourredine@ai.univ-paris8.fr, jjm@ai.univ-paris8.fr, gilles.bernard@iedparis8.net

Résumé. Les méthodes d'évaluation actuelles des représentations vectorielles de mots utilisent un jeu de données restreint. Pour pallier à ce problème nous présentons une nouvelle approche, basée sur la similarité entre les synsets associés aux mots dans la grande base de données lexicale WordNet. Notre méthode d'évaluation consiste dans un premier temps à ranger les représentions vectorielles de mots dans des clusters par un algorithme de clustering, puis à évaluer la cohérence sémantique et syntaxique des clusters produits. Cette évaluation est effectuée en calculant la similarité entre les mots de chaque cluster pris deux à deux, en utilisant des mesures de similarité entre les mots dans WordNet proposées par NLTK (path_similarity ou wap_similarity). On obtient pour chaque cluster une valeur entre 0 et 1, un cluster dont la valeur est 1 est un cluster dont tout les mots appartiennent au même synset. On calcule ensuite la moyenne des mesures de tous les clusters. Nous avons utilisé notre nouvelle approche pour étudier et comparer trois méthodes de représentations vectorielles: une méthode traditionnelle WebSOM et deux méthodes récentes, word2vec (Skip-Gram et CBOW) et GloVe, sur trois corpus: en anglais, en français et en arabe. Les résultats montrent que la méthode de word2vec surpasse les deux autres méthodes sur les trois corpus.

1 Introduction

Une méthode de représentation vectorielle a pour but d'associer à chaque mot dans un corpus textuel, un vecteur à valeurs réelles tel que les composantes de ce vecteur décrivent le mieux possible le sens de ce mot dans son contexte. Cependant, la tâche la plus difficile est de vérifier la qualité des vecteurs produits par ces méthodes. L'évaluation se fait généralement par un travail manuel ou avec des tâches d'analogie de mots. Les auteurs de word2vec (ou GloVe) font l'évaluation de leurs méthodes avec des fichiers textuels qui contiennent des couples de mots similaires sémantiquement ou syntaxiquement, en calculant la distance euclidienne (ou distance de cosinus) entres les vecteurs correspondants, afin de mesurer le taux de correction du système. Une étude récente de (Baroni et al., 2014) conduit un ensemble d'expériences en comparant la méthode de word2vec aux autres méthodes traditionnelles. (Levy et al., 2015) ont trouvé des résultats importants, en comparant word2vec, GloVe et d'autres méthodes. Or, toutes ces méthodes d'évaluation utilisent un petit jeu de données comme : "TOEFL" (Landauer et Dutnais, 1997) constitué de 80 questions à choix multiples, "Google's

analogy dataset" contient 19 544 analogies ou "MSR's analogy dataset" qui contient 8000 analogies morpho-syntaxiques. C'est pour cela que nous proposons une nouvelle approche basée sur l'idée d'utiliser comme référence (Gold Standard) la grande base de données lexicale WordNet (Miller, 1995), qui a été faite manuellement par des experts linguistes de l'université de Princeton, et qui contient plus de 200 000 mots, plus leurs relations sémantique et lexicale. Pour expérimenter notre approche, nous avons choisi trois méthodes de représentations vectorielles de mots : word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) et WebSOM (Kohonen et al., 1998). elles sont toutes fondées sur l'hypothèse de (Harris, 1968), selon laquelle les mots apparaissant dans des contextes similaires ont un sens similaire. Les modèles de représentation vectorielle de mots transforment l'analyse distributionnelle d'un corpus en espace vectoriel, dans lequel deux vecteurs proches géométriquement représentent deux mots dont la sémantique est proche. Nous avons exploité les mesures de similarités entre les mots de WordNet présentées dans (Pedersen et al., 2004), et qui sont implémentées dans NLTK [1] .

2 Représentation vectorielle de mots

Dans ce paragraphe, nous présentons brièvement les trois méthodes choisies pour nos expérimentations.

2.1 WebSOM

Cette technique de représentation a été utilisée dans le système WEBSOM (Kohonen et al., 1998). Dans un premier temps, on associe à chaque mot m_i un vecteur x_i de dimension d avec des composantes de nombres réels, initialisés aléatoirement entre 0 et 1. Dans un second temps, le mot m_i sera représenté par un autre vecteur X_i, qui est déterminé de la façon suivante :

$$X_i = \begin{pmatrix} p_N(x_i) \\ . \\ . \\ . \\ p_1(x_i) \\ \epsilon.x_i \\ s_1(x_i) \\ . \\ . \\ . \\ s_N(x_i) \end{pmatrix}$$

Où : $p_1(x_i)$ et $s_1(x_i)$ sont respectivement les vecteurs moyens des vecteurs qui correspondent à tous les mots prédécesseurs immédiats et successeurs immédiats du mot m_i dans l'ensemble de corpus ($p_1(x_i)$ et $s_1(x_i)$ sont évidement de dimension d). La fenêtre contextuelle est de $(2N + 1)$ mots (le mot courant, N mots précédents et N mots suivants). ϵ est un réel positif inférieur à 1.0. Il sert à contrôler le rôle du mot m_i dans son contexte. Le vecteur X_i est de dimension $(2N + 1)d$

1. NLTK (Natural Language Toolkit) est une bibliothèque logicielle en Python

2.2 word2vec

Une méthode proposée par (Mikolov et al., 2013) fondée sur les réseaux de neurones, a été implémentée dans un outil qui s'appelle word2vec. Deux modèles de représentation des mots sont implémentés dans word2vec, qui sont le continuous bag-of-words (CBOW) et le Skip-Gram.

1. CBOW a pour objectif de prédire la probabilité d'un mot sachant ses contextes, Cette représentation de mots consomme moins de temps en entraînement que le skip-gram.
2. Skip-Gram contrairement aux CBOW, vise a prédire la probabilité des contextes d'un mot sachant ce mot.

2.3 GloVe

GloVe (Global Vectors for Word Representation) (Pennington et al., 2014) est un modèle proposé par l'équipe NLP de l'université de Stanford. Cette méthode combine les avantages de la factorisation matricielle globale et des méthodes de contexte local. Le contexte est une fenêtre de longueur fixe d'éléments lexicaux centrés sur le mot. On cherche à représenter chaque mot i et chaque mot j appariaient dans le même contexte par des vecteur v_i et v_j respectivement, de dimension d tels que : $v_i.v_j + b_i + b_j = log(X_{ij})$. Où X_{ij} représente le nombre de fois le mot j se produit dans le contexte du mot i. b_i et b_j sont des biais scalaires associés aux mots i et j respectivement.

3 WordNet et la similitude entre les mots

WordNet (Miller, 1995) est une grande base de données lexicale de l'anglais, développée par des linguistes de l'université de Princeton. Les mots sont regroupés en ensembles de synonymes cognitifs (synsets). Les synsets sont interconnectés au moyen de relations conceptuelles-sémantiques et lexicales. WordNet est distribué sous une licence libre, la dernière version 3.1 répertorie plus de 200 000 mots. (Pedersen et al., 2004) présentent plusieurs algorithmes de similarité entre mots, en se basant sur la structure et le contenu de WordNet. Parmi ces algorithmes, nous en avons testé deux qui sont implémentés dans NLTK.

path_similarity : retourne un score entre 0 et 1, indiquant le degré de similitude entre deux mots, en fonction du chemin le plus court qui les relie dans une taxonomie. Dans le cas d'utilisation avec WordNet, la similitude entre deux mots appartenant au même synset est égale à 1. path_similarity est définie par l'équation (1) :

$$path\_similarity(s_1, s_2) = \frac{1}{1 + longChemin(s_1, s_2)} \quad (1)$$

$longChemin(s_1, s_2)$ est le nombre d'arêtes dans le chemin le plus court entre s_1 et s_2 dans un graphe. s_i est un noeud dans un graphe, il peut être un concept, mot ou un synset.

wup_similarity (Wu et Palmer, 1994) : renvoie un score entre 0 et 1, en fonction des profondeurs des deux mots et celle de leur dernier ancêtre commun dans une taxonomie. Elle est définie par l'équation (2)

$$wup\_similarity(s_1, s_2) = \frac{2 * depth(lcs(s_1, s_2))}{depth(s_1) + depth(s_2)} \tag{2}$$

$lcs(s_1, s_2)$: le dernier ancêtre commun entre s_1 et s_2 (pour l'anglais : least common subsumer). Il correspond au dernier noeud du graphe taxonomique à partir duquel divergent les branches de s_1 et s_2. $depth(s_i)$ est la profondeur de s_i (le nombre d'arêtes entre la racine et s_i, $depth(racine) = 1$).

4 Méthodologie

Après la représentation vectorielle de mots, nous utilisons un algorithme de clustering, afin de mettre les mots qui ont un sens sémantique ou syntaxique commun dans le même cluster (leurs vecteurs sont proches les uns aux autres vu une distance euclidienne ou distance de cosinus). Pour conforter nos résultats, nous avons testé deux algorithmes de clustering : notre implémentation [2] de Self-Organizing Maps (Kohonen, 1982) et Kmeans++ (Arthur et Vassilvitskii, 2007), qui a été implémenté dans Scikit-learn [3]. Nous proposons une mesure basée sur la similarité entre les synsets associés aux mots dans wordnet :

Simwords : nous calculons la similarité entre les mots de chaque cluster pris deux à deux en utilisant wup_similarity. Nous définissons $Simwords(C_i)$ la similarité entre les mots du cluster C_i par l'équation :

$$Simwords(C_i) = \frac{\sum_{k=1}^{n_{C_i}-1} \sum_{j=k+1}^{n_{C_i}} wup\_similarity(m_k, m_j)}{n_{C_i}(n_{C_i} - 1)/2} \tag{3}$$

Où les mots m_k et m_j appartiennent au clusters C_i

On calcule ensuite la moyenne des mesures de tous les clusters, par l'équation :

$$Simwords = \frac{\sum_{i=1}^{N} Simwords(C_i)}{N} \tag{4}$$

Où n_{C_i} est le nombre de mots du cluster C_i, N est le nombre de clusters (nous ne prenons pas les clusters singleton, qui contiennent un seul mot).

Pour utiliser WordNet avec NLTK, nous utilisons l'instruction :
from nltk.corpus import wordnet.
Pour obtenir l'ensemble des synonymes d'un mot donné en anglais w, nous utilisons l'instruction : *wordnet.synsets(w)*.

2. https ://gitlab.com/Data-Liasd/SOM
3. Scikit-learn est une bibliothèque libre Python dédiée à l'apprentissage automatique

En revanche, pour le français l'instruction : *wordnet.synsets(m, lang='fra')*, renvoie les synonymes de la traduction en anglais du mot français m (pour l'arabe, *lang = 'arb'*).

La Figure 1 donne une vue générale de la méthode proposée :

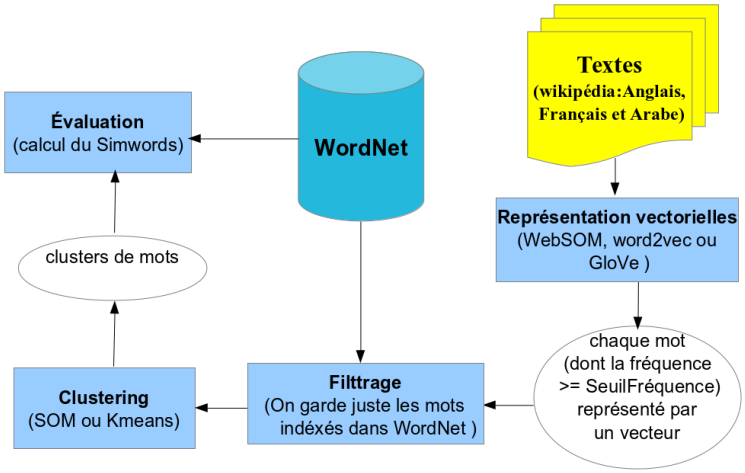


FIG. 1 – *Le principe de fonctionnement de notre système*

5 Corpus

Pour évaluer et comparer les trois méthodes de représentations vectorielles de mots choisies. nous avons sélectionné trois corpus textuels (Wikipédia dump 2017) en trois langues différentes : anglais, français et arabe. Les propriétés détaillées de ces corpus textuels sont indiquées dans le tableau 1.

Corpus	Nombre de mots	Vocabulaire : nombre de mots uniques	Fréquence de mots	Nombre de mots indexés dans WordNet
Wikipédia anglais 2017	2409291852	9045033	$\geq 600 = 96967$	47118
Wikipédia français 2017	804476834	4348227	$\geq 300 = 86697$	18600
Wikipédia arabe 2017	117472209	2140757	$\geq 300 = 33974$	1646

TAB. 1 – *Propriétés des corpus*

6 Expérimentations et résultats

Pour comparer les trois méthodes choisies, nous avons joué sur les paramètres suivants :

1. la taille de la fenêtre : 4, 6, 8, 10, 12 pour les trois méthodes,
2. dimension des vecteurs : 50, 100, 150, 200, 250. pour les trois méthodes,
3. pour controler le mot dans son contexte ϵ : 0.1, 0.3, 0.5 pour WebSOM,
4. negative sampling pour word2vec avec les deux modèles : Skip-gram et CBOW,
5. itération : 5, 10, 15 pour word2vec et GloVe,
6. normalisation L_2 des vecteurs : avec et sans pour les trois méthodes,
7. distance : distance euclidienne et distance cosinus pour les trois méthodes,
8. nombre de clusters : varie entre 150 et 3000, tout dépend du nombre de mots à classifier.

Nous avons choisi pour chaque méthode, les paramètres qui donnent le meilleur résultat (Simwords maximum). Pour le même corpus, on donne le même nombre de clusters pour les trois méthodes de vectorisation. les meilleurs résultats rapportés sont présentés dans le tableau 2 (corpus anglais), tableau 3 (corpus fançais) et tableau 4 pour le corpus arabe.

		SOM : 45x55 (2475 clusters)			Kmeans++ (2475 clusters)		
		Simwords	n_cp	n_cn	Simwords	n_cp	n_cn
word2vec	skip gram	**0.61627**	**101**	26	**0.60402**	**105**	21
	cbow	0.59336	98	15	0.58338	99	13
GloVe		0.54853	54	30	0.54950	51	22
WebSOM		0.47308	13	21	0.47244	17	16

TAB. 2 – *Corpus Wikipédia en anglais : évaluation des méthode de representations vectorielles de mots*

		SOM : 25x35 (875 clusters)			Kmeans++ (875 clusters)		
		Simwords	n_cp	n_cn	Simwords	n_cp	n_cn
word2vec	skip gram	0.50842	18	20	0.52058	11	15
	cbow	**0.52539**	**27**	26	**0.54128**	18	11
GloVe		0.52195	23	19	0.53662	**19**	14
WebSOM		0.46211	17	24	0.47218	10	17

TAB. 3 – *Corpus Wikipédia en français : évaluation des méthode de representations vectorielles de mots*

		SOM : 15x25 (375 clusters)			Kmeans++ (375 clusters)		
		Simwords	n_cp	n_cn	Simwords	n_cp	n_cn
word2vec	skip gram	0.37984	10	14	0.38112	11	8
	cbow	**0.42859**	**14**	14	**0.41221**	14	12
GloVe		0.39759	12	10	0.39841	13	14
WebSOM		0.37429	7	17	0.37233	**15**	12

TAB. 4 – *Corpus Wikipédia en arabe : évaluation des méthode de representations vectorielles de mots*

n_cp est le nombre de clusters dont tous les mots appartiennent au même synset, n_cn est le nombre de clusters dont les mots n'ont aucune relation sémantique ou syntaxique dans le WordNet. On peux constater que word2vec avec ses deux modèles, représente mieux les mots selon la mesure proposée, skip-gram est plus performant pour le corpus anglais et CBOW pour le français et l'arabe.
Nous ne prenons pas en compte les clusters singleton (simwords(cluster_singleton) = 1, car il contient un seul mot) dans le calcul de ces mesures. Car ceux-ci augmentent les valeurs sans pour autant signifier une bonne méthode de vectorisation de mots. Par exemple en segmentant les vecteurs produits par la méthode WebSOM, on obtient plusieurs clusters singletons (435/875), si nous les comptons, on trouve simwords = 0.73471 au lieu 0.47244. en revanche avec la méthode word2vec-skip-gram, on obtient 153 clusters singleton sur 875 clusters, qui donne simwords = 0.67325 au lieu 0.60402 sans les compter.

L'augmentation du nombre de clusters, augmente Simwords (similarité entre mots) (voir la figure 2)

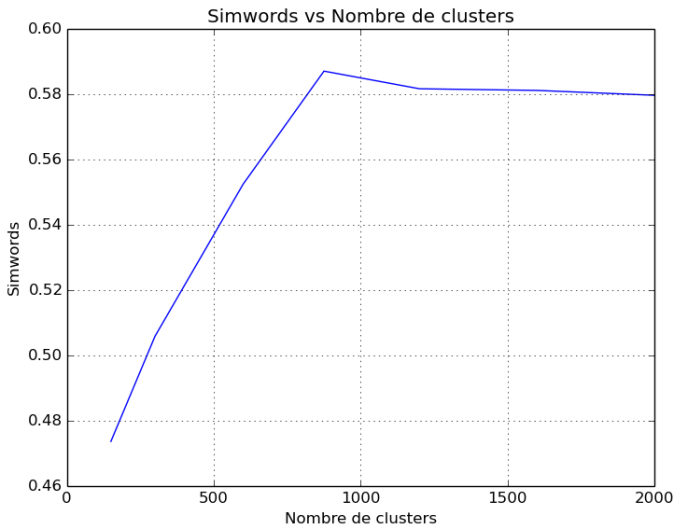


FIG. 2 – *Simwords vs Nombre de clusters : échantillon de 5000 vecteurs du corpus anglais, avec word2vec(CBOW), taille de la fenêtre = 8, dimension des vecteurs = 200.*

L'augmentation de la dimension des vecteurs pour word2vec et GloVe, augmente les performances (figure 3).

L'augmentation de la taille de la fenêtre pour word2vec-CBOW, influence négativement les performances. Et pour GloVe une fenêtre de 10 fait mieux que celles de 4 ou de 14 (figure 4)

Contrairement à (Levy et al 2015), la normalisation des vecteurs diminue les performances pour les trois méthodes. La distance euclidienne est mieux adaptée pour le clustering que la distance de cosinus (1-cosinus).

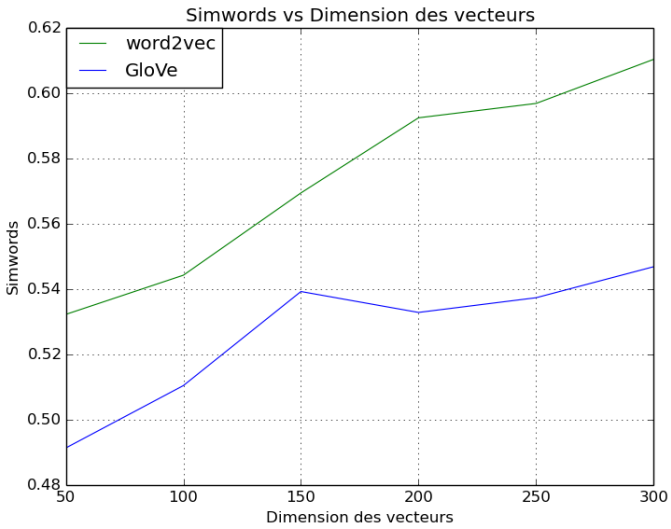


FIG. 3 – *Simwords vs Dimension des vecteurs : échantillon de 5000 vecteurs du corpus anglais, avec word2vec (CBOW) et GloVe, nombre de clusters = 875, taille de la fenêtre = 8.*

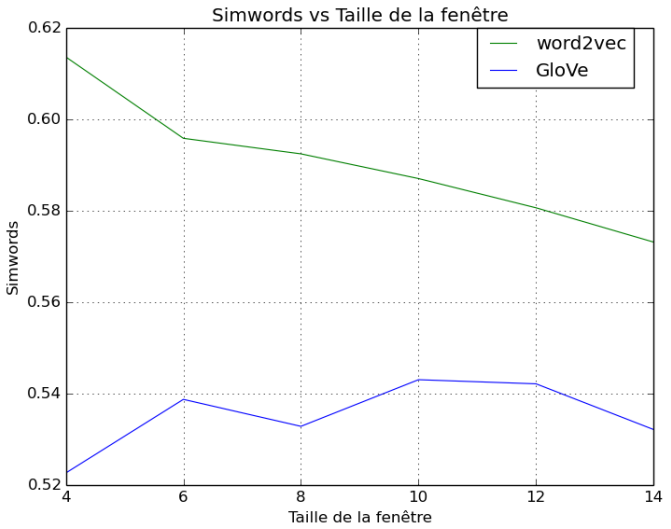


FIG. 4 – *Simwords vs Taille de la fenêtre :échantillon de 5000 vecteurs du corpus anglais, avec word2vec (cbow) et GloVe, nombre de clusters = 875, dimension des vecteurs = 200.*

Les trois méthodes produisent quelques clusters parfaits (dont les mots appartiennent au même synset) voici quelques exemples produits par word2vec (skip-gram) avec Wikipédia anglais : [comparison compare], [wide broad wider], [made making makes], [main primary principal] , [take took taken takes taking], [play played playing plays], [run running runs ran], [example instance], [almost nearly virtually].

7 Conclusion

Nous avons proposé une solution alternative aux méthodes d'évaluation actuelles, qui permet de mesurer les performances d'une méthode de representation vectorielle de mots ou d'optimiser les paramètres d'une méthode pour augmenter ses performances. Notre approche est plus générale vu la richesse sémantique des synonymes de WordNet et avec la possibilité de l'utiliser avec différentes langues. Il reste à étuder et comparer d'autres méthodes de vectorisations de mots traditionnelles au récentes comme celle proposée par (Lebboss, 2016) qui est fondée sur les travaux de (Bernard, 1997) sur l'importance des marqueurs grammaticaux dans le texte.

Références

Arthur, D. et S. Vassilvitskii (2007). K-means++ : the advantages of careful seeding. In *In Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*.

Baroni, M., G. Dinu, et G. Kruszewski (2014). Don't count, predict ! a systematic comparison of context-counting vs. context-predicting semantic vectors. *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference 1*, 238–247.

Bernard, G. (1997). Experiments on distributional categorization of lexical items with Self Organizing Maps. In *International Workshop on Self Organizing Maps WSOM'97*, pp. 304–309.

Harris, Z. S. (1968). *Mathematical Structures of Languages (Interscience Tracts in Pure and Applied Mathematics)*. USA : Interscience Publishers New York.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological cybernetics 43*(1), 59–69.

Kohonen, T., S. Kaski, K. Lagus, et J. Salojarvi (1998). Websom - self-organizing maps of document collection. *Helsinki University of Technologie, Finland*.

Landauer, T. K. et S. T. Dutnais (1997). A solution to platoâs problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *PSYCHOLOGICAL REVIEW 104*(2), 211–240.

Lebboss, G. (2016). *Contribution à l'analyse sémantique des textes arabes*. Ph. D. thesis, Université de Paris 8.

Levy, O., Y. Goldberg, et I. Dagan (2015). Improving distributional similarity with lessons learned from word embeddings. *TACL 3*, 211–225.

Mikolov, T., K. Chen, G. Corrado, et J. Dean (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.

Miller, G. A. (1995). Wordnet : A lexical database for english. *Commun. ACM 38*(11), 39–41.

Pedersen, T., S. Patwardhan, et J. Michelizzi (2004). Wordnet : :similarity : Measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, HLT-NAACL–Demonstrations '04, Stroudsburg, PA, USA, pp. 38–41. Association for Computational Linguistics.

Pennington, J., R. Socher, et C. D. Manning (2014). Glove : Global vectors forword representation. In *EMNLP*, Volume 14, pp. 1532–1543.

Wu, Z. et M. Palmer (1994). Verbs semantics and lexical selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, Stroudsburg, PA, USA, pp. 133–138. Association for Computational Linguistics.

Summary

Current evaluation methods for word representation use a small dataset. To overcome this problem, we present a new approach, based on the similarity between the synsets associated with words in the large lexical database WordNet. Our evaluation method consists first of all in arranging the vector representations of words in clusters by a clustering algorithm, then in evaluating the semantic and syntactic coherence of the clusters produced. This evaluation is performed by calculating the similarity between the words of each cluster taken two by two, using similarity measures. between the words in WordNet proposed by NLTK (path_similarity or wap_similarity). For each cluster, we obtain a value between 0 and 1, a cluster whose value is 1 is a cluster in which all the words belong to the same synset. The average of the measurements of all the clusters is then calculated. We used our new approach to study and compare three vector representation methods: a traditional WebSOM method and two recent methods, word2vec (Skip-Gram and CBOW) and GloVe, on three corpora: in English, French and Arabic. The results show that the word2vec method overpasses the other two methods on the three corpora.

Graph2Bots: Assistance automatisée à la conception d'agents dialoguants

Jean-Leon Bouraoui*, Vincent Lemaire*

*2 Avenue Pierre Marzin, 22300 Lannion
{jeanleon.bouraoui,vincent.lemaire}@orange.com,
http://vincentlemaire-labs.fr/

Résumé. Nous décrivons la démonstration d'un prototype permettant la modélisation non supervisée de la structure de dialogues finalisés; ces dialogues appartiennent à un domaine donné (par exemple réservation de trains). Ceci présente de nombreux intérêts, le principal étant de servir de base à la conception de l'architecture d'un agent dialoguant. Un graphe modélise les principales étapes des dialogues et les transitions entre elles. La technique adoptée consiste à appliquer du CoClustering sur le corpus cible de dialogues, afin d'obtenir les principaux thèmes qui y figurent. On calcule ensuite les transitions entre thèmes dans chaque dialogue. Notre outil permet d'obtenir le graphe correspondant et de le manipuler de manière ergonomique. Nous présentons en détail les différentes fonctionnalités démontrées.

1 Introduction

En intelligence artificielle, les agents dialoguants connaissent un gain de popularité auprès du grand public ; et ce d'autant plus qu'ils bénéficient des avancées dans la compréhension des contenus et du contexte. Cela est le fait notamment d'applications mobiles telles que Siri (Apple), Google Now (Google), ou Cortana (Microsoft) ou Alexa (Amazon). Cet intérêt grandissant pour la technologie des interfaces dialoguantes et des agents dialoguants en particulier est décrit par de nombreuses études telles que celle du cabinet d'analyse Gartner [1].

Une des tendances actuelles est de proposer des dispositifs logiciels de conception d'agents dialoguants, personnalisables selon les besoins et le domaine d'application (par exemple, réservation de voyages, commande de produits ou de services, etc.). L'un des enjeux de ces dispositifs est de pouvoir être mis en place rapidement, sachant qu'il n'existe actuellement pas de système générique, et qu'une adaptation de l'agent à un domaine d'application donné prend du temps.

Dans ce contexte nous présentons une solution d'assistance semi-automatique à la création ou l'adaptation d'un agent dialoguant pour un domaine applicatif donné. Dans un premier temps, nous décrivons la problématique abordée. Nous présentons ensuite notre solution, d'abord du côté backend, puis de l'interface utilisateur correspondante (frontend), telles qu'elles seront présentées à l'occasion de la démonstration.

1. http://www.gartner.com/smarterwithgartner/top-trends-in-the-gartner-hype-cycle-for-emerging-technologies-2017

2 Description de la problématique

Dans le cadre de notre démonstration, nous appellerons dialogue un échange d'informations entre deux interlocuteurs (sachant qu'un dialogue peut faire intervenir plus de deux interlocuteurs). Un interlocuteur peut être un humain ou une machine (au sens large : un système artificiel, logiciel ou matériel). Nous nous intéressons aux dialogues finalisés, qui cherchent à atteindre un but : les interlocuteurs vont collaborer pour l'atteinte de ce but.

On appelle dans cet article "corpus textuel" un ensemble de n dialogues relatifs à un domaine particulier, (par exemple transcriptions de dialogues de réservations de trains, ou tchats d'interactions entre un téléconseiller et un client). Chaque dialogue est composé de t tours de parole, un tour de parole correspondant à ce que dit l'un des interlocuteurs sans interruption (la plupart du temps, une ou plusieurs phrases).

Notre but est de déterminer automatiquement, et ce au sein de chaque dialogue : (i) les différentes phases du dialogue (incluant notamment les intentions exprimées ; nous les désignerons désormais par le terme "thème", qui correspond aussi bien à des thématiques génériques qu'à des sous-buts du dialogue) ; (ii) les transitions entre les phases. Le but est d'obtenir une représentation du déroulement typique des dialogues du corpus. La représentation souhaitée est celle d'un graphe orienté montrant les principales transitions entre thèmes, comme celui représenté (et simplifié) sur la figure 1. Notre postulat est que, selon la position dans le dialogue, un tour de parole donné présente plus de probabilité d'appartenir à une phase donnée (i.e. un cluster), qu'un autre ; cette information est donc prise en compte lors du processus.

Le graphe ainsi obtenu présente de multiples intérêts. Le principal est l'initialisation de la conception de l'agent dialoguant : il pourra servir de base à la modélisation de l'architecture d'un agent dialoguant spécialisé sur le domaine cible, et ainsi en faciliter l'exécution. A l'heure actuelle, cette tâche est la plupart du temps effectuée manuellement : soit a priori, à partir de la représentation que le concepteur se fait des dialogues possibles portant sur une tâche et un domaine donné ; soit a posteriori, à partir de la consultation de corpus existants ; dans les deux cas, le processus est coûteux en temps.

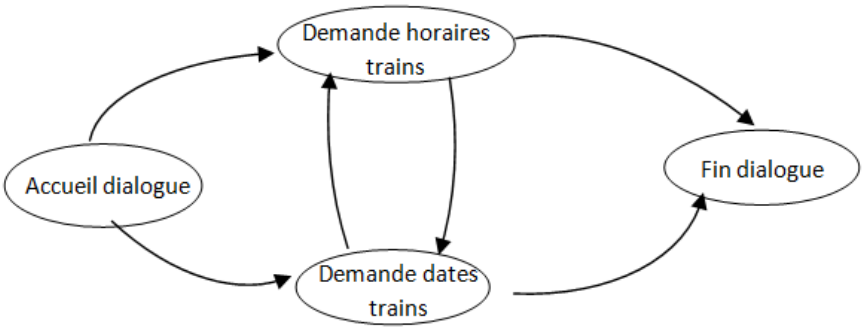


FIG. 1 – *Représentation graphique des principales transitions entre thèmes*

De plus, le graphe, ainsi que les étapes parcourues pour l'obtenir, permettra au concepteur, sans connaissance préalable du domaine d'application, d'avoir une première compréhension du contenu thématique des dialogues, de leur structuration, et plus généralement la connaissance des informations les plus pertinentes pour la réalisation de l'agent dialoguant.

3 Description de la démonstration

Un exemple prototypique d'utilisation de notre outil est le suivant : un concepteur souhaite mettre en place un agent dialoguant relatif pour un domaine d'application spécifique. Il dispose de corpus de dialogues relatifs à ce domaine. Dans un premier temps, il va utiliser le backend de notre outil pour identifier, de manière non supervisée et sans annotation préalable, les phases de dialogues sous-jacentes, et leurs transitions. Il va ensuite manipuler les graphes obtenus à l'aide de l'interface visuelle fournie. Nous décrivons ces différentes étapes ci-dessous.

3.1 Approche pour le backend

Au préalable, les mots du corpus sont filtrés pour supprimer les mots dits "outils" ou "creux", a priori inutiles pour le traitement. Le filtre utilisé est la liste de "stopwords" français utilisés pour le stemming par la librairie NLTK [2]. A moyen terme, nous envisageons d'autres prétraitements des mots du corpus, notamment leur lemmatisation.

Pour identifier les phases de dialogue, on utilise une technique de CoClustering qui permet d'obtenir une "copartition" de la matrice mots x tour de paroles. Étant données, deux (ou plus) variables catégorielles ou numériques, on réalise un partitionnement simultané des variables : les valeurs de variables catégorielles sont groupées en clusters et les variables numériques sont partitionnées en intervalles – ce qui revient à un problème de coclustering. La méthode employée est basée sur l'approche MODL décrite dans (Boullé et al., 2014). Il reste ensuite à déterminer les transitions entre les phases de dialogue.

Dans notre approche, une phase correspond à un cluster de tours de parole, homogènes par rapport à une thématique donnée. Une phase définie ainsi peut être reliée à une ou plusieurs autres, en fonction de la fréquence observée de leurs successions dans les dialogues du corpus.

Le CoClustering fait initialement perdre la séquentialité des dialogues, puisque les tours de paroles sont regroupés par thèmes, indépendamment de leur ordre dans le dialogue. Pour retrouver cet aspect temporel, le backend projette ensuite les identifiants de cluster sur chaque tour de parole correspondant. La représentation obtenue est un graphe orienté, dont les nœuds sont les clusters, et les arcs sont les successions entre clusters.

2. `http://www.nltk.org/nltk_data`

3.2 Présentation du frontend

L'interface utilisateur permet de visualiser interactivement et en temps réel les données traitées en backend sous la forme de graphes de dialogue. Nous en décrivons les principales caractéristiques ci-dessous.

3.2.1 Affichage interactif et en temps réel

Le concepteur de l'agent dialoguant peut observer et analyser le graphe obtenu. Les principales fonctionnalités appartenant à cette catégorie sont les suivantes :

— Choix de la granularité d'affichage du graphe, selon le nombre de clusters et/ou de leur relations, et modification dynamique du graphe correspondant ;
— Possibilité de manipuler les graphes avec le pointeur de la souris : par exemple pouvoir « tirer » un cluster à l'écart des autres, sélectionner un ou plusieurs clusters, zoomer et dézoomer, etc.
— Affichage des noms de clusters [3], de leur nombre, et de la fréquence de chaque relation (avec éventuellement affichage du pourcentage sur l'ensemble des dialogues)
— Possibilité de renommer les clusters, et de rajouter des noms aux relations entre clusters ;
— Possibilité d'afficher le contenu d'un cluster, composé de plusieurs tours de parole. Ceux-ci peuvent être affichés sous la forme d'une liste, avec éventuellement des informations supplémentaires.

3.2.2 Manipulation de l'architecture du dialogue

Ces fonctionnalités permettent de modifier l'architecture locale ou globale du graphe. Le concepteur de l'agent dialoguant peut ainsi adapter et raffiner l'architecture obtenue en fonction de ses besoins. Il est ainsi possible de modifier :

— Le contenu d'un cluster donné. Notamment en supprimant un ou plusieurs tours de parole qui ne seraient pas homogènes thématiquement avec le cluster.
— L'architecture elle-même. Deux principales fonctionnalités sont utilisables. D'une part la fusion de deux clusters (par exemple si ces clusters sont similaires thématiquement et donc redondants). D'autre part la possibilité de sélectionner plusieurs tours de paroles d'un cluster donné, pour créer un nouveau cluster (ce qui revient à scinder en deux le cluster courant) ; cela est utile dans le cas où les tours de parole en question sont similaires sémantiquement, mais hétérogènes par rapport à la principale thématique exprimée dans le cluster. Si l'une de ces fonctionnalités est utilisée,l'affichage du nombre de clusters et de leurs relations est mis à jour.

Nous avons utilisé ce prototype pour plusieurs usescase réels, dont celui décrit dans (Bouraoui et Lemaire, 2017). La figure 2 présente la version actuelle du frontend du prototype. Il sera présenté plus en détails lors de la démonstration.

3. Le nom attribué initialement à un cluster par le backend est une suite non significative de caractères.

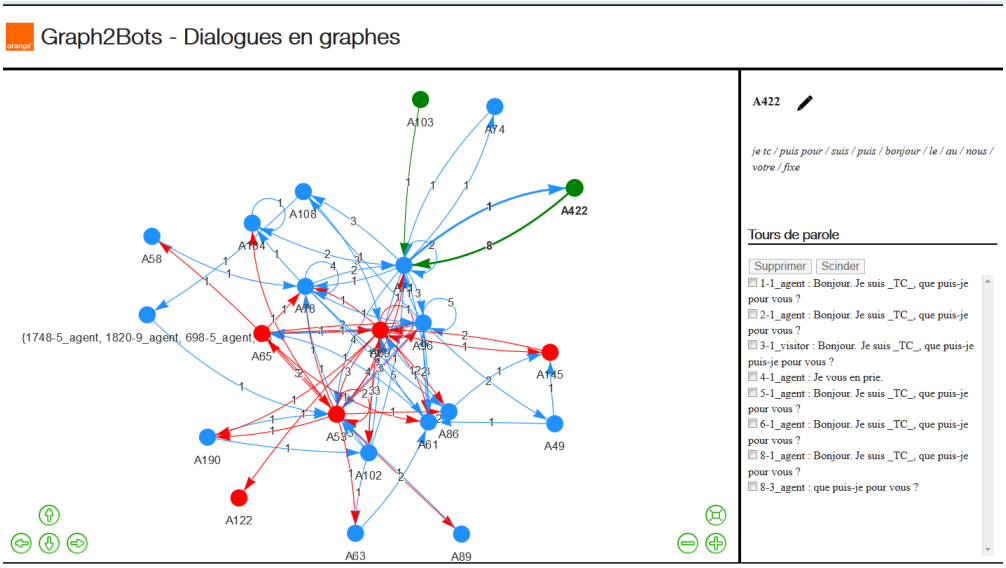


FIG. 2 – *Graphe pour la conception d'architecture*

Références

Boullé, M., R. Guigourès, et F. Rossi (2014). Analyse exploratoire par k-coclustering avec khiops coviz. In *Advances in Knowledge Discovery and Management*, Volume 527, pp. 15–35.

Bouraoui, J.-L. et V. Lemaire (2017). Cluster-based graphs for conceiving dialog systems. In *Workshop DMNLP at European Conference on Machine Learning (ECML)*.

Summary

We demonstrate a prototype allowing the unsupervised modeling of the structure of task-oriented dialogues. The dialogues are related to a given domain (for instance, train booking). This presents several advantages; notably its use as a basis for conceiving a conversational agent architecture. The modeling is represented by a graph. It displays the main stages of the dialogues and the transitions between them. Our approach consists in applying the coclustering to the targeted dialogue corpus. Thus we obtain the main themes that appear in the corpus. We then compute the theme transitions within each dialogue. Our tool allows to obtain and manipulate the corresponding graph. We detail the various fonctionalities demonstrated.

Index