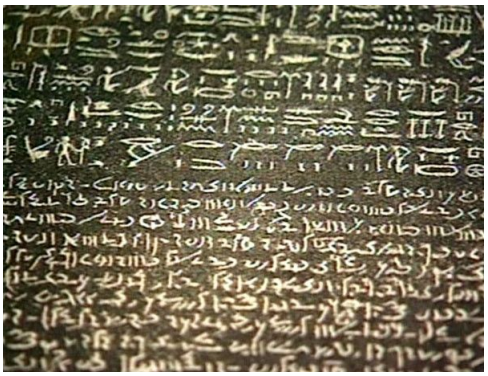
TextMine

Atelier sur la Fouille de Textes



Organisateurs :

Pascal Cuxac (INIST - CNRS),
Vincent Lemaire (Orange Labs),
Jean-Charles Lamirel (Loria)

Organisé conjointement à la conférence EGC
(Extraction et Gestion des Connaissances)
le 24 janvier 2017 à Grenoble

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

Jean-Charles Lamirel - LORIA - SYNALP Research Team
Campus Scientifique, BP. 239, 54506 Vandoeuvre les Nancy Cedex
Email : jean-charles.lamirel@loria.fr

Publisher:

Vincent Lemaire, Pascal Cuxac, Jean-Charles Lamirel
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...
En France, des conférences comme TALN, CORIA, JADT par exemple sont centrées sur l'analyse et le traitement des textes, mais avec des approches plus ciblées soit TAL, soit RI, soit statistiques. Cet atelier se veut plus fédérateur autour d'approches et d'applications aussi diverses que possibles.
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
J.-Ch. Lamirel — Loria

Membres du comité de lecture

Le Comité de Lecture est constitué de:

Patrice Bellot (LSIS, Marseille)
Guillaume Cabanac (IRIT, Toulouse)
Martine Cadot (Loria, Nancy)
Mariane Clausel (LJK , Grenoble)
Vincent Claveau (IRISA, Rennes)
Guillaume Cleuziou (LIFO, Orléans)
Dominique Gay (U. Réunion, Saint Denis de la Réunion)
Natalia Grabar (STL - Lille3, Lille)
Brigite Grau (LIMSI, Orsay)
Mustapha Lebbah (LIPN, Paris)
Denis Maurel (LIT, Tours)
Patrick Paroubeck (LIMSI, Orsay)
David Reymond (i3N , Toulon - Nice)
Mathieu Roche (LIRMM, Montpellier)
Jacques Savoy (U. Neuchatel, Neuchatel, Suisse)
Isabelle Tellier (Sorbonne, Paris)
Julien Velcin (ERIC, Lyon)

TABLE DES MATIÈRES

Exposé Invité

Extension d'un corpus d'articles scientifiques par recherche de similarités sémantiques : application à une problématique des sciences du sport
Fabrice Muhlenbach 1

Session Exposés

Enhanced Verbatim Analysis (EVA) : Une chaîne d'analyse sémantique de verbatims
Aleksandra Guerraz, Nathalie Legay, Rémi Bars 3

Co-clustering pour la fouille de textes : le package CoClust
François Role, Stanislas Morbieu, Mohamed Nadif 9

Aide à l'automatisation de conception de systèmes de dialogue
Jean-Leon Bouraoui, Vincent Lemaire 13

Exploratory Text Segmentation through Joint Distribution Estimation
Dominique Gay, Romain Guigourès, Marc Boullé, Fabrice Clérot 23

Préliminaire à la construction d'un réseau de signalisation en biologie systémique
F. Landomiel, A. Gupta, D. Maurel, A. Poupon 31

Archives numériques et fouille de textes : le projet ISTEX
Pascal Cuxac, Nicolas Thouvenin 43

Index des auteurs **53**

Extension d'un corpus d'articles scientifiques par recherche de similarités sémantiques : application à une problématique des sciences du sport

Fabrice MUHLENBACH

Univ. Lyon, UJM-Saint-Etienne, CNRS,
Laboratoire Hubert Curien UMR 5516, F-42023 Saint Etienne, France
fabrice.muhlenbach@univ-st-etienne.fr

La recherche scientifique menée dans les milieux académiques ou industriels est productrice de connaissances toujours plus nombreuses. Les chercheurs, pour apporter leurs contributions originales à ce savoir, doivent pouvoir accéder aux connaissances existantes (publications, brevets) et retrouver parmi elles les éléments pertinents (théories, démonstrations, méthodologies, résultats expérimentaux...) qui servent de socle à leurs travaux et permettent de délimiter les cadres de leurs propres apports scientifiques.

À l'heure actuelle, les publications scientifiques sont souvent accessibles soit directement sur des sites *web* de chercheurs, d'institutions, d'entreprises réalisant de la R & D ou d'éditeurs scientifiques, soit à travers des plates-formes spécifiques tels que les projets *Gallica*, *Google Livres* ou *ISTEX* [1]. Ces bibliothèques numériques sont à la fois l'instrument et la matière première de la recherche et de l'innovation scientifique. Par conséquent, la maîtrise et l'utilisation efficace de ces sources constituent un enjeu stratégique pour le développement de la science, l'accroissement de la richesse économique et plus largement l'évolution de la société.

Néanmoins, l'exploration de ces gigantesques bibliothèques numériques n'est pas efficace en raison notamment des limitations cognitives humaines et du manque de temps. Il en résulte une exploitation faible de la richesse des bibliothèques numériques avec une focalisation à la fois sur les articles les plus récents (alors que des articles anciens pourraient pourtant s'avérer pertinents) et des articles limités à la communauté scientifique d'appartenance du chercheur (alors que des articles venant de disciplines complémentaires pourraient être intéressants).

Cette présentation propose plusieurs pistes menées par une équipe de recherche stéphanolyonnaise afin de répondre à ce problème. À partir d'un corpus de base de quelques textes scientifiques de référence concernant un sujet donné, au moyen de techniques de fouille de textes, nous montrons qu'il est possible d'obtenir d'autres articles pertinents (dans notre cas, issus de la base ISTEX) liés au sujet du corpus d'entrée par des relations de similarités sémantiques. De plus, ces articles scientifiques nouveaux extraits des bibliothèques numériques viennent enrichir le corpus de départ en étant puisés dans des disciplines allant au-delà du seul domaine des articles fournis en exemple. Nous illustrons cette proposition par une problématique intéressant les sciences du sport, domaine par nature pluridisciplinaire puisque les pratiques sportives constituent un objet d'étude qui peut être abordé par différentes disciplines, en particulier sous les angles de la physiologie, de la psychologie ou de psycho-sociologie.

1. ISTEX : Initiative d'excellence de l'Information Scientifique et Technique, voir : `http://www.istex.fr/`

Enhanced Verbatim Analysis (EVA)
Une chaîne d'analyse sémantique de verbatims

Aleksandra Guerraz, Nathalie Legay, Rémi Bars

Orange Labs
2, avenue Pierre Marzin, 22307 Lannion Cedex, France
{aleksandra.guerraz,nathalie.legay,remi.bars}@orange.com

Résumé. Cette démonstration présente **EVA** (**E**nhanced **V**erbatim **A**nalysis), une application web qui permet d'explorer et de classifier les verbatims écrits par les clients (sondage, SAV, médias sociaux...). Simple à utiliser, cette application est destinée aux utilisateurs des entités « métiers » telles que le Marketing Produit, la Satisfaction Client, la Qualité des Offres... Elle ne nécessite aucune expertise en statistiques, linguistique ou sémantique. Dans cet article de démonstration EVA est présentée, de manière illustrative, à travers des données issues de tchats entre clients et téléconseillers.

1 Introduction

Aujourd'hui des centaines de milliers de verbatims sont écrits par des clients ou des téléconseillers lors des interactions avec les nombreuses entités d'un grand groupe comme par exemple le Groupe Orange (sondages, appels SAV, enquêtes...). Bien qu'étant une mine d'information, ces verbatims restent encore très peu exploités. L'une des voies possibles pour « comprendre » la teneur du contenu de ces verbatims est l'utilisation d'outils de classification statistique et/ou d'analyse sémantique mais qui nécessitent des compétences pointues en linguistique, statistiques et informatique.

La compréhension en temps réel des milliers de verbatims par les entités « métier » permettrait de mettre en oeuvre très rapidement des actions marketing, commerciales et RH appropriées, sans faire appel à des spécialistes rares et coûteux. Cela suppose de disposer d'un outil simple à utiliser par des non-experts. L'application EVA répond à ce besoin et permet très facilement de charger un corpus de verbatims, de les analyser et de présenter des résultats lisibles immédiatement. L'application est actuellement utilisée par plusieurs entités « métiers ».

2 Classifier des verbatims avec EVA

L'application EVA permet de produire un modèle de classification à base de règles en s'appuyant sur une analyse exploratoire. Pour cela, EVA intègre deux outils spécialisés :

– Un outil de data mining dédié au co-clustering : Khiops Coclustering décrit dans (Boullé, 2011),

– Un outil d'analyse sémantique qui permet de mettre au point un ensemble de règles sémantiques : Semantic Analyzer de Disserto décrit dans (Laroche, 2010).

Pour classifier les verbatims avec EVA, l'utilisateur devra passer par trois grandes étapes. La première étape est celle de l'exploration des données où l'utilisateur crée automatiquement une partition des données à l'aide de Khiops Coclustering, puis prend le temps d'analyser les co-clusters construits par l'outil. Il peut affiner cette partition (les co-clusters découverts) et la valider dans la deuxième étape appelée étape d'expérimentation. Finalement, l'utilisateur peut déployer le modèle obtenu sur d'autres corpus (provenant de la même source, par exemple, les réponses à un sondage mensuel) dans l'étape d'exploitation.

2.1 Exploration de données

L'étape d'exploration est effectuée à travers un co-clustering à deux dimensions $(Textes \times Mots)$. On obtient alors des groupes de verbatims définis par le fait qu'ils contiennent des mots similaires en commun. Le co-clustering obtenu est hiérarchique et peut être visualisé dans un outil expert présenté dans (Guerraz et al., 2016). Afin de simplifier l'exploration de données, dans EVA, un seul niveau de hiérarchie est présenté à l'utilisateur. Ce niveau peut comporter entre dix et quarante groupes de verbatims.

Dans EVA, le co-clustering est visualisé sous forme de bulles. Les textes sont regroupés en catégories et présentés à l'utilisateur sur une carte sous forme de bulles bleues. Les mots similaires sont regroupés en groupes de mots et visualisés sur la carte sous forme de bulles rouges. La figure 1 présente la visualisation du co-clustering obtenu sur un corpus de tchat entre clients et téléconseillers (agents).

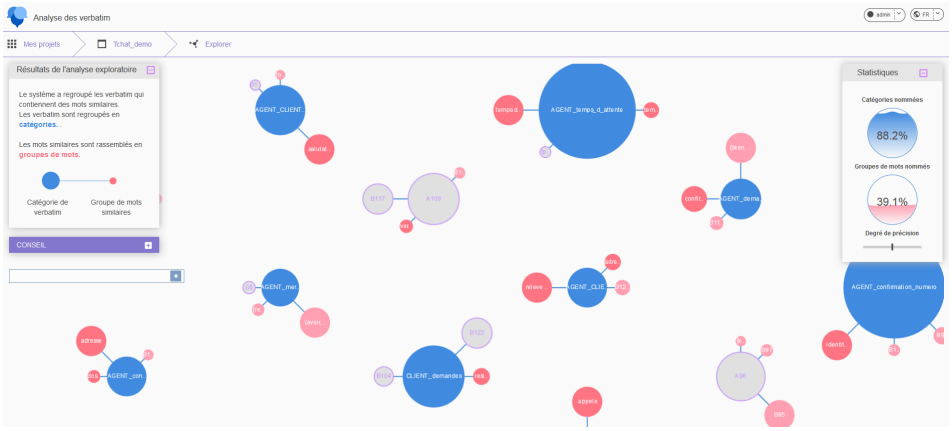


FIG. 1 – *Visualisation du co-clusetring dans EVA*

Pour chaque catégorie, on affiche à l'utilisateur les trois premiers groupes de mots les plus informatifs selon un critère dérivé de l'information mutuelle (se référer à (Guigourès, 2013) pour une définition complète du critère). Pour chaque groupe de mots on édite les dix premiers mots les plus typiques du groupe. Avec la connaissance de son domaine « métier » l'utilisateur peut nommer et valider les groupes de textes et les groupes de mots qui ont du sens pour lui

en examinant leurs contenus. Les mots les plus typiques pour une catégorie donnée facilitent le nommage. Les catégories nommées et marquées par l'utilisateur comme ayant du sens sont retenues dans le modèle de classification final.

2.2 Création de règles

Dans l'étape d'expérimentation, le résultat du regroupement automatique peut être amélioré à l'aide de règles sémantiques présentes dans Disserto. Ces règles permettent d'affiner le périmètre d'une catégorie d'une façon déterministe. Les regroupements réalisés par le co-clustrering permettent d'initier la construction d'un modèle de classification à base de règles qui sera appliqué pour classer les verbatims. Les règles sémantiques de Disserto sont des associations et combinaisons des étiquettes sémantiques. Les étiquettes sémantiques regroupent l'ensemble des mots significatifs pour un corpus. Une étiquette sémantique regroupe différents mots. Ainsi, les mots tels que « téléphone », « téléphoner », « téléphonique » peuvent appartenir à une seule étiquette sémantique. Une étiquette sémantique peut regrouper des synonymes, les différentes formes dérivationnelles d'un mot, etc... La combinaison des étiquettes sémantiques est faite par l'intermédiaire de trois opérateurs ET, OU, et PUIS (l'ordre des mots est important).

Dans EVA, pour chaque catégorie une première règle est automatiquement générée en se basant sur la combinaison des mots faisant partis des groupes de mots les plus représentatifs de la catégorie. L'utilisateur peut ajouter ses propres règles pour améliorer le résultat du regroupement automatique. Les règles permettent de retravailler les catégories dans le cas où le regroupement obtenu avec le co-clustering ne correspond pas pleinement aux attentes de l'utilisateur. La figure 2 présente l'ensemble des règles pour la catégorie « AGENT accès au dossier ». La première règle a été générée automatiquement et peut se lire de la manière suivante : si le texte contient un mot du groupe de mots « dossier » et un mot du groupe de mots « vous prie de patienter » alors le texte sera classé dans la catégorie « AGENT accès au dossier ».

L'application de l'ensemble des règles obtenues permet de classifier les verbatims dans les catégories définies par l'utilisateur. L'utilisateur peut suivre la performance du modèle (en termes de précision) en passant par l'étape de validation. Cette étape consiste à valider, catégorie par catégorie, la classification réalisée avec les règles. Quand la classification est satisfaisante, le modèle de classification créé peut être appliqué sur d'autres verbatims, dans l'étape d'exploitation.

3 L'interface utilisateur

Une conception orientée « utilisateur » a permis de réaliser une interface web simple et intuitive, qui masque aux utilisateurs la complexité des nombreux traitements effectués par le système sur les verbatims (prétraitements linguistiques, traitements statistiques et sémantiques...). Le domaine de l'optimisation des Interfaces Homme-Machine (IHM) est souvent lié à une problématique de communication entre l'humain et le système. Il est primordial d'orienter la conception des IHM en fonction des futurs utilisateurs. Une conception orientée « utilisa-

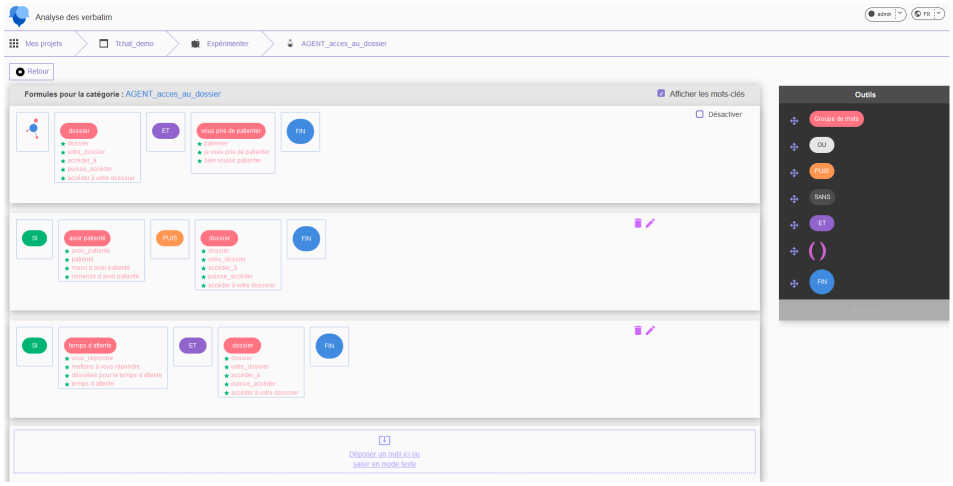


FIG. 2 – *Règles pour la catégorie « AGENT accès au dossier »*

teurs » s'appuie sur deux critères majeurs : l'utilisabilité [1] et l'utilité [2] (définis par la norme ISO 9241-210 [3]), afin que les IHM présentées aux utilisateurs leur permettent de mieux atteindre les objectifs qui leurs sont fixés, en leur apportant une meilleur satisfaction.

Les grandes étapes de la conception d'EVA ont été les suivantes :

- analyse des besoins : caractéristiques des futurs utilisateurs (compétences, habitudes, âge, attentes...), contexte d'utilisation, tâches confiées aux utilisateurs avec l'analyse d'une situation existante de référence (utilisation de différents logiciels non intégrés). Les nombreuses tâches fastidieuses et complexes à réaliser, dans la situation de référence, ont été volontairement supprimées ou bien rendues invisibles (explicites-expliquées à l'utilisateur ou bien implicites-cachées à l'utilisateur car jugées inutiles). Les tâches simples restent à réaliser par l'utilisateur : chargement du corpus de verbatims, lancement de l'analyse automatique, éventuellement amélioration de l'analyse, avant d'afficher les résultats de classification des verbatims.
- veille sur l'existant : ergonomie des IHM (Boucher, 2011) d'applications existantes, ergonomie des termes utilisés (pour simplifier les termes complexes), veille sur l'analyse sémantique et en design graphique avec datavisualisation.
- production de solutions de conception : architecture d'information (structuration et hiérarchisation des contenus, arborescence simple de l'application) et agencement des contenus et des commandes, production de gabarits d'écrans avec un design graphique simple et épuré, réalisation des écrans finaux).

1. L'utilisabilité est le degré selon lequel un produit peut être utilisé, par des utilisateurs identifiés, pour atteindre des buts définis avec efficacité, efficience et satisfaction, dans un contexte d'utilisation spécifié
2. Les fonctionnalités proposées par une application web doivent être utiles et servir des besoins. Un système peut respecter tous les critères d'utilisabilité mais être inutile. C'est l'adéquation entre l'activité et l'application qui permet de dire que l'outil est utile.
3. ISO 9241-210 (2010) : Ergonomics of human-system interaction – Part 210 : Human-centred design for interactive systems

– évaluation des solutions mises en oeuvre auprès d'utilisateurs réels, en situation réelle, dans des contextes de travail précis (RH, marketing, opérationnel...).

L'objectif était de fournir un outil-métier utilisable très facilement par tout utilisateur connaissant bien son domaine métier (par exemple marketing produit, qualité ciblée d'une offre...), et disposant de milliers de verbatims à analyser.

Références

Boucher, A. (2011). *Ergonomie web : Pour des sites web efficaces*. Paris : Eyrolles.

Boullé, M. (2011). Data grid models for preparation and modeling in supervised learning. In I. Guyon, G. Cawley, G. Dror, et A. Saffari (Eds.), *Hands-On Pattern Recognition: Challenges in Machine Learning*, Volume 1, pp. 99–130. Microtome Publishing.

Guerraz, B., M. Boullé, D. Gay, V. Lemaire, et F. Clérot (2016). Analyse exploratoire par k-coclustering avec khiopscoviz. *EGC RNTI-E-30*, 493–498.

Guigourès, R. (2013). *Utilisation des modèles de co-clustering pour l'analyse exploratoire des données*. Thèse de doctorat, Université Paris 1 Panthéon-Sorbonne.

Laroche, R. (2010). *Raisonnement sur les incertitudes et apprentissage pour les systèmes de dialogue conventionnels*. Thèse de doctorat.

Summary

The **EVA** (**E**nhanced **V**erbatim **A**nalysis) semantic analysis system is user friendly web application for the verbatim exploration and classification. The interface is intuitive and does not require any expertise in statistics, linguistics or semantics. EVA can be used by users from different Business Units (e.g. Product Marketing, Quality of Customer Service...). In this article EVA is presented through data from chat sessions between clients and agents.

Co-clustering pour la fouille de textes : le package CoClust

François Role * Stanislas Morbieu* Mohamed Nadif*

*LIPADE, Université Paris Descartes, 75006 Paris, France
prénom.nom@parisdescartes.fr

Résumé. En fouille de textes, la classification croisée ou *co-clustering* sert à analyser des matrices document-terme pour créer simultanément des ensembles de lignes (documents) et des ensembles de colonnes (termes). Le package Python **CoClust** fournit des implémentations de différents algorithmes dédiés à la tâche du *co-clustering*. Les composants du package respectent une interface homogène de manière à faciliter l'utilisation et la comparaison entre les différents algorithmes.

1 Introduction

Pour prendre en compte le volume toujours croissant de données numériques, les techniques de classification automatique (*cluster analysis*) sont plus que jamais à l'ordre du jour par leur capacité à regrouper des entités de nature variée (documents, gènes, clients, etc.). Cependant, si un assez grand nombre d'implémentations sont disponibles pour effectuer un *clustering*, il n'en va pas de même pour la classification croisée (*co-clustering*) de matrices de co-occurrences, comme par exemple les matrices document-terme utilisées en fouille de textes. Dans ce domaine, il n'existe pas de bibliothèque proposant plusieurs types d'algorithmes alternatifs. A titre d'exemple, le package Python **scikit-learn** (Pedregosa et al., 2011) supporte seulement deux algorithmes de *co-clustering* spectral : (1) le bien connu "Spectral Co-Clustering" (Dhillon, 2001) et (2) le "SpectralBiclustering" (Kluger et al., 2003) qui est également disponible dans le package R **biclust**.

Le package **CoClust** présenté dans cet article ambitionne donc de fournir une série d'implémentations d'algorithmes alternatifs conçus pour traiter efficacement de telles matrices. Dans sa première version, **CoClust** propose des implémentations d'algorithmes de *co-clustering* relevant de deux approches différentes. Les premières considèrent une matrice document-terme comme un graphe biparti et utilisent la modularité de graphe, adapté au cas biparti, comme critère pour effectuer un *co-clustering*. Les secondes sont basées sur la théorie de l'information. Dans ce type de méthodes, une distribution de probabilité jointe est tout d'abord dérivée de la matrice des co-occurrences. La fonction de coût à minimiser est alors la perte d'information mutuelle entre cette distribution de probabilité jointe et une distribution définie sur un tableau de contingence réduit, obtenu en fusionnant les lignes et les colonnes selon les partitions produites par l'algorithme de *co-clustering*. Des études comparatives récentes (Ailem et al., 2016) ont clairement montré que des implémentations basées sur ces méthodes sont beaucoup plus efficaces pour les textes que les méthodes spectrales souvent utilisées. Par ailleurs, les algorithmes inclus dans **CoClust** ont été choisis pour permettre d'effectuer un *co-clustering* diago-

nal ou un *co-clustering* non diagonal. Les algorithmes de *co-clustering* diagonal cherchent à découvrir une structure diagonale (présentée à l'utilisateur sous la forme de blocs diagonaux) ce qui facilite l'interprétation puisqu'à un groupe de documents est automatiquement associé un groupe de termes. Evidemment, la recherche de diagonalité impose qu'on ait demandé au départ un même nombre de *clusters* de documents et de *clusters* de termes. Les algorithmes non diagonaux n'ont pas cette contrainte : le nombre de groupes de documents demandé peut être différent du nombre de groupes de termes, mais c'est alors à l'utilisateur de décider des appariements entre groupes de documents et groupes de termes. Cette différence est illustrée par la figure 1 où les lignes (documents) et les colonnes (mots) ont été réorganisées selon les partitions renvoyées par un algorithme diagonal et un algorithme non diagonal inclus dans **CoClust**. Une telle visualisation comparative, en blocs (*co-clusters*) homogènes, peut être facilement obtenue avec **CoClust**, ce qui est dans l'esprit de ce package : permettre de comparer facilement les performances de différents algorithmes de *co-clustering* sur les textes. Tous les algorithmes ont ainsi été implémentés en respectant une interface homogène, compatible avec la bibliothèque **scikit-learn**, ce qui encourage à explorer les différentes méthodes et facilite la reproductibilité des expériences.

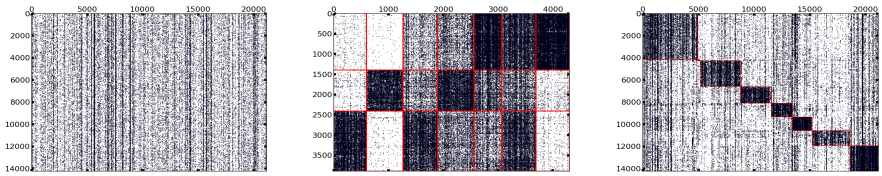


FIG. 1 – *(a) Données d'origine - (b) Données réorganisées avec un co-clustering non diagonal - (c) Données réorganisées avec un co-clustering diagonal.*

2 Algorithmes disponibles pour la fouille de textes

Dans sa version actuelle, **CoClust** propose trois algorithmes de *co-clustering*. Les deux premiers, `CoclustMod` et `CoclustSpecMod` s'appuient sur la maximisation du critère de la modularité de graphe appliqué au cas bipartite. `CoclustSpecMod` utilise une approche spectrale (Labiod et Nadif, 2011) tandis que `CoclustMod` procède à une maximisation directe de la modularité (Ailem et al., 2016). Ces deux algorithmes conduisent à une structure de *co-clustering* diagonale. Le troisième algorithme, `CoclustInfo`, repose quant à lui sur la maximisation de l'information mutuelle (voir Govaert et Nadif (2013)). `CoclustInfo` est un exemple d'algorithme de *co-clustering* non diagonal.

Ces algorithmes sont très adaptés au traitement de matrices document-terme comme l'ont montré les études présentées dans les articles précédemment cités. A titre d'exemple la figure 2 montre les performances en termes de NMI (*Normalized Mutual Information*) obtenues par `CoclustMod` en comparaison avec plusieurs algorithmes connus et décrits dans (Ailem et al., 2016) : *FNMTF*, χ-*Sim*, *Block*, *Spec* et *SpecCo*. Les résultats ont été obtenus en lançant chaque algorithme sur différentes versions [1] de 10 corpus textuels de référence et en retenant pour chaque algorithme la solution optimisant son critère. On peut alors compter combien de fois chaque algorithme a permis d'obtenir le meilleur score NMI dans chaque situation (matrice originale, binarisée ou pondérée).

1. Matrices originales, binarisées ou pondérées avec TF-IDF.

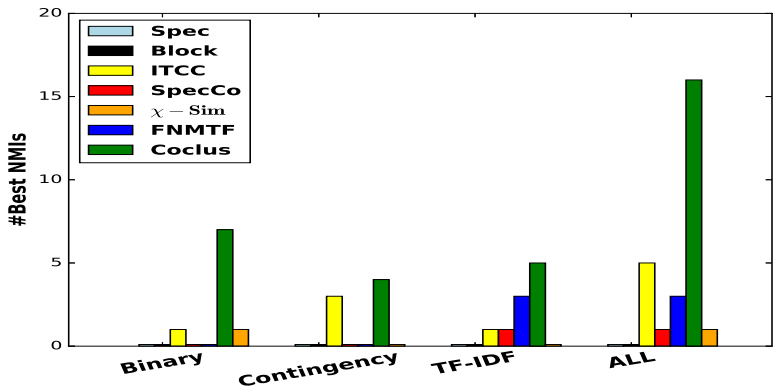


FIG. 2 – *Nombre de fois où chaque algorithme a permis d'obtenir le meilleur score NMI dans chaque situation (matrice originale, binarisée ou pondérée).*

3 Compatibilité avec la bibliotèque scikit-learn

CoClust utilise les conventions de **scikit-learn** (Pedregosa et al., 2011) : pour utiliser un algorithme, il suffit d'instancier un objet correspondant à cet algorithme puis d'appeler la méthode $fit(X)$ où X est la matrice à traiter. Dans l'exemple ci-dessous, on applique trois algorithmes différents à la matrice X :

```
c_1 = CoclustMod(n_clusters=4) ; c_1.fit(X)
c_2 = CoclustSpecMod(n_clusters=4) ; c_2.fit(X)
c_3 = CoclustInfo(n_row_clusters=3, n_col_clusters=4); c_3.fit(X)
```

Ceci permet de tester très facilement différents algorithmes de *co-clustering* sur un même jeu de données. **CoClust** permet aussi d'obtenir des informations sur la taille des *clusters* obtenus ou le contenu des *clusters* de termes (figure 3). Des représentations sous forme de graphe

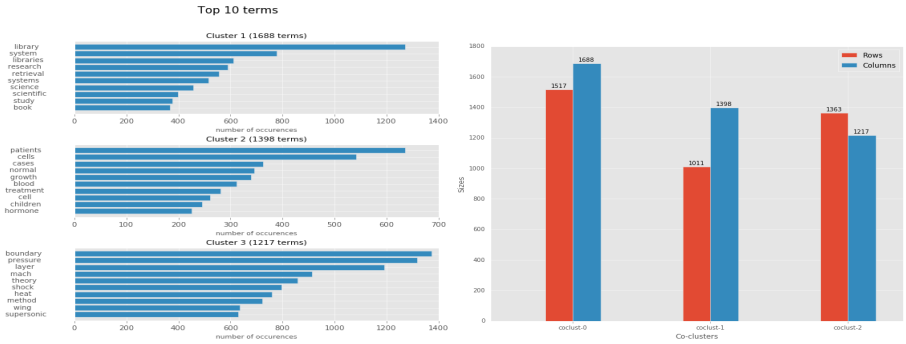


FIG. 3 – *(a) Termes représentatifs des 3* clusters *de termes de CLASSIC3 - (b) Tailles comparées des trois* co-clusters *obtenus.*

pour chaque *cluster* de termes, inspirées de Role et Nadif (2014), sont également disponibles (figure 4).

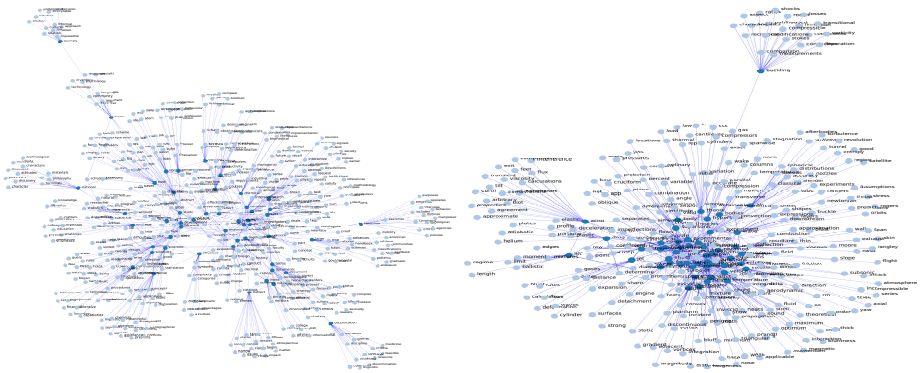


FIG. 4 – *CoclustMod : représentation de deux* clusters *de termes sous forme de graphes en utilisant la fonction get_term_graph du package. On voit que le* cluster *de droite est plus centré thématiquement que le* cluster *de gauche plus général.*

4 Conclusion et perspectives

Le package **Coclust** dispose déjà de trois algorithmes de *co-clustering* bien adaptés au matrices document-terme. D'autres algorithmes s'appuyant sur les modèles des blocs latents sont en cours de développement. Ce package offrira de cette façon un large éventail d'algorithmes efficaces, accompagnés de visualisation et faciles à utiliser.

Références

Ailem, M., F. Role, et M. Nadif (2016). Graph modularity maximization as an effective method for co-clustering text data. *Knowledge-Based Systems 109*(C), 160–173.

Dhillon, I. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. KDD'01, pp. 269–274.

Govaert, G. et M. Nadif (2013). *Co-Clustering*. John Wiley & Sons.

Kluger, Y., R. Basri, J. T. Chang, et M. Gerstein (2003). Spectral biclustering of microarray cancer data : Co-clustering genes and conditions. *Genome Research 13*, 703–716.

Labiod, L. et M. Nadif (2011). Co-clustering for binary and categorical data with maximum modularity. ICDM'11, pp. 1140–1145.

Pedregosa, F., G. Varoquaux, A. Gramfort, et ... (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Role, F. et M. Nadif (2014). Beyond cluster labeling : Semantic interpretation of clusters' contents using a graph representation. *Knowledge-Based Systems 56*, 141–155.

Summary

The Python **Coclust** package provides implementations of various co-clustering algorithms. The developed code has a consistent interface so as to facilitate the compared use of different co-clustering algorithms. **CoClust** also offers utilities for the exploratory analysis and interpretation of the obtained co-clusters.

Aide à l'automatisation de conception de systèmes de dialogue

Jean-Leon Bouraoui*, Vincent Lemaire*

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

Résumé. L'article décrit un processus industriel en cours de recherche / développement. L'objectif est d'obtenir une modélisation non supervisée de la structure de dialogues finalisés appartenant à un domaine donné (par exemple réservation de trains). Ce type de modélisation présente de nombreux intérêts, le principal étant de servir de base à la conception de l'architecture d'un agent dialoguant. La modélisation obtenue est représentée par un graphe présentant 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, pour obtenir le graphe décrivant les principaux thèmes du corpus et leur séquentialité. Nous décrivons en détail le processus mis en place, et situons cette approche par rapport aux travaux connexes. Enfin, nous présentons les verrous scientifiques restants.

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) . Pour quantifier cet intérêt grandissant pour la technologie des interfaces dialoguantes et des agents dialoguants en particulier, nous pouvons citer la récente étude du cabinet d'analyse Gartner, qui place les systèmes conversationnels parmi les 10 technologies stratégiques pour 2017 [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, comme base à discussion au cours de l'atelier TextMine d'EGC 2017, une méthodologie ayant pour but de mettre en place une solution d'assistance

1. http://www.gartner.com/newsroom/id/3482617

semi-automatique à la création ou l'adaptation d'un agent dialoguant pour un domaine applicatif donné.

2 Description de la problématique

Dans cet article, 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).

Dans un premier temps, on cherche à associer chaque tour de parole à une "classe" donnée notée L_c. Une classe correspond à l'intention que l'interlocuteur exprime dans son tour de parole ; prenons comme exemple un dialogue entre un client et le service client d'un opérateur téléphonique : le tour de parole où l'agent demande au client de s'identifier appartient à une classe spécifique (nommée par exemple *DmdeIdent*) ; celui où le client procède à son identification est relatif à une autre classe (nommée par exemple *RepIdentClient*).

Ensuite, on détermine les thèmes T_t, qui regroupent un ensemble de classes relatives à un sujet commun. Reprenons notre exemple du service client : les thèmes pourront être l'identification du client (*IdentClient* ; classes correspondantes : la demande par le téléconseiller, et la réponse du client), la discussion du problème (*ExpProb* ; classes correspondantes : la présentation par le client, et la demande de précision par le téléconseiller), etc.

L'association des classes aux tours de parole des différents dialogues du corpus pourra donc être représentée comme indiqué dans la figure 1, et le regroupement des classes en thèmes comme représenté dans le tableau 1.

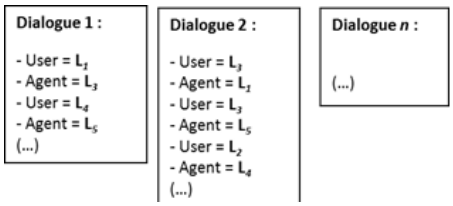


FIG. 1 – *Association de classes aux tours de paroles*

Notre but est de déterminer automatiquement : (i) les classes ; (ii) les thèmes ; (iii) les transitions entre thèmes et ceci au sein de chaque dialogue, de manière à 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 2. Notre postulat est que, selon la position dans le dialogue, un tour

Thèmes	T_1	T_2
Classes	L_1, L_4, L_5	L_3, L_6, L_7

TAB. 1 – *Regroupement des classes en fonction des thèmes auxquelles elles correspondent*

de parole donné présente plus de probabilité d'appartenir à une classe donnée (i.e. un cluster), qu'un autre ; cette information est donc prise en compte lors du processus, et l'un des objectifs de notre travail est de vérifier la validité de ce postulat.

Le graphe ainsi obtenu présente de multiples intérêts. Le principal est l'initialisation de la conception du système de dialogue : 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.

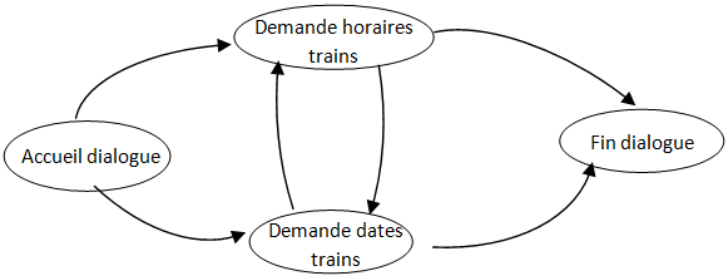


FIG. 2 – *Représentation graphique des principales transitions entre thèmes*

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 Travaux connexes

Dans la littérature, on peut regrouper sous trois catégories les différentes approches employées pour répondre à notre problématique, selon qu'elles procèdent par l'identification des thèmes et de leur séquentialité, qu'elles utilisent le Deep Learning, ou qu'elles adoptent une approche *ad hoc*.

— Identification des thèmes puis des séquences de thèmes : les travaux appartenant à cette catégorie se différencient principalement selon la méthode utilisée pour identifier les

thèmes des dialogue. Ainsi, Bangalore et al. (2008) et Chotimongkol (2008) utilisent des clusters à cette fin, tandis que Paul (2012) et Zhai et Williams (2015) utilisent le modèle de la Latent Dirichlet Allocation (LDA), qui permet d'identifier les "thèmes" dans un document ou ensemble de documents textuels. Dans tous les cas de figure, ces auteurs utilisent les Hidden Markov Models (HMM) pour modéliser les transitions entre thèmes.

— Deep Learning : A notre connaissance, le premier article ayant proposé l'utilisation, pour les agents dialoguants, de techniques appartenant au Deep Learning, est Vinyals et Quoc (2015). Les auteurs utilisent le modèle *seq2seq* pour modéliser l'enchaînement des tours de parole entre interlocuteurs, dans un réseau neuronal récurrent. Deux domaines d'application sont décrits, dont l'un relève du dialogue finalisé (tchats de dépannage informatique).

— Autres approches : Nous regroupons dans cette catégorie les travaux qui utilisent des méthodes *ad hoc* à la place ou en complément d'une modélisation des données. Ainsi, D'Haro et al. (2009) appliquent des heuristiques logicielles pour constituer un modèle de dialogue à partir d'une base de données applicatives. Des heuristiques sont également utilisées dans Chalamalla et al. (2008) et Negi et al. (2009) pour passer d'une représentation en clusters à une modélisation du dialogue. Nous citons aussi Laroche (2015) qui intègre des techniques de Recherche d'Information dans l'architecture d'un agent dialoguant.

4 Description de notre approche

4.1 Détermination des classes et des thèmes

On utilise ici 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. Le produit des partitions uni-variées forme une partition multivariée de l'espace de représentation, i.e., une grille ou matrice de cellules et il représente aussi un estimateur de densité jointe des variables. Afin de choisir la "meilleure" grille (connaissant les données) de l'espace de modèles, une approche Bayésienne dite Maximum A Posteriori (MAP) est utilisée. La méthode employée est basée sur l'approche MODL décrite dans (Boullé et al., 2014). Comme pour toute méthode d'apprentissage automatique, notre approche nécessite une quantité de données : un volume minimum est requis pour que le clustering soit pertinent.

Un outil de Visualisation des CoClusters trouvés est ensuite utilisé (Guerraz et al., 2015), ce qui permet une analyse fine et un "profilage" des clusters obtenus. Nous ne décrivons pas ici tous les détails mais l'utilisateur peut choisir la granularité de la grille nécessaire à son analyse tout en contrôlant soit le nombre de parties soit le taux d'information mais aussi réaliser un taggage (par exemple à l'aide de mots clefs) des clusters obtenus.

4.2 Détermination des transitions entre thèmes

Dans notre approche de la problématique, un thème correspond à un cluster de tours de parole. Un thème défini ainsi peut être relié à un ou plusieurs autres thèmes, en fonction de la fréquence observée de leurs successions dans les dialogues du corpus.

La représentation obtenue est un graphe orienté, dont les nœuds sont les clusters, et les arcs sont les successions entre clusters. La génération du graphe se fait en plusieurs phases successives, schématisées dans la figure 3.

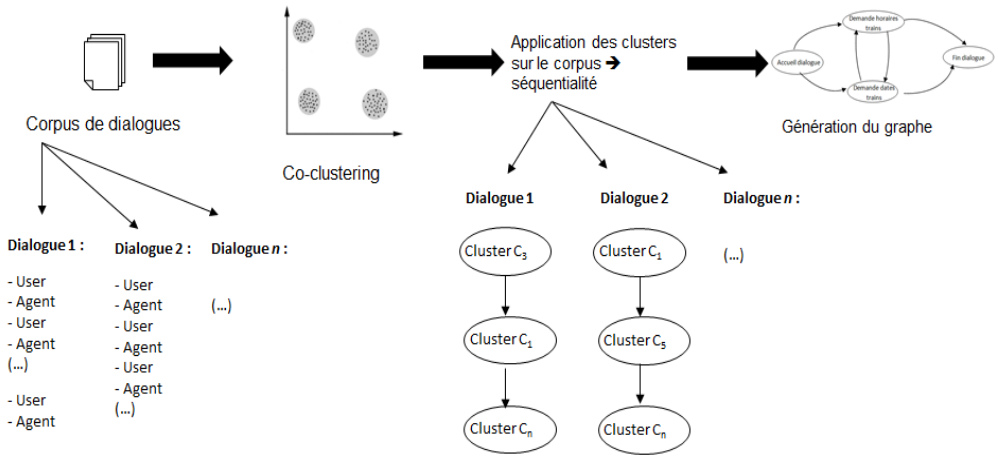


FIG. 3 – *Chaîne de traitement*

La première consiste à obtenir un corpus standardisé, à l'aide d'un outil de prétraitement de texte (interne à Orange Labs et non décrit ici) paramétré avec la liste de stopwords utilisée pour le stemming français par la librairie NLTK. Cette liste de stopwords permet de supprimer des mots a priori inutiles pour l'apprentissage (par exemple articles, prépositions, etc.).

La représentation ainsi obtenue est ensuite utilisée par l'outil de CoClustering mentionné ci-dessus, qui génère les clusters auxquels appartient chaque tour de parole du corpus. Les énoncés sont associés à des classes (ici : des clusters de tours de paroles) en fonction de la maximisation du contraste entre la distribution des co-parties (ici : des CoClusters de mots) et la distribution espérée sous l'hypothèse d'indépendance (connaissant les marginales). Les partitions induites par un CoClustering sur les deux entités sont des clusterings. La notion de similarité associée est liée à la façon dont les individus d'un cluster d'une entité se distribuent sur les clusters de l'autre entité.

On obtient ainsi un ensemble de clusters de tours de paroles, chaque cluster correspondant à un thème donné, et dénommé par un identifiant unique.

Les identifiants de cluster sont ensuite projetés sur le corpus initial, afin de retrouver la séquentialité des clusters. Autrement dit, à chaque tour de parole du corpus initial est désormais associé l'identifiant du cluster auquel le tour de parole appartient.

A partir de là, chaque dialogue peut ainsi être parcouru comme une séquence de cluster/thèmes. On peut ainsi calculer, sur l'ensemble des dialogues, les transitions entre chaque

cluster ; le début et la fin de chaque dialogue sont pris en compte de manière à éviter des transitions erronées.

Les fréquences de transition entre clusters ainsi calculées sur l'ensemble des dialogues sont stockées dans une matrice. Celle-ci est utilisée pour générer le graphe correspondant. Il est possible de choisir un seuil minimum de fréquence de transitions, à partir duquel les transitions sont affichées : cela permet de générer des graphes plus ou moins complexes à lire. D'autre part, la possibilité de n'afficher que des transitions supérieures à un seuil permet de limiter la prise en compte de transitions non pertinentes, provoquées par des erreurs dans le clustering.

4.3 Illustration sur un cas concret

Nous avons utilisé l'approche décrite dans la section précédente sur différents corpus de dialogues, appartenant à des domaines variés. A titre d'illustration, évoquons ici un corpus de 7 407 dialogues de "tchats" d'assistance en ligne de clients d'Orange avec des téléconseillers ; chaque dialogue comporte une quinzaine de tours de paroles en moyenne.

Nous avons appliqué à ce corpus le pré-traitement puis le CoClustering évoqués dans section précédente. On obtient 44 clusters différents, regroupant les tours de parole du corpus. Les identifiants des clusters sont ensuite associés, dans le corpus même, aux tours de paroles correspondants ; on récupère ainsi la séquentialité des clusters, qui va ensuite permettre la génération des graphes.

En appliquant le seuillage défini plus haut (avec différentes valeurs de seuil par graphe), on obtient des visualisations graphiques globalement pertinentes : l'analyse manuelle du contenu des clusters obtenus à partir d'un seuil élevé montre que la plupart des clusters sélectionnés correspondent bien à des phases bien spécifiques de ce type de dialogue : phases de salutations, d'identification du client, de description du problème rencontré, de l'identification de la solution à apporter, et phases de remerciement et de fin de dialogue.

L'ordonnancement de ces clusters/thèmes en termes d'architecture des dialogues montre également un déroulement séquentiel typique de ces différentes phases, nonobstant certaines erreurs de clustering. Celles-ci consistent en deux grandes catégories : soit l'hétérogénéité du cluster (un cluster correspond à plusieurs thèmes au lieu d'un seul), soit au contraire la redondance de clusters (plusieurs clusters différents correspondent au même thème).

A titre d'exemple, la figure 4 présente un graphe obtenu à partir de l'application de la chaîne de traitement sur le corpus mentionné ci-dessus. A fins de lisibilité, ce graphe n'affiche que 32 clusters : ceux dont le nombre de transitions est supérieur à 300 ; les nombres apparaissant à côté de chaque arc correspond aux fréquences de transition ; chaque cluster est désigné par l'identifiant attribué automatiquement par Khiops.

La figure 5 présente un autre graphe, obtenu à partir des mêmes données, cette fois après analyse du contenu des clusters du graphe de la figure 4. Pour le générer, nous avons utilisé un outil (interne à Orange Labs) de visualisation et de manipulation des données. A partir de l'analyse, cet outil nous a permis de renommer les clusters, fusionner certains d'eux, et en rendre la présentation encore plus lisible, comme pourrait le faire le concepteur d'un agent dialoguant pour obtenir une première architecture sans connaissance préalable du domaine.

Pour l'instant nous évaluons les résultats obtenus selon deux critères. Le premier est l'homogénéité des clusters obtenus. Le second est la qualité et la régularité des transitions entre

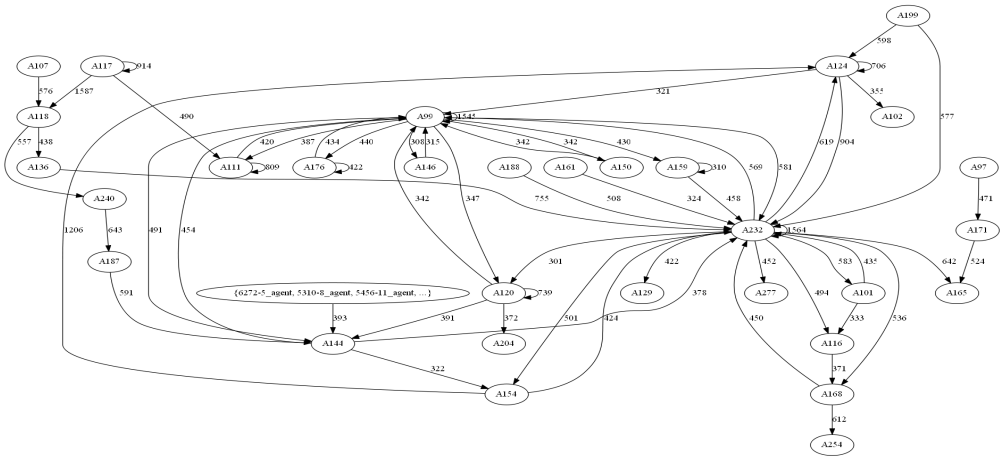


FIG. 4 – *Graphe initial (seuil de fréquence minimum : 300)*

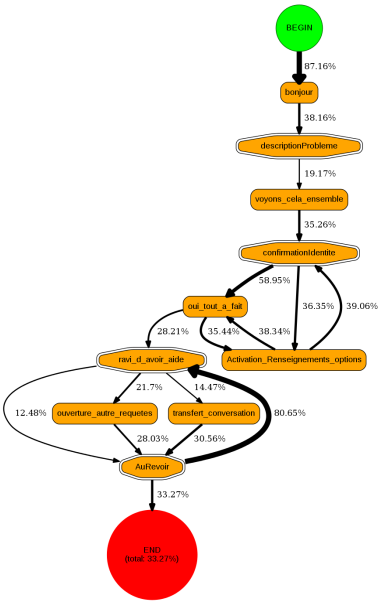


FIG. 5 – *Graphe pour conception d'architecture*

clusters ; ce second critère dépend fortement du premier. L'évaluation est en cours par les auteurs ainsi que par une ergonome.

5 Illustration lors de l'atelier

Lors de l'atelier une présentation plus détaillée des résultats obtenus sera effectuée pour illustrer concrètement notre approche. Nous présenterons les différents corpus utilisés jusqu'ici, les clusters obtenus, et la manière d'en étudier le contenu en regard des corpus initiaux. Nous présenterons aussi les graphes obtenus en fonction des fréquences de succession, et l'interface que nous utilisons pour les manipuler afin d'en déduire une architecture de systèmes de dialogues.

6 Conclusion

Dans cet article, nous avons présenté une méthode non supervisée pour obtenir une première version d'une architecture d'un système de dialogue, indépendamment du domaine couvert par celui-ci. La méthode comporte deux grandes étapes : d'une part l'application d'un algorithme de CoClustering sur un corpus de dialogues appartenant au domaine ciblé ; d'autre part, la prise en compte de la séquentialité des clusters obtenus pour représenter le déroulement prototypique d'un dialogue.

Il reste de nombreux verrous scientifiques autour de cette approche. Plusieurs sont relatifs à la technique même de clustering. Nous pensons notamment à la problématique de la thématisation des cluster : comment déterminer les mots ou tours de paroles les plus représentatifs du cluster ? La question de la sélection des clusters les plus pertinents, c'est à dire les plus homogènes au regard d'une thématique donnée, se pose aussi, en corrélation avec la granularité/taille optimale des clusters. Il serait également intéressant d'étudier l'efficacité d'optimisations linguistiques du corpus pour diminuer le nombre de paramètres utilisés pour le clustering : par exemple lemmatisation des mots, neutralisation des Entités Nommées.

En ce qui concerne la modélisation de la succession des clusters, de nombreuses problématiques se posent aussi. Nous avons vu que dans la littérature, c'est souvent un HMM qui est utilisé à cette fin ; mais nous pensons qu'il serait intéressant d'étudier d'autres approches telles que celles des Conditional Random Fields (CRF).

Enfin, la question se pose de la capacité de réutilisabilité des informations obtenues (notamment les classes et thèmes) d'un domaine à l'autre. Un déploiement de ces informations vers des domaines proches du domaine initial est faisable et prévu dans la suite Khiops. Pour des domaines plus éloignés, une approche relevant de l'apprentissage par transfert (*transfer learning*) est envisageable.

Références

Bangalore, S., G. DiFabbrizio, et A. Stent (2008). Learning the structure of task-driven human-human dialogs. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, 201–208.

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.

Chalamalla, A., S. Negi, S. Joshi, et L. V. Subramaniam (2008). Identification of class specific discourse patterns. *CIKM '08*.

Chotimongkol, A. (2008). *Learning the Structure of Task-Oriented Conversations from the Corpus of In-Domain Dialogs*. Phd thesis, Carnegie Melon University.

D'Haro, L. F., R. Cordoba, J. M. Lucas, R. Barra-Chicote, et R. San-Segundo (2009). Speeding up the design of dialogue applications by using database contents and structure information. *SIGDIAL 2009*, 160–169.

Guerraz, B., M. Boullé, D. Gay, V. Lemaire, et F. Clérot (2015). Analyse exploratoire par k-coclustering avec khiops coviz. In *Atelier CluCo, Extraction et Gestion des Connaissances (EGC)*.

Laroche, R. (2015). Speeding up the design of dialogue applications by using database contents and structure information. *18th International Conference on Intelligence in Next Generation Networks*, 231–238.

Negi, S., S. Joshi, A. Chalamallay, et L. V. Subramaniam (2009). Automatically extracting dialog models from conversation transcripts. *2009 Ninth IEEE International Conference on Data Mining*.

Paul, M. (2012). Mixed membership markov models for unsupervised conversation modeling. *EMNLP-CoNLL '12*, 231–238.

Vinyals, O. et V. L. Quoc (2015). A neural conversational model. *International Conference on Machine Learning*, 231–238.

Zhai, K. et J. Williams (2015). Discovering latent structure in task-oriented dialogues. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 36–46.

Annexe

Nous remercions Nicolas Voisine, qui nous a fourni Patatext, l'outil de prétraitement de texte pour Khiops, ainsi que Fabien Dupont et Laurent Roussarie pour leur participation à la chaîne de clustering et de visualisation des graphes, et pour leur aide à l'utilisation.

Nous remercions également Nathalie Legay pour sa participation à l'étude des graphes obtenus.

Nous remercions enfin Romain Laroche, qui a contribué à la rédaction de l'introduction de cet article.

Summary

The paper describes an on-going industrial process of research / development. Our goal is to obtain an unsupervised modeling of the structure of task-oriented dialogues. The dialogues are related to a given domain (for instance, train booking). This modeling 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. The resulting graph describes the main themes in the corpus, and their overall sequential organization. We describe the details of the process that we set up, and our position in regard to the related works. We eventually present the remaining scientific problems to address.

Exploratory Text Segmentation through Joint Distribution Estimation

Dominique Gay*, Romain Guigourès**
Marc Boullé*,*** Fabrice Clérot***

*Université de La Réunion
**Zalando
***Orange Labs

Résumé. We suggest a novel way for exploratory topic segmentation based on data grid models. In this context, a text can be represented as a data set of two-dimensional points; each point is defined by two variables: a word (categorical value) and the placement of the word in the text (numerical value). Instantiating data grid models to the 2D-points turns the problem into coclustering. Simultaneously, the words are partitioned into clusters and the placement (or time) variable is discretized into intervals/segments, following a parameter-free Bayesian model selection approach. We also suggest several criteria for exploiting the resulting grid through agglomerative hierarchies, for interpreting the clusters of words and characterizing their components through insightful visualizations. Experiments on the Bible show the relevance of our approach.

1 Exploratory topic segmentation

Text Segmentation has been extensively studied over the past years since it is a prequel to further text analytics like e.g., text summarization or information retrieval. Since pioneering work Hearst (1997) based on lexical cohesion, many text segmentation techniques have been suggested in the literature (see Purver (2011) for a well-structured survey) : e.g, among others, Utiyama et Isahara (2001), MCSeg Malioutov et Barzilay (2006), BayesSeg Eisenstein et Barzilay (2008), HierBayes Eisenstein (2009), APS Kazantseva et Szpakowicz (2011), TSM Du et al. (2013), etc.
In this paper, we focus on the long text segmentation problem and we suggest a new approach for *exploratory topic segmentation*. Pragmatically, a resulting text segmentation should hold the following features :

— *(i)* the segmentation technique should give a global picture of the underlying structure of the text and show the evolution of the detected topics along the running text
— *(ii)* the segmentation technique should highlight groups of words that are characteristic of segments ;
— *(iii)* for the sake of ergonomy, computing the segmentation should not involve parameter tuning ;
— *(iv)* the whole methodology should allow to explore the resulting segmentation at various granularities.

To the best of our knowledge, there is no long text segmentation technique having all these features. The methodology we suggest fulfills all the previous requirements and uses recent progress in joint distribution estimation based on data grid models Guigourès et al. (2015); Gay et al. (2015).

2 Data Grid Models in a Nutshell

Data grid models Boullé (2011) aim at estimating the joint distribution between K variables of mixed-types (categorical as well as numerical). The main principle is to simultaneously partition the values taken by the variables into groups/clusters of categories for categorical variables and into intervals for numerical variables. In this context, a text is represented by two variables : W for the words and T for the placement (or time) of the word in the text. Instantiating data grid models for text segmentation, the result is a two-dimensional data grid whose cells (or co-clusters) are defined by a part of each partitioned variable value set, i.e., a cluster of words and a time/placement interval.

In order to choose the "best" data grid model M^* (given the data) from the model space $\mathcal{M}$, we use a Bayesian Maximum A Posteriori (MAP) approach. We explore the model space while minimizing a Bayesian criterion, called cost. The cost criterion implements a trade-off between the accuracy and the robustness of the model and is defined as follows :

$$cost(M) = -\log(\underbrace{p(M \mid D)}_{\text{posterior}}) \propto -\log(\underbrace{p(M)}_{\text{prior}} \times \underbrace{p(D \mid M)}_{\text{likelihood}}) \tag{1}$$

where D is the underlying data. Thus, the optimal grid M^* is the most probable one (maximum a posteriori) given the data. Considering a data-dependent hierarchical prior (on the parameters of the grid model) that is uniform at each stage of the hierarchy, Boullé (2011) has shown that we can obtain an exact analytical expression of the cost criterion. The full details about the $cost$ criterion and the optimization algorithm are available in Boullé (2011). The key features to keep in mind are : *(i)* the algorithm is parameter-free, i.e., there is no need for setting the number of clusters/intervals per dimension ; *(ii)* using a greedy bottom-up strategy coupled with pre and post-optimization heuristics and Variable Neighbourhood Search meta-heuristic, it provides an effective locally-optimal solution to the data grid model construction efficiently, in sub-quadratic time complexity ($O(N\sqrt{N}\log N)$ where N is the number of data points). Notice that in the case of two categorical variables (e.g., texts $\times$ words for text categorization), the criterion is an exact density estimation estimator, that asymptotically converges to the mutual information between both partitions. In other words, it could be compared as a regularized version of the Information Theoretic Coclustering Dhillon et al. (2003) – while, in addition it can deal with mixed-typed variables.

2.1 Data grid exploitation and visualization

When facing long texts, the optimal grid M^* can be made of hundreds of parts per dimension, i.e., many thousands of cells, which is difficult to exploit and interpret. To alleviate this issue, we suggest a grid simplification method together with several criteria that allow us to choose the granularity of the grid for further analysis, to rank words in clusters and to gain

insights in the underlying text through meaningful visualizations.

Dissimilarity index and grid structure simplification. In order to simplify the structure, we propose to apply an agglomerative hierarchical clustering on top of the optimal model M^*. We derive a dissimilarity measure between clusters adjacent intervals from the exact criterion described above.

Definition 1 (Dissimilarity index) *Let $c_{.1}$ and $c_{.2}$ be two parts of a variable partition of a grid model M. Let $M_{c_{.1} \cup c_{.2}}$ be the grid after merging $c_{.1}$ and $c_{.2}$. The dissimilarity $\Delta(c_{.1}, c_{.2})$ between the two parts $c_{.1}$ and $c_{.2}$ is defined as the difference of cost before and after the merge :*

$$\Delta(c_{.1}, c_{.2}) = cost(M_{c_{.1} \cup c_{.2}}) - cost(M) \tag{2}$$

Building such a hierarchy on top of a grid obtained using a regularized approach has the advantage to provide an interesting exploratory analysis tool for exploring the results at different level of granularity, while ensuring reliable results. It has been shown, in Guigourès et al. (2015), that asymptotically, Δ converges to the Jensen-Shannon divergence between the distributions of two clusters over the time intervals (resp. two intervals over the clusters). This approach could then be regarded as an agglomerative information bottleneck Slonim et Tishby (2000) approach starting from an optimal level and preventing the probability estimation errors occurring at the first merges of agglomerative approaches based on divergence. In order to control the degradation of the quality of our model, we introduce the information ratio of the grid M', defined as follows :

$$IR(M') = \frac{cost(M') - cost(M_\emptyset)}{cost(M^*) - cost(M_\emptyset)} \tag{3}$$

where $M_\emptyset$ is the null model (the grid with a single cell).

Typicality for ranking words in a cluster. When the grid is coarsen during the hierarchical agglomerative process, the number of clusters of words decreases and the number of words per cluster increases. It is useful to focus on the most representative words among thousands of words of a cluster. In order to rank words in a cluster, we define the typicality of a word as follows.

Intuitively, the typicality evaluates the average impact in terms of $cost$ on the grid model quality of removing a word from its cluster and reassigning it to another cluster. Thus, a word is representative (say typical) if it is "close" to the cluster it belongs to and "different in average" from other clusters. The typicality is actually very similar to the *Silhouettes* approach Rousseeuw (1987), an efficient technique for validation and interpretation of a clustering.

Insightful visualizations with Mutual Information. It is common to visualize 2D coclustering results using 2D frequency matrix or heat map. We also suggest an insightful measure for co-clusters to be visualized, namely, the Contribution to Mutual Information (CMI) – providing additional valuable visual information inaccessible with only frequency representation.

Definition 2 (Contribution to mutual information) *The mutual information between two partitioned variables W^M of size J_W and T^M of size J_T (from the partition M of W and T*

variables induced by the grid model M) is defined as :

$$MI(W^M;T^M) = \sum_{i_1=1}^{J_W}\sum_{i_2=1}^{J_T} MI_{i_1 i_2} \text{ where } MI_{i_1 i_2} = p(c_{i_1 i_2}) \log \frac{p(c_{i_1 i_2})}{p(c_{i_1 .})p(c_{.i_2})} \quad (4)$$

where $MI_{i_1 i_2}$ represent the contribution of cell $c_{i_1 i_2}$ to the mutual information, $p(c_{i_1 i_2})$ is the observed joint probability of points in cell $c_{i_1 i_2}$ and $p(c_{i_1 .})p(c_{.i_2})$ is the expected probability in case of independence, i.e., the product of marginal probabilities.

3 Application

Data. For our experiments, we use the text of The Bible which is among the best-selling books of all time – and as far as we know, surprisingly, it has never been automatically segmented. It is also one of the most studied book. Thus, the following findings through exploratory text segmentation can easily be asserted by common knowledge on the book – which is taken as ground truth. The wide-spread King James version of The Bible (without apocrypha) is composed of 66 books [1] ramified in *(i)* the *Old Testament*, subdivided into the Pentateuch (5 books), the Historical Books (12), the Poetical Books (5), the Prophets (17) and *(ii)* the *New Testament*, subdivided into the Gospels (4), the Acts of the Apostles (1), the Pauline and other Epistles (21) and the Revelation (1). Each book is divided into chapters, then each chapter into verses. The Bible originally contains $|W| = 12918$ unique words ($N = 789628$ total) and $|T| = 31102$ verses. For this experiment, we work at the verse level (i.e., cut points are allowed only between verses) and pre-process the text by removing stop words and grouping lexical items by stemming Porter (1980). Thus, the 2D input for data grid models is like :
$(1, begin), (1, god), (1, creat), (1, heaven), \ldots,$
$(|T|, lord), (|T|, jesus), (|T|, christ), (|T|, amen).$

The big picture. Thanks to Khiops Coclustering [2], an effective locally-optimal grid is obtained in 67 minutes and is made of 329 segments and 252 clusters of words. At this scale (see Figure 1), the analysis of the summary provided by the 2D-segmentation is not an easy task for a non-expert. However, we can highlight two clusters of words : In green (rectangular), the cluster of words whose most typical word is "*begat*", which relates to the genealogy of some characters and is a recurrent topic in the Bible (see circles) : e.g., Adam and Noah in the Genesis book, Saul in the Chronicles and Jesus in the Gospels. In pink, the clusters of words whose most typical words are "*angel, repent, satan and throne*" which relates to the apocalypse described in the Revelation book at the end of the Bible.

Zoom into the Gospels. Now, we zoom into the Gospels part (see Figure 2). The storyline of Gospels is Jesus' life. Whereas the gospels are perfectly separated from other books, segmentations *between* Matthew, Mark, Luke and John show a mismatch of about 1-3 verses. However, the recurrent topics of the various Gospels are well-identified. In blue (rectangular), cluster of words $C1$ relates to typical characters of the Gospels : "Jesus, discipl, Peter, John, Simon" who are recurrent along the text ; while $C2$ relates to the various acts and encounters

1. https://en.wikipedia.org/wiki/List_of_books_of_the_King_James_Version
2. Khiops CoClustering is available as a shareware for research purpose at http://www.khiops.com

FIG. 1: Visualisation of CMI for the resulting (x-axis = 329 segments $\times$ y-axis = 252 clusters of words)-grid obtained on the whole Bible. Red cells indicate positive CMI, i.e., excess of interactions between T and W in the cell. Meta-segmentations in black lines are manually added and annotated.

of Jesus. Typical words of $C2$ are "*ask, temple, sit, pharise, whosoever, ship, heal*". We also observe a strong similitude between the so-called *synoptic* Gospels (Matthew, Mark and Luke) contrasting with the Gospel of John : indeed, considering cells in $C2$, the cumulative CMI in each synoptic Gospel is above 25.10^{-4} while only 5.10^{-4} for the fourth Gospel. Notice also that the genealogy of Jesus is reported only by Matthew and Luke (see green ellipses). Despite the variations between Gospels, they all have in common the passion and resurrection (with typical words : "*pilat, mari, crucifi, sepulchr, betray, ...*", placed at the end of each gospel (see blue ellipses).

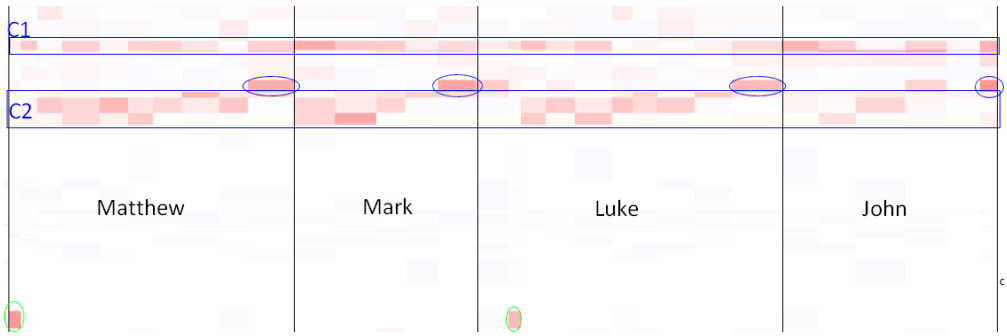


FIG. 2: Zoom into the Gospels.

Agglomerative hierarchy. Since the Bible is made of 66 books, we build a agglomerative hierarchy on top of the computed grid, as defined in previous section, in order to have 66 segments. The resulting segmentation matches perfectly with 13 books boundaries (i.e., matching the end of the books of Ruth, Kings2, Nehemiah, Esther, Ecclesiastes, Song of Solomon, Jeremiah, Ezekiel, Daniel, Zephaniah, Malachi, Gospel of John and Philemon). With a tolerance of ± 5 verses, the end of the books of Genesis, Exodus, Leviticus, Job, Psalms, Gospel of Luke,

Acts of the Apostles and the third Epistle of John are also detected. At this granularity, the book of Genesis is still *over-segmented*, since they relate several different stories from the origins of the earth and human kind to the life of characters such as Abraham, Jacob and Joseph – involving a specific vocabulary in each story which is quite rare in the rest of the Bible. Similar observations stand for the book of Exodus – explaining the mismatches.
Continuing towards the null model (see figure 3), the two last steps of segments agglomeration highlight perfect cuts (at the top of the hierarchy) between the Old and the New testament and inside the Old Testament between {Pentateuch, Historical Books} and {Poetic Books, Prophets}.

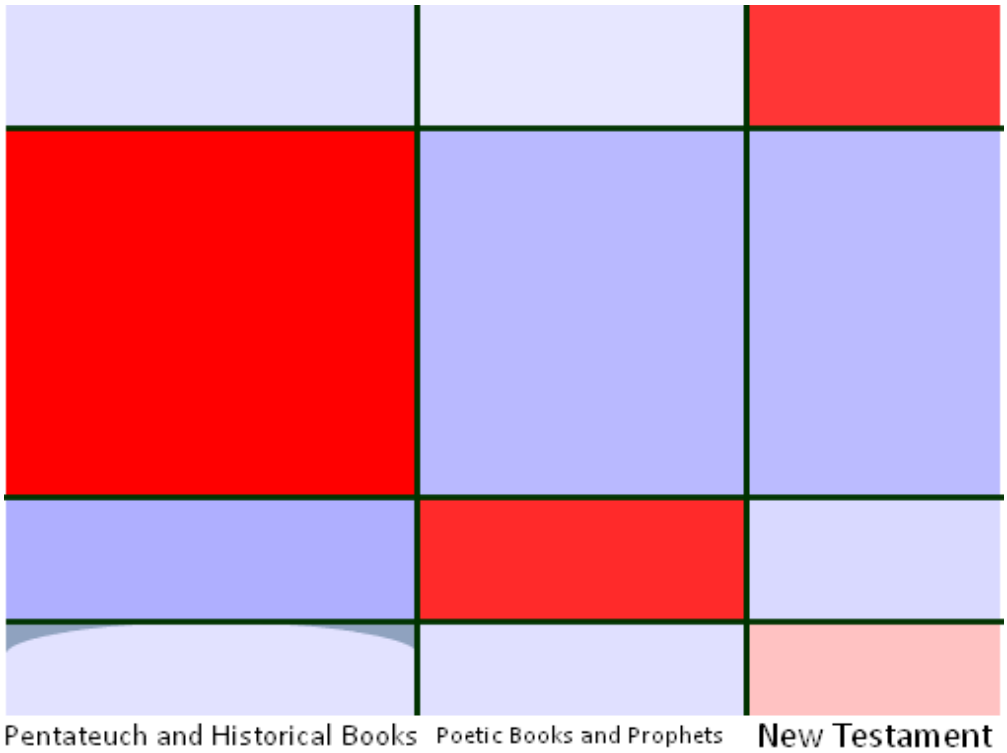


FIG. 3: (3×4)-grid, where the New and Old Testament are perfectly separated and where the Old Testament is perfectly divided into {Pentateuch, Historical Books} and {Poetic Books, Prophets}.

4 Discussion

We have suggested a relevant application nugget of data grid models for exploratory topic segmentation. Data grid models provide an effective 2D segmentation of a given long text. The method allows to efficiently get the big picture of the underlying text and to explore the segmentation at multiple levels of granularity while highlighting significant topics of the text.

While applying the method on the Bible has been successful, we plan to extend the experiments to multiple comparisons with state-of-the-art text segmentation techniques on artificially generated data and on other long texts.

Références

Boullé, M. (2011). Data grid models for preparation and modeling in supervised learning. In I. Guyon, G. Cawley, G. Dror, et A. Saffari (Eds.), *Hands-On Pattern Recognition : Challenges in Machine Learning, vol. 1*, pp. 99–130. Microtome.

Dhillon, I. S., S. Mallela, et D. S. Modha (2003). Information-theoretic co-clustering. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 24 - 27, 2003*, pp. 89–98.

Du, L., W. L. Buntine, et M. Johnson (2013). Topic segmentation with a structured topic model. In *Human Language Technologies : Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pp. 190–200.

Eisenstein, J. (2009). Hierarchical text segmentation from multi-scale lexical cohesion. In *Human Language Technologies : Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, May 31 - June 5, 2009, Boulder, Colorado, USA*, pp. 353–361.

Eisenstein, J. et R. Barzilay (2008). Bayesian unsupervised topic segmentation. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 334–343.

Gay, D., R. Guigourès, M. Boullé, et F. Clérot (2015). TESS : temporal event sequence summarization. In *2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015, Campus des Cordeliers, Paris, France, October 19-21, 2015*, pp. 1–10.

Guigourès, R., M. Boullé, et F. Rossi (2015). Discovering patterns in time-varying graphs : a triclustering approach. *Advances in Data Analysis and Classification*, 1–28.

Guigourès, R., D. Gay, M. Boullé, F. Clérot, et F. Rossi (2015). Country-scale exploratory analysis of call detail records through the lens of data grid models. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part III*, pp. 37–52.

Hearst, M. A. (1997). TextTiling : segmenting text into multi-paragraph subtopic passages. *Computational Linguistics 23*(1).

Kazantseva, A. et S. Szpakowicz (2011). Linear text segmentation using affinity propagation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 284–293.

Malioutov, I. et R. Barzilay (2006). Minimum cut model for spoken lecture segmentation. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual*

Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program 14*(3).

Purver, M. (2011). Topic segmentation. In G. Tur et R. de Mori (Eds.), *Spoken Language Understanding : Systems for Extracting Semantic Information from Speech*, pp. 291–317. Wiley.

Rousseeuw, P. J. (1987). Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics 20*, 53–65.

Slonim, N. et N. Tishby (2000). Document clustering using word clusters via the information bottleneck method. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 208–215. ACM.

Utiyama, M. et H. Isahara (2001). A statistical model for domain-independent text segmentation. In *Association for Computational Linguistic, 39th Annual Meeting and 10th Conference of the European Chapter, Proceedings of the Conference, July 9-11, 2001, Toulouse, France.*, pp. 491–498.

Summary

We suggest a novel way for exploratory topic segmentation based on data grid models. In this context, a text can be represented as a data set of two-dimensional points; each point is defined by two variables: a word (categorical value) and the placement of the word in the text (numerical value). Instantiating data grid models to the 2D-points turns the problem into co-clustering. Simultaneously, the words are partitioned into clusters and the placement (or time) variable is discretized into intervals/segments, following a parameter-free Bayesian model selection approach. We also suggest several criteria for exploiting the resulting grid through agglomerative hierarchies, for interpreting the clusters of words and characterizing their components through insightful visualizations. Experiments on the Bible show the relevance of our approach.

Préliminaire à la construction d'un réseau de signalisation en biologie systémique

F. Landomiel*, A. Gupta**,***
D. Maurel**, A. Poupon*

*Équipe BIOS, INRA, UMR85, Unité Physiologie de la Reproduction et des Comportements, 37380 Nouzilly, France
flavie.landomiel@gmail.com; anne.poupon@tours.inra.fr
**Université François-Rabelais de Tours, Laboratoire d'informatique
anubhav.gupta@etu.univ-tours.fr; denis.maurel@univ-tours.fr
***DIST CNRS

Résumé. Dans le domaine scientifique, la littérature est un outil indispensable à la connaissance. Cependant, l'avancée des recherches et l'édition de documents scientifiques ne fait que progresser de manière exponentielle. En ce sens, il devient de plus en plus ardu pour un scientifique d'être à jour dans son domaine d'expertise. Afin de palier à cette difficulté, le projet Biosystémique a permis de développer une méthode dans le but d'extraire les résultats expérimentaux dans les publications scientifiques concernant la biologie systémique et, plus précisément, concernant les voies de signalisation des récepteurs couplés aux protéines G (RCPG). Par la suite, ces données seront utilisées dans un moteur d'inférence afin de créer un réseau de signalisation et ainsi mettre en relation des données qui semblaient indépendantes. Dans cet article, nous avons mis en évidence la possibilité d'extraire des phrases clés d'un article indépendamment de l'interprétation de l'auteur.

1 Introduction

1.1 Motivation

Le domaine biomédical est un domaine d'application intéressant de la fouille de texte par son attractivité et l'enrichissement constant des bases de données mises à la disposition du public. Dans l'optique du projet Biosystémique, notre axe de recherche s'est orienté sur les voies de signalisation en biologie cellulaire, plus particulièrement celles impliquant les RCPG.

Les RCPG sont des protéines situées dans la membrane des cellules. Ces protéines sont des récepteurs, c'est-à-dire qu'elles se lient de manière spécifique à un ligand tel que des hormones, des ions, des lipides, etc. Cette liaison du ligand au récepteur déclenche un grand nombre de réactions à l'intérieur de la cellule qui correspond ainsi au réseau de signalisation.

Les RCPG étant des protéines spécifiques et membranaires, ils représentent des cibles idéales pour les médicaments. En effet, presque la moitié des médicaments sur le marché

ciblent ces récepteurs. Cependant, la connaissance concernant ces RCPG reste limitée car seulement 15 % des récepteurs sont ciblés par la pharmacopée. De plus, des effets secondaires peuvent se déclarer suite à la mauvaise connaissance du réseau de signalisation spécifique du RCPG.

La littérature scientifique est répertoriée dans de nombreuses bases de données. Les voies biologiques sont considérées comme très critiques dans notre compréhension du mécanisme des fonctions biologiques. Pour collecter des données sur les voies, pendant longtemps, la curation manuelle a été la méthode la plus populaire pour recueillir des informations nécessaires à la construction de ces voies de signalisation. L'inconvénient de cette méthode concerne le temps passé sur l'analyse : elle nécessite l'intervention d'experts pour identifier, collecter des informations et appliquer leurs connaissances biologiques pour organiser les interactions acquises de telle sorte qu'elles réalisent ensemble une fonction biologique commune comme une voie de signalisation. Le travail de Tari et al. (2010) se rapproche sensiblement de notre thématique de recherche dans le sens où ils ont cherché et réussit à automatiser des voies pharmacocinétiques (Tari et al., 2010). Ils ont même pu intégrer des faits qui n'étaient pas annotés manuellement sur PharmKGB (voir ci-après). Contrairement à nous, ils ne se sont basés que sur les résumés des textes et non sur leur contenu intégral pour la reconstruction des réseaux. Dès lors, il est possible que des éléments puissent manquer car le résumé ne décrit que ce que l'auteur a décidé.

Face à cette observation, la construction du réseau qui nous intéresse va rassembler tous les résultats expérimentaux décrits dans la littérature, relatifs à l'activation d'un RCPG et puis de pouvoir assembler le réseau. En 2011, Gloaguen et al. ont d'ailleurs commencé ce travail mais de manière semi-automatique sur un corpus à taille humaine de 300 textes (Gloaguen et al., 2011). Pour parvenir à ces réseaux de signalisation, de nombreuses bases de données ont été créées. Nous retrouvons notamment Reactome (Joshi-Tope et al., 2005), KEGG (Kanehisa et al., 2004) et HumanCyc (Romero et al., 2005) qui regroupent les voies métaboliques tandis que Biocarta [1] et Panther (Mi et al., 2007) se concentrent sur les voies de signalisation. Enfin, PharmGKB (Klein et al., 2001) qui met en lien les effets des drogues sur les variations génomiques en incluant les voies pharmacocinétiques et pharmacodynamiques. De même, UniProt (Bateman et al., 2015) recense toutes les protéines existantes ainsi que leurs divers noms et leurs homologues. En revanche, l'expansion des publications scientifiques ne facilite pas ce travail. En effet, pour le seul mois de Septembre 2015, 1 993 articles ont été recensés concernant les voies de signalisation. De plus, dans ces publications, les auteurs interprètent leurs résultats expérimentaux et seule une petite partie des données nécessaires à la reconstruction du réseau est présentée (Heiner et al., 2004; Kell et Oliver, 2004; Ananiadou et al., 2006; Luciano et Stevens, 2007; Ye et Doak, 2011).

Dès lors, l'idée est venue de développer une méthode d'inférence automatique où seuls les résultats expérimentaux seraient pris en compte et non leur interprétation. De même, la taille des textes utiles à la reconstruction d'un réseau ne serait plus limitant dans cette thématique de recherche. Enfin, l'extraction des informations pourra se faire sans l'aide d'aucun lecteur. Suite à cela, le réseau nouvellement construit pourrait même apporter de nouvelles hypothèses de travail concernant un aspect de la voie à creuser (nouvelle relation entre deux protéines par exemple). Ce dernier aspect nécessitera donc une collaboration avec un biologiste fondamental.

1. www.biocarta.com

Ce travail s'inscrit dans la continuité de la plateforme MEDIE [2] qui permet de retrouver entre autre des relations protéines-protéines, protéine-maladie dans la littérature mais seulement dans les abstracts alors que nous nous focalisons sur l'article intégral, ou, plus précisément, sur la partie "Résultats" de ces articles.

La suite de cet article présente donc les résultats préliminaires obtenus concernant seulement une partie précise d'un réseau de signalisation, le réseau dépendant de ERK, une protéine importante pour la régulation cellulaire.

1.2 Expérimentation

La fouille de texte en biologie est un processus établi. Cependant, le processus d'automatisation nécessite d'avantage de développement. Comme cité précédemment, de nombreux articles sur la fouille de texte en biologie se basent sur les résumés fournis par MEDLINE (Tari et al., 2010; Kemper et al., 2010) alors que nous nous intéressons à la partie "Résultats" qui contient tous les résultats expérimentaux du texte. Dans notre cas, nous avons interrogé les deux bibliothèques numériques que sont ISTEX et PubMed/PMC (PubMed Central) sur le NCBI (National Center for Biotechnology Information) pour trois mots-clé précis, *ERK*, *arrestin* et *phosphorylation* (*ERK* et *arrestin* sont deux protéines et *phosphorylation* un état de la protéine).

Cependant, l'accès à ces documents n'existe pas sous tous les formats. En effet, dans certains cas, il n'est pas possible d'avoir accès au format balisé des textes (XML ou HTML pour PMC et TXT pour ISTEX), mais seulement au format non-balisé (PDF pour PubMed). Face à cette difficulté, nous nous sommes donc demandés si le format du texte avait un impact sur notre travail. C'est pourquoi pour la suite, nous avons comparé les résultats obtenus sur les fichiers ".txt" issus d'une conversion PDF et ceux obtenus sur les fichiers ".xml" ou ".html" (voir section 3.1).

Au final, nous avons donc choisi de ne télécharger que des articles existants sous un format non PDF, soit, avec ces mots-clés, 3 255 documents scientifiques. Dès lors, deux opérations successives ont été réalisées afin de trier les documents. En effet, dans un premier temps seule la partie d'intérêt de l'article (la partie "Résultats") a été gardée, puis les mots-clés ont été recherchés à nouveau sur cette partie. Sur les 3 255 textes initiaux, seuls 1 282 textes ont été gardés. En effet, parmi les textes non retenus sont concernés les revues littéraires, les résumés seuls, mais aussi les documents qui ne possèdent pas nos mots-clés dans la partie "Résultats".

Au début de notre expertise, nous avons utilisé un corpus d'entraînement constitué de cinq articles scientifiques (parmi les 1 282). Ces derniers ont alors été annotés manuellement par nous-même afin d'extraire les éléments principaux présents dans la constitution d'une phrase clé de le partie "Résultats".

L'intérêt de ne garder que la partie "Résultats" dans notre phase de tri concerne le fait que certaines phrases recherchées à l'aide de nos graphes pourraient être retrouvées, soit dans l'introduction, qui fait l'état de l'art de l'article ou dans la partie "Discussion" où suite aux résultats démontrés, les auteurs font leur propre interprétation et leurs propres hypothèses. Or, ce n'est pas ce que le chercheur vise. Ainsi, pour éviter toute recherche de phrase hors cadre, nous réalisons un "cache" en amont et en aval de la partie d'intérêt. Dès lors, nous ciblons les

2. http://www.nactem.ac.uk/medie/

phrases mettant en relation deux protéines minimum et utilisant des verbes démonstratifs. Par exemple, pour l'article de (Wang et al., 2005), nous cherchons les phrases ci-dessous :

- *We found that only phosphorylated ERK bound to Cdc25A.*
- *These data provide evidence that the Cdc25A-ERK interaction can be independent of EGFR activation.*
- *As shown in Figure 1B, GST-Cdc25A bound to ERK in vitro, whereas glutathione almost completely blocked this binding.*

Ces phrases font directement référence à l'article lui-même, soit en mentionnant la figure démontrant la phrase conclusive "*As shown in Figure 1B*", soit en résumant les diverses démonstrations citées dans le paragraphe "*These data provide*".

A l'inverse, nous voulons éviter de retrouver des phrases parasites. Parmi celles-ci, nous retrouvons les phrases descriptives annonçant ce que les auteurs ont l'intention de montrer, par exemple :

- *We next examined whether Cpd 5-induced ERK phosphorylation can be independent of MEK, its direct up-stream kinase activator.*

Mais aussi des phrases qui font le lien entre des résultats précédemment publiés ou de la bibliographie directement telles que :

- *We previously reported that Cpd 5, a Cdc25A inhibitor, caused prolonged EGFR activation, which in turn triggered ERK phosphorylation and cell growth inhibition (Wang et al., 2000, 2002).*

Une fois cette annotation manuelle terminée, nous avons utilisé un outil spécialisé dans l'analyse de corpus : Unitex [3] (Paumier, 2003).

2 Méthode

Le logiciel Unitex permet l'analyse lexicale, multilingue et grammaticale de corpus. De plus, c'est une plate-forme en libre accès et qui est régulièrement mise à jour par les modifications apportées par les divers utilisateurs et contributeurs.

2.1 Prétraitements

Une des fonctions principales concerne la création et l'application de dictionnaires spécifiques pour l'analyse des corpus d'intérêts. Dans le cas du projet Biosystémique, quatre dictionnaires ont été nécessaires pour prendre en compte la totalité des termes propices à notre analyse. Nous retrouvons notamment celui concernant les protéines (issu de la base de données UniProt), mais aussi celui rassemblant les diverses techniques mises en place en laboratoire, ainsi que tout le lexique inhérent au domaine (issu de la bibliographie ainsi que des sites spécialisés) et, enfin, les dictionnaires comprenant les divers systèmes cellulaires et les composés chimiques usuels (issus des sites spécialisés et du NCBI). Chaque terme du dictionnaire peut être qualifié par autant d'attributs que nécessaire. Par exemple (tableau 1), le verbe "confirm" se trouve dans le dictionnaire "Biosystemic" (V+Biosystemic), c'est un verbe démonstratif (+demonstration) et au présent (:P).

Ce dictionnaire est fléchi et le graphe de la figure 1 complète la flexion verbale par la reconnaissance des formes conjuguées avec des auxiliaires.

3. `http://www-igm.univ-mlv.fr/~unitex/`

Dictionnaire	Entrée
Biosystemic	co-elutions, co-elution.N+Biosystemic+experimentation:p confirm,confirm.V+Biosystemic+demonstration:P
Cell	CAKI-1,.Cell+kidney_carcinoma CCRF-CEM,.Cell+T_cell_leukemia
Compound	BAPTA/AM,.Compound:s:p carvedilol,.Compound:s:p
Protein	Spy1,Speedy protein A.Protein+Biosystemic:s:p Src family-associated phosphoprotein 1, Src kinase-associated phosphoprotein 1.Protein+Biosystemic:s:p

TAB. 1 – *Quelques exemples extraits des quatre dictionnaires spécifiques.*

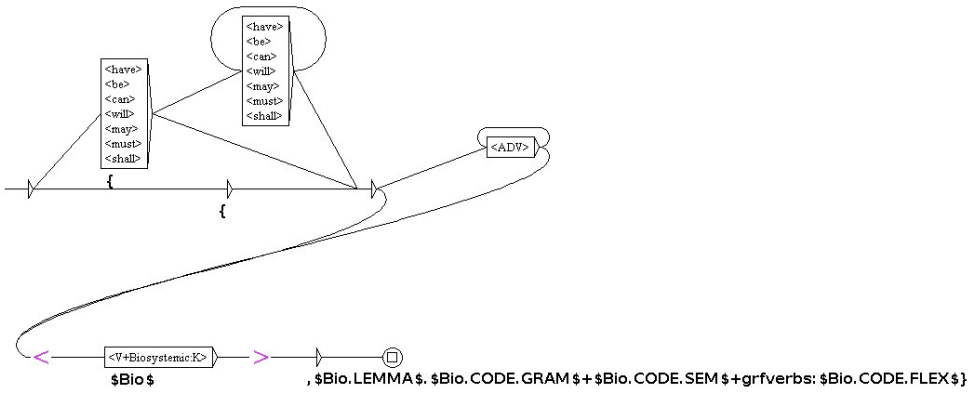


FIG. 1 – *Graphe des groupes verbaux.*

Enfin, certaines formes polylexicales spécifiques aux anticorps ne figurent pas dans le dictionnaires, mais sont reconnues par le graphe de la figure 2.

Cependant ces deux graphes sont précédés du graphe de découpage en phrases, version anglaise inspirée de Friburger et al. (2000).

2.2 Graphes d'analyse

Suite à l'annotation manuelle qui nous a permis de mettre en lumière les constructions de phrases récurrentes dans les articles et la création des dictionnaires, nous avons créé trois graphes pour remplir nos objectifs. Un de ces graphes est présenté figure 3.

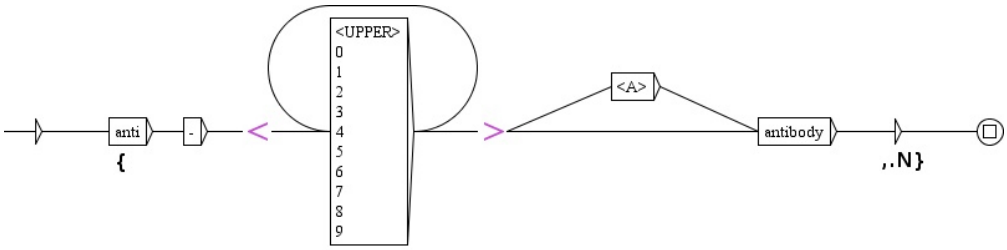


FIG. 2 – *Graphe des formes polylexicales spécifiques aux anticorps.*

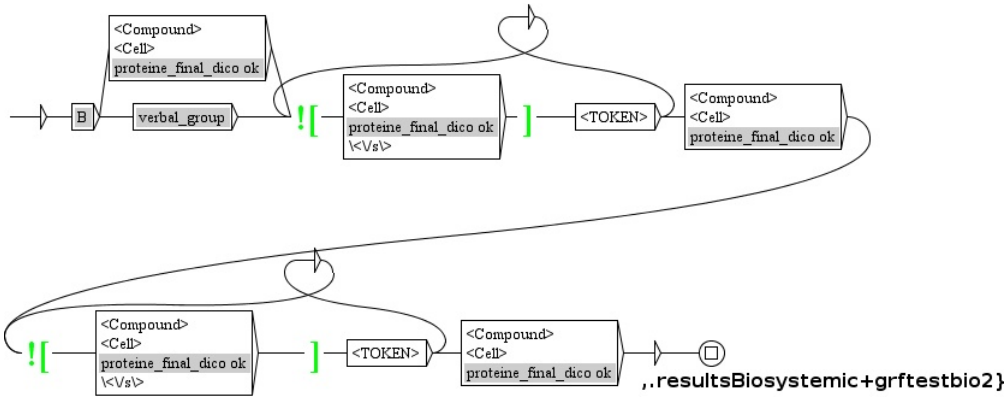


FIG. 3 – *Un des graphes désignant les relations entre les protéines.*

2.3 Cascade

Tous ces graphes (de prétraitement et d'analyse des relations entre protéines) sont passés successivement sur les textes à traiter suivant le principe de cascade (Abney, 1996; Friburger et Maurel, 2004), rappelé figure 4. Le menu *CasSys* d'Unitex permet le passage de graphes en cascade.

Notre cascade est subdivisée en trois sous-cascades, comme présenté tableau 2.

À la fin de la première sous-cascade, toutes les phrases d'intérêt sont identifiées en entier et le texte est balisé au format XML propre à CasSys.

Dans la deuxième sous-cascade, les balises extra-numéraires sont supprimées ainsi que les phrases qui pourraient être retrouvées par leur structure mais dont le sens n'est pas recherché. C'est pourquoi nous avons créé un graphe "exclusion" qui recense les termes de ces phrases. Les verbes trop descriptifs tels que *was collected*, *was probed* mais aussi les phrases commençant par *We next*, *To further* ou encore *In order* qui ne vont pas aboutir à des phrases conclusives y sont intégrés.

Enfin, la dernière sous-cascade élimine toutes les balises précédemment insérées et crée un fichier rassemblant les phrases clés à la suite les une des autres.

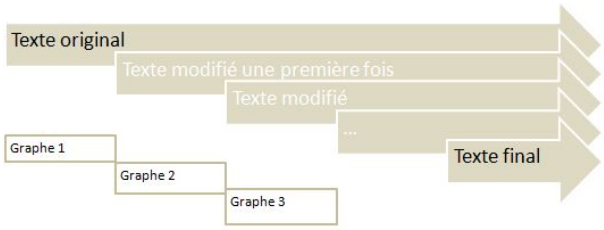


FIG. 4 – *Principe d'une cascade de graphes.*

Analyse	Synthèse	Extraction
toolPhraseTei	suppression	extract
verbs	balisage	final
polylexical	exclusion	
test_bio		
no_verb		
no_protein		
sentenceResult		

TAB. 2 – *Présentation des cascades.*

3 Résultats

3.1 Choix du corpus

Une fois le travail d'annotation et de constitution de graphes réalisés, un corpus d'évaluation a été constitué. La première question revenait de savoir l'impact du choix du format du texte (XML ou PDF) sur l'analyse car tous les textes ne sont pas disponibles au format (XML/HTML). Nous nous sommes donc tout d'abord focalisés sur des textes au format PDF. Suite à la conversion de ces textes en format ".txt" nécessaires à l'analyse par Unitex par un convertisseur en ligne [4], nous nous sommes rendus compte de la difficulté à les utiliser directement, car la structure du PDF a un impact sur sa conversion. En effet, les articles extraits sont présentés sur deux colonnes, ce qui ne pose pas de problème au convertisseur, sauf si les figures avec leur légende couvrent toute la largeur de la page ou empiètent partiellement sur les deux colonnes. La conversion va alors mélanger le texte de la légende de la figure avec le texte du paragraphe (Kim et al., 2009). La Figure 5 illustre ce phénomène avec des phrases comme :

– *FIGURE 1. Mobilization of calcium, DNA synthesis, and phosphorylation of ERK stimulated by ANG and contribution of -arrestin-mediSII in VSMCs. A, VSMCs were loaded with the calcium-binding dye...*

4. `http://document.online-convert.com/fr/convertir-en-txt`

FIGURE 1. **Mobilization of calcium, DNA synthesis, and phosphorylation of ERK stimulated by ANG and SII in VSMCs.** *A*, VSMCs were loaded with the calcium-binding dye Fura-2, and stimulated either with ANG (100 nM) or SII (10 μM) in the presence or absence of pretreatment with the AT1R antagonist (*AT1RB*) valsartan (50 μM) or AT2R antagonist (*AT2RB*) PD123319 (30 μM). Calcium fluorimetric traces are shown with the 340/380 nm excitation ratio (*y* axis) plotted as a function of time (*x* axis). Results displayed are mean ± S.E. of three independent experiments. *B*, VSMCs were serum starved for 24 h to arrest cycling and pretreated with dimethyl sulfoxide (*DMSO*) or MEK inhibitor PD98056 (20 μM). ANG (100 nM), SII (10 μM), or EGF (10 ng/ml) along with ^{3}H-labeled thymidine were then added to the media, and 24 h later cells were harvested as described under "Experimental Procedures." *NS* indicates no stimulation. Results depicted represent the mean ± S.E. of count per minute (cpm) values from four independent experiments. Statistical analysis was done using a one-way ANOVA (PRISM software) to correct for multiple comparisons (Bonferroni's multiple comparison test) with post test. The PD98056-pretreated condition shows significant reduction compare with the dimethyl sulfoxide-pretreated condition for each stimulant (*, $p < 0.05$). *C*, VSMCs with endogenous AT1R were treated with 100 nM ANG or 10 μM SII for the indicated times. Equal amounts of cell lysate were separated by SDS-PAGE and analyzed for phosphorylated ERK (*p-ERK*) and total ERK (*ERK*) by Western blotting. *IB*, immunoblot. *D*, signals were quantified by densitometry and p-ERK was normalized to a loading control (ERK). p-ERK activation was expressed as percentage of the maximal phosphorylated ERK obtained by ANG stimulation at 5 min. Each data point represents the mean ± S.E. from eight independent experiments.

contribution of β-arrestin-mediated ERK activation to VSMC proliferation, we studied ANG- and SII-mediated DNA synthesis in VSMCs as an indicator of cell proliferation. VSMCs were pretreated or not with a pharmacological inhibitor (PD98059) of mitogen-activated protein kinase kinase (MEK) followed by stimulation with 100 nM ANG, 10 μM SII, or 10 ng/ml EGF. Although SII cannot activate classical G-protein-dependent Ca^{2+} fluxes (Fig. 1*A*), it significantly increased DNA synthesis (thymi-

FIG. 5 – *Illustration d'une conversion de PDF où légende et paragraphe se mélangent.*

C'est suite à ces observations que trois analyses distinctes ont été réalisées :

1. le texte brut converti ;
2. le texte converti et réorganisé manuellement, mais en incluant les légendes ;
3. le texte converti et réorganisé manuellement en supprimant les légendes.

Pour chacun des quinze textes, deux mesures sont calculées :

- la mesure de rappel qui correspond au ratio du nombre de phrases correctes retrouvées par rapport au nombre de phrases attendues ;
- la mesure de précision qui représente le ratio du nombre de phrases correctes retrouvées par rapport au nombre de phrases totales trouvées ;

Nous avons constaté qu'après cette première analyse, les mesures de rappel et de précision augmentent fortement dès lors que le texte est réorganisé. La mesure de rappel passe donc de 0,66 (texte brut) à 0,9 (texte réorganisé) tandis que la précision est presque doublée passant de 0,48 à 0,8 (cf. tableau 3).

	PDF		
	brut	avec légende	sans légendes
Rappel	0,66	0,90	0,90
Precision	0,48	0,6	0,80

TAB. 3 – *Rappel et précision suivant les divers formats de textes analysés.*

3.2 Évaluation

Pour l'évaluation, nous avons constitué un corpus de vingt-sept articles pris au hasard parmi les 1 282 documents disponibles. Ce corpus représente des articles divers et variés en terme d'année de parution, de nationalité de l'auteur et de journal de publication.

Les mesures de rappel et de précision sont présentées tableau 4. Il est à noter que seuls vingt-sept textes ont été utilisés pour l'évaluation, mais que les 1 282 documents ont été analysés, ce qui nous donne au total 62 655 phrases extraites.

Malgré des mesures de rappel et de précision très encourageantes dans cette phase préliminaire, il est toujours possible d'augmenter ces valeurs. D'une part, nous avons dû apporter une légère correction au graphe des fins de phrase, en ajoutant la possibilité d'un début de phrase par un chiffre suivi d'une minuscule, ce qui n'avait pas été prévu, mais peut correspondre au nom d'une protéine. D'autre part, le passage du corpus de test (5 articles) au corpus d'évaluation (27 articles), a posé la question d'ajouter ou non les phrases commençant par *when* et celles contenant les mots *previously* et *et al.* qui correspondent à environ 1/3 des phrases manquantes et qui avait été éliminée par l'annotateur. En effet, dans les articles tests, ces mots semblaient diminuer le rappel. A l'avenir, il revient donc de savoir si le rétablissement de ces termes va augmenter la valeur de rappel sans impacter négativement la mesure de précision.

	Textes
Rappel	0,90
Precision	0,81

TAB. 4 – *Rappel et précision sur les vingt-sept textes du corpus d'évaluation.*

4 Conclusion

L'objectif initial de ce projet est la recherche de résultats expérimentaux dans les publications scientifiques concernant la biologie systémique afin de détecter les voies de signalisation des RCPG.

Avec l'étendue des données connexes à la biologie systémique, la nécessité de développer une méthode automatique d'analyse textuelle complémentaire à l'analyse manuelle est devenue une évidence. Cette combinatoire renforce le domaine des connaissances. Avec notre approche, nous nous affranchissons des interprétations des auteurs et laissons la place à de nouvelles hypothèses, autres que celles présentées.

L'évaluation que nous avons réalisée a permis de mettre en lumière un rappel et une précision équivalentes entre un texte non balisé mais réorganisé (le PDF revu) et un texte balisé (en XML ou HTML). Cependant, il revient de noter que, dans la mesure du possible, il est fortement recommandé d'utiliser le format balisé car, d'une part, il est moins chronophage que le texte soumis à conversion et, d'autre part, l'analyse du corpus est alors complètement automatique.

Après l'évaluation, nous avons appliqué notre cascade sur un plus gros volume de textes, en téléchargeant tous les articles en format balisé présents dans ISTEX et le NCBI avec les mots-clés *ERK*, *arrestin* et *phosphorylation*, soit 3 255 documents. Comme ces trois mots-clés n'étaient pas forcément présents dans notre partie d'intérêt (la partie "Résultats" de l'article), il a été nécessaire de réaliser un filtrage sur le corpus afin d'éliminer les textes qui ne correspondaient pas à notre attente (soit près de 40 % du corpus initial). Les résultats obtenus vont nous permettre de faire de l'inférence automatique, c'est-à-dire de relier divers passages extraits de la lecture de plusieurs articles, en utilisant les 62 000 phrases retrouvées et en créant les règles permettant de construire le réseau de signalisation adéquat à notre recherche. En effet, dans notre approche préliminaire, toutes les phrases possèdant un lien avec nos mots-clés sont re-

trouvées. C'est lors de l'inférence automatique que le choix des phrases va être réalisé. Nous avons donc une approche large qui sera la cible des règles créées par l'utilisateur. A notre stade, nous avons développé un réel outil permettant au lecteur d'identifier les données expérimentales présentes dans un article suite à une requête. Les phrases clés de la partie "Résultats" de l'article seraient référencées et visibles facilement. Cette approche existe déjà sur la plateforme Textpresso [5] mais de manière contrainte (certaines espèces, certaines maladies, certains textes) et sans privilégier la partie "Résultats" de l'article.

Notre travail s'inscrit donc dans la continuité de la plateforme MEDIE et pourrait même la surpasser lorsque notre projet sera finalisé.

Pour conclure, l'exploration de textes en biologie est une analyse en pleine expansion. Bien que certains opus couvrent le domaine (Schölkopf et al., 2004; Tan et al., 2007; Han et al., 2012), de nombreuses études continuent à se développer. Notre travail s'inscrit dans cette thématique. Pour l'instant établi dans le domaine de la signalisation cellulaire, il pourra facilement être transposable à d'autres domaines de la biologie. En effet, la constitution des dictionnaires est immédiate à partir des bases de données existantes et la partie spécifique à la signalisation cellulaire correspond à trois graphes sur douze.

Remerciement

Le projet Biosystémique est financé par le projet ISTEX [6] (*Initiative d'excellence en Information scientifique et technique*).

Références

Abney, S. (1996). Partial parsing via finite-state cascades. In *Proceedings of the Workshop on Robust Parsing, 8th European Summer School in Logic, Language and Information*, pp. 8–15. Prague, Czech Republic.

Ananiadou, S., D. B. Kell, et J. Tsujii (2006). Text mining and its potential applications in systems biology. *Trends Biotechnol 24*(12), 571–579.

Bateman, A., M. J. Martin, C. O'Donovan, M. Magrane, R. Apweiler, et E. Alpi (2015). UniProt : A hub for protein information. *Nucleic Acids Research 43*(D1), D204–D212.

Friburger, N., A. Dister, et D. Maurel (2000). Améliorer le découpage des phrases sous intex. *Revue Informatique et Statistique dans les Sciences Humaines 36*(1–4), 181–200.

Friburger, N. et D. Maurel (2004). Finite-state transducer cascades to extract named entities in texts. *Theoretical Computer Science 313*(1), 93 – 104. Implementation and Application of Automata.

Gloaguen, P., P. Crépieux, D. Heitzler, A. Poupon, et E. Reiter (2011). Mapping the follicle-stimulating hormone-induced signaling networks. *Frontiers in Endocrinology 2*(OCT), 1–13.

5. http://www.textpresso.org/
6. http://www.istex.fr/istex-excellence-initiative-of-scientific-and-technical-information/

Han, J., M. Kamber, et J. Pei (2012). Data Mining : Concepts and Techniques. *Waltham : Morgan Kaufmann*.

Heiner, M., I. Koch, et J. Will (2004). Model validation of biological pathways using Petri nets - Demonstrated for apoptosis. *BioSystems 75*(1-3), 15–28.

Joshi-Tope, G., M. Gillespie, I. Vastrik, P. D&apos ;Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney, et L. Stein (2005). Reactome : A knowledgebase of biological pathways. *Nucleic Acids Research 33*(DATABASE ISS.), 428–432.

Kanehisa, M., S. Goto, S. Kawashima, Y. Okuno, et M. Hattori (2004). The KEGG resource for deciphering the genome. *Nucleic acids research 32*(Database issue), D277–80.

Kell, D. B. et S. G. Oliver (2004). Here is the evidence, now what is the hypothesis ? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *BioEssays 26*(1), 99–105.

Kemper, B., T. Matsuzaki, Y. Matsuoka, Y. Tsuruoka, H. Kitano, S. Ananiadou, et J. Tsujii (2010). PathText : A text mining integrator for biological pathway visualizations. *Bioinformatics 26*(12), 374–381.

Kim, J., S. Ahn, K. Rajagopal, et R. J. Lefkowitz (2009). Independent β-arrestin2 and Gq/protein kinase Cζ pathways for ERK stimulated by angiotensin type 1A receptors in vascular smooth muscle cells converge on transactivation of the epidermal growth factor receptor. *Journal of Biological Chemistry 284*(18), 11953–11962.

Klein, T. E., J. T. Chang, M. K. Cho, K. L. Easton, R. Fergerson, M. Hewett, Z. Lin, Y. Liu, S. Liu, D. E. Oliver, D. L. Rubin, F. Shafa, J. M. Stuart, et R. B. Altman (2001). Integrating genotype and phenotype information : an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. *The pharmacogenomics journal 1*(3), 167–170.

Luciano, J. S. et R. D. Stevens (2007). e-Science and biological pathway semantics. *BMC bioinformatics 8 Suppl 3*, 1–21.

Mi, H., N. Guo, A. Kejariwal, et P. D. Thomas (2007). PANTHER version 6 : Protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic Acids Research 35*(SUPPL. 1), 247–252.

Paumier, S. (2003). *De la Reconnaissance de Formes Linguistiques é l'Analyse Syntaxique*. Thèse de Doctorat en Informatique, Université de Marne-la-Vallée.

Romero, P., J. Wagg, M. L. Green, D. Kaiser, M. Krummenacker, et P. D. Karp (2005). Computational prediction of human metabolic pathways from the complete human genome. *Genome biology 6*(1), R2.

Schölkopf, B., K. Tsuda, et J.-P. Vert (2004). Kernel Methods in Computational Biology. *Cambridge : The MIT Press*.

Tan, P., K. Steinbach, et V. Kumar (2007). Introduction to Data Mining. *Pearson Education*.

Tari, L., S. Anwar, S. Liang, J. Hakenberg, et C. Baral (2010). Synthesis of pharmacokinetic pathways through knowledge acquisition and automated reasoning. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing 476*, 465–76.

Wang, Z., B. Zhang, M. Wang, et B. I. Carr (2005). Cdc25A and ERK interaction : EGFR-

independent ERK activation by a protein phosphatase Cdc25A inhibitor, Compound 5. *Journal of Cellular Physiology 204*(2), 437–444.

Ye, Y. et T. G. Doak (2011). A Parsimony Approach to Biological Pathway Reconstruction/Inference for Metagenomes. *Handbook of Molecular Microbial Ecology I : Metagenomics and Complementary Approaches 5*(8), 453–460.

Summary

In the scientific domain, the literature is a useful tool for the knowledge. However, the advancement in the research and publication of the scientific documents has progressed exponentially. In this sens, it has become more and more labourious for a researcher to stay up to date in his/her domain of expertise. In order to overcome this difficulty, the project Biosystemic permits to develop a method to extract the experimental results in the scientific publications concerning the systemic biology, and more precisely concerning the signalisation pathways of the G-protein coupled receptors (GPCR). Subsequently, these data will be used in an inference engine to create a signaling network and thus connect data that appeared to be independent. In this article, we have showed the possibility of extracting key phrases of the article without the interpretation of the author.

Archives numériques et fouille de textes : le projet ISTEX

Pascal Cuxac*, Nicolas Thouvenin*

*INIST-CNRS
2, allée du parc de Brabois
CS 10310
54519 Vandœuvre lès Nancy Cedex

(prénom).(nom)@inist.fr

Résumé. A travers cet article nous souhaitons à la fois faire connaître la ressource ISTEX à des fins de "Text Mining" mais également les traitements appliqués à une bibliothèque numérique de grande taille. Plusieurs challenges sont à relever, dont le passage à l'échelle sur plus de 18 millions de documents, l'intégration de différents outils dans une même chaîne de traitement, et la modélisation des données obtenues pour leur mise à disposition.

1 Introduction

A l'ère de l'Internet nous assistons au développement des données en libre accès (OpenData), des collections issues de bibliothèques traditionnelles sont maintenant accessibles librement : Gallica, Europeana, Digital Public Library of America. À ce type de bibliothèques numériques s'ajoutent les publications savantes, qui occupent une part non négligeable des publications numériques. De récentes initiatives nationales ont également permis le développement d'importantes archives scientifiques (ISTEX en France, SwissBib en Suisse, GBV en Allemagne, Scholars Portal en Ontario). Le projet ISTEX [1] (initiative d'excellence en Information Scientifique et Technique) a pour objectif de permettre à la communauté ESR française (Enseignement Supérieur et Recherche) d'accéder, via un accès en ligne, à une bibliothèque numérique regroupant l'essentiel des publications scientifiques mondiales dans toutes les disciplines scientifiques. Ce projet offre tous les moyens de recherche d'information et d'accès aux documents en texte intégral. La plate-forme ISTEX fournit l'ensemble de ses services sous la forme d'une API Web [2] mais également via un démonstrateur [3] qui permet de se familiariser avec les formats et la syntaxe d'interrogation. Ce réservoir de publications scientifiques est bien entendu à destination des documentalistes et chercheurs ayant un besoin documentaire mais est également une ressource unique pour tous les chercheurs gravitants autour des thématiques de la fouille de texte, du TAL, de la Recherche d'Information...etc. Cet axe recherche/développement autour de la plateforme ISTEX s'est concrétisé par un appel à projet

1. http://www.istex.fr/
2. https://api.istex.fr/documentation/
3. http://demo.istex.fr/

«Chantiers d'usages» [4] afin de «*Créer une dynamique de recherche/développement autour de la plateforme ISTEX qui puisse servir de déclencheur à des activités plus larges d'appropriation par les chercheurs des contenus d'ISTEX pour développer des recherches de Text and Data Mining (TDM) de qualité.*» (Pierrel, 2016). Mais il est également intégré à la plateforme à travers le projet d'enrichissements des données mené par une équipe de l'INIST-CNRS en collaboration avec le LI de Tours [5], le LINA de Nantes [6] et ScienceMiner [7], dans le but d'intégrer dans les données ISTEX des enrichissements complémentaires à partir du plein texte et à l'aide de plusieurs outils ou méthodes issus de la recherche pour les mettre à disposition d'autres projets ou initiatives. Enfin, trois types de services à valeur ajoutée sont développés par différents partenaires universitaires : un moteur de réponse offrant des outils de classification automatique (CLEE/IRIT, Toulouse) [8], la caractérisation de l'évolution des recherches et des connaissances dans le temps grâce à la construction de cartes diachroniques (LORIA-ATILF-INIST, Nancy), une bibliothèque open source de composants XML d'exploitation du corpus ISTEX (Université de Lorraine).

Dans cet article nous aborderons uniquement les traitements visant à enrichir les données ISTEX et accessibles via l'API ISTEX. Les technologies déployées dans l'API proprement dite ne seront pas abordées ici, toutefois l'utilisateur pourra trouver un outil de moissonage en ligne de commande développé en NodeJS à l'adresse suivante `https://github.com/istex/istex-api-harvester`. Dans la suite de cet article nous allons passer brièvement en revue tous les programmes utilisés pour enrichir les données ; ils doivent répondre à un certain nombre d'exigences dont le passage à l'échelle sur plus de 18 millions de documents (optimisation des temps de traitements), l'intégration dans une même chaîne de traitement (compatibilité des programmes), des données en sortie au format TEI.

2 La catégorisation des documents

ISTEX contient actuellement 18,2 millions de documents en texte intégral couvrant tous les domaines de recherche ; un marquage de tous ces textes par une ou plusieurs catégories scientifiques est rapidement apparu nécessaire à la fois pour le comité de pilotage, afin d'avoir des statistiques sur le fonds, et également pour les utilisateurs afin de cibler un domaine particulier lors de l'interrogation. A cette fin deux approches complémentaires ont été implémentées :

- une catégorisation par appariement. Le principe en est simple puisqu'il s'agit de mettre en correspondance un identifiant de publication (ISSN ou ISBN par exemple) avec une ou plusieurs catégories attribuées à cette publication par un organisme reconnu.

A ce jour, deux ressources ont été choisies : celle du Web of Science [9] et celle de ScienceMetrix [10] (Archambault et al., 2011). Mais les domaines scientifiques attribués à une revue ne sont pas toujours adaptés à catégoriser tous les articles de la même revue. C'est pour cela que nous avons complété ces résultats par ceux obtenus en utilisant une méthode de classification

4. `http://www.istex.fr/seminaire-technique-25-et-26-avril-2016/`
5. `http://tln.li.univ-tours.fr/`
6. `http://www.lina.univ-nantes.fr/-TALN-.html`
7. `http://science-miner.com/`
8. `http://cillex.kodexlab.com`
9. `http://ip-science.thomsonreuters.com/mjl/scope/scope_scie/`
10. `http://www.science-metrix.com/fr/classification`

```
−<revisionDesc>
    <change when="24-06-2016" who="istex-rd" xml:id="rd-multicat">catégorisation par appariement</change>
  </revisionDesc>
</teiHeader>
<listAnnotation type="rd-multicat">
−<annotationBlock xmls="https//www.tei-c.org/ns/1.0">
  −<keywords change="#istex-rd" resp="#istex-rd" scheme="#wos">
      <term level="1">SOCIAL SCIENCE</term>
      <term level="2">TRANSPORTATION</term>
    </keywords>
  −<keywords change="#istex-rd" resp="#istex-rd" scheme="#wos">
      <term level="1">SCIENCE</term>
      <term level="2">TRANSPORTATION SCIENCE & TECHNOLOGY</term>
      <term level="2">ENGINEERING, CIVIL</term>
    </keywords>
  −<keywords change="#istex-rd" resp="#istex-rd" scheme="#science-metrix">
      <term level="1">ECONOMIC & SOCIAL SCIENCES</term>
      <term level="2">ECONOMICS & BUSINESS </term>
      <term level="3">LOGISTICS & TRANSPORTATION</term>
    </keywords>
  </annotationBlock>
</listAnnotation>
```


FIG. 1 – *Un exemple de catégories Wos et Science-Metrix associées à un objet documentaire*

supervisée. Une illustration du résultat au format TEI est donnée figure 1.

- une catégorisation par apprentissage automatique. Nous avons développé le module RD-NB basé sur un Bayésien naïf avec un apprentissage sur les bases PASCAL/FRANCIS. Le module s'appuie sur un apprentissage en cascade : il commence par déterminer si le document est SHS (sciences humaines et sociales = FRANCIS) ou STM (sciences techniques et médecine = PASCAL), puis par exemple dans le cas de STM l'étape suivante sera de déterminer si on est dans le cadre des sciences de la vie ou non et ainsi de suite (figure 3) .

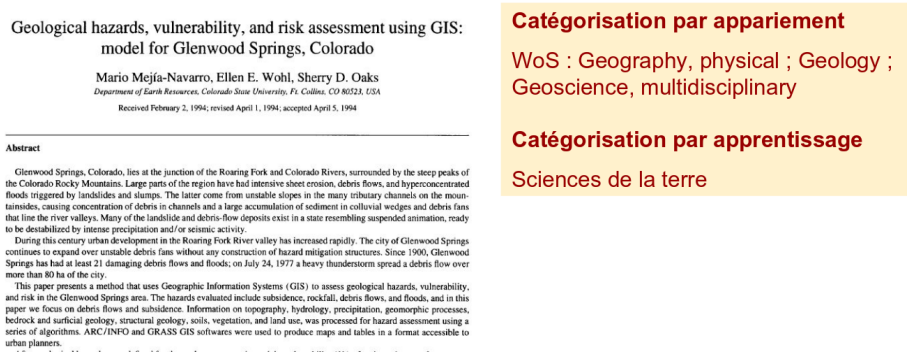


FIG. 2 – *Un article catégorisé par appariement et par apprentissage automatique*

La figure 2 donne un exemple d'article catégorisé à la fois par la méthode par appariement et par apprentissage automatique. La catégorisation automatique par apprentissage apporte un plus quand la catégorie associée à la revue est trop générique ou quand celle-ci n'existe pas.

Nous avons également testé des classifications s'appuyant sur une représentation vectorielle des textes de type "word embedding" : - un SVM couplé à doc2vec (Le et Mikolov, 2014) - l'algorithme FastText et sa classification basée sur un softmax hiérarchique (Joulin et al., 2016) Cette dernière solution donne des résultats plus performants en termes de temps d'exécution et devrait nous permettre d'avoir une classification plus précise que celle obtenue avec l'approche Bayésienne.

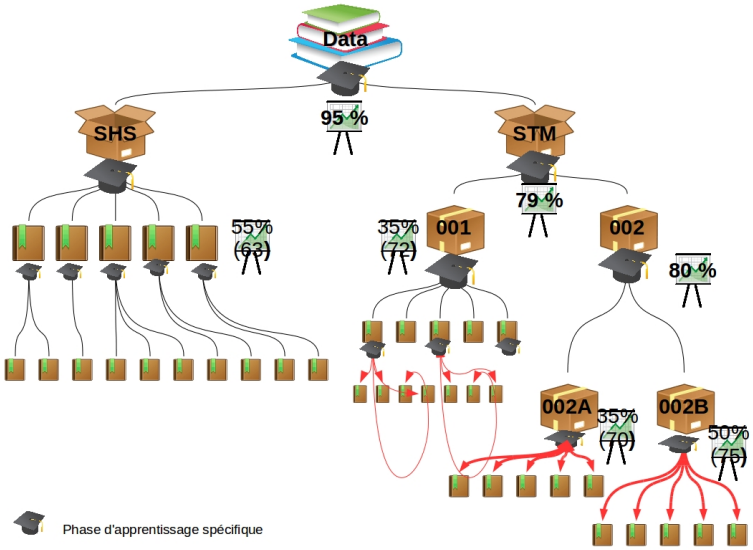


FIG. 3 – *La cascade d'apprentissage pour le Bayésien Naif*

3 Les entités nommées

Dans le cadre d'un partenariat avec le Laboratoire d'Informatique de l'Université François Rabelais de Tours [11] la plateforme Unitex [12] a été adaptée et complétée par un système de cascades de graphes CasSys afin de traiter de gros volumes de textes en français et en anglais. Les dix types suivants d'entités nommées ont été choisis pour être détectées et extraites :

— Les noms de personnes : <persName>
— Les noms de lieux : <placeName> et <geogName>
— Les noms d'organisations : <orgName>
— Les dates : <date>
— Les organismes financeurs et projets financés : <orgName type= 'funder'>

11. http://tln.li.univ-tours.fr/
12. http://www-igm.univ-mlv.fr/~unitex/

— Les URL : <ref type='url'>

— Les citations : <ref type='bibl'> et les références bibliographiques : <bibl>

— Les organismes hébergeurs de ressources : <orgName type='provider'>

Actuellement, plus de 12 millions de documents ont été traités pour un total de plus de 362 millions d'entités détectées.

4 Les références bibliographiques

Dans le réservoir ISTEX un nombre important de documents en texte intégral n'est accessible qu'en format non structuré. Dès le début du projet il est apparu important de pouvoir détecter et structurer les références bibliographiques afin de les rendre 'cliquables' et pouvoir ainsi naviguer dans le réservoir ISTEX ou faire le lien avec d'autres ressources extérieures. Nous avons pour cela utilisé l'outil Grobid (GeneRation Of BIbliographic Data) [13] développé par Science-Miner (Lopez, 2009). Partant du pdf, l'outil va utiliser des CRF (Conditional Random Fields) en cascade pour découper le document et baliser les références bibliographiques. Un ré-entraînement complet a permis d'atteindre une Fmesure de 0,76.

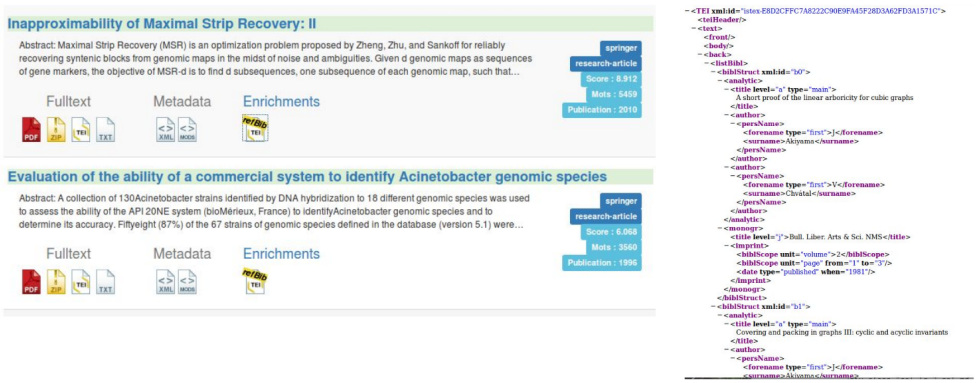


FIG. 4 – *Affichage des références bibliographiques détectées*

La figure 4 illustre la prise en compte dans l'API des références bibliographiques détectées, mises en évidence par un logo et en parallèle le format TEI de représentation utilisé. Pour chaque référence détectée nous avons un découpage en auteur, titre, année, publication...etc.

5 L'indexation

L'indexation automatique de documents sélectionne -par des méthodes logicielles- des termes extraits d'un texte pour donner une représentation de ce texte. Cette indexation peut être utilisée pour aider à la recherche de documents pertinents mais également dans des tâches

13. `http://grobid.readthedocs.io/en/latest/`

de classification qui pourraient être appliquées à des sous corpus. Du fait de la diversité des documents, une indexation supervisée impliquant des ressources spécialisées couvrant tous les domaines n'est pas envisagée. Deux méthodes ont été adaptées à notre chaîne de traitement :
- RD-TEEFT [14] est un outil d'indexation développé par l'équipe Istex-RD ; il traite les documents plein texte en format texte en anglais pour produire une liste de termes extraits et leur spécificité.
- Keyterm (Lopez et Romary, 2010) traite des documents pdf français ou anglais et utilise Wikipedia dans une étape de désambiguïsation. Keyterm est actuellement en cours d'intégration.

6 La structuration XML/TEI des documents

Pour la plupart des documents existe différents formats : mods (métadonnées), pdf, texte, XML (TEI). Pour l'instant le XML-TEI d'un document ne structure pas le corps du document qui se retrouve en format texte dans une balise «<body>». Avoir des documents entièrement au format XML-TEI aurait des intérêts multiples : faciliter les manipulations de textes pour l'utilisateur, permettre un 'nettoyage' des documents avant divers traitements (par exemple, ôter les figures, tableaux, en-têtes...afin d'avoir une extraction d'entités nommées, ou une indexation plus performante), utiliser la structure même du document pour améliorer certains algorithmes. Par exemple tenir compte de la structure d'un document peut améliorer significativement son indexation : nous avons testé deux approches l'une en pondérant différemment les termes extraits suivant leur position dans le texte, l'autre en mettant en concurrence les parties du texte. Dans cette dernière approche (figure 5) nous considérons le document par parties (le document peut être assimilé à une succession de classes) puis nous appliquons une sélection de variables inspirée de la méthode développée par J.C. Lamirel (Lamirel et al., 2010) pour terminer par une étape de re-pondération des termes.

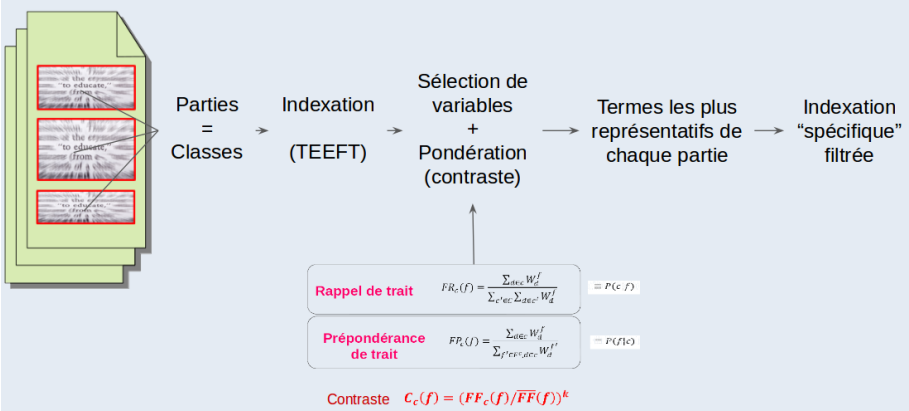


FIG. 5 – *Schémas d'indexation prenant en compte les parties d'un document*

14. RD-TEEFT en Node.js est dérivé de TEEFT, un programme python dont un exécutable peut être chargé sur le site `https://sites.google.com/site/pascalcuxac/outils`

Cette méthode est encore en phase expérimentale [15]. Si les performances en termes de vitesse de traitement sont bonnes (10 documents plein textes traités en 3 secondes), il reste à valider les améliorations apportées et surtout arriver à produire un XML de qualité à partir d'un pdf. L'outil Grobid peut être adapté pour faire ce traitement à condition d'avoir un corpus d'apprentissage performant.

7 Le reversement des données

A travers l'API ISTEX, les données sont déchargeables aux formats pdf, texte et XML-TEI. Tous les enrichissements produits et reversés doivent pouvoir être visibles par l'utilisateur final voire être interrogeables. Nous avons décidé d'enrichir le document XML-TEI : pour chaque document les enrichissements produits sont placés dans une balise <standOffs> après les métadonnées du document et conforme aux standards de la TEI (Text Encoding Initiative) [16]. La TEI est un format XML de description de textes. Elle permet de décrire la structuration du texte tel qu'il a été conçu et non son rendu final (figure 6).

```
<!--exemple_tei_categorisation_istex_rd....-->
- <standOff>
  - <teiHeader>
    + <fileDesc></fileDesc>
    - <revisionDesc>
      - <change when="24-06-2016" who="istex-rd" xml:id="rd-nb">
          catégorisation par approche statistique - Bayésien Naïf
        </change>
      </revisionDesc>
    </teiHeader>
  - <listAnnotation type="rd-nb">
    - <annotationBlock corresp="abstract" xmlns="https//www.tei-c.org/ns/1.0">
      - <keywords change="#rd-nb" resp="#istex-rd" scheme="http://inist-category.lod.istex.fr">
          <term cert="0.651637563297372" key="refCode : STM" level="1">SCIENCES APPLIQUEES, TECHNOLOGIE ET MEDECINE</term>
          <term cert="0.5541006144477459" key="refCode : 001" level="2">SCIENCES EXACTES ET TECHNOLOGIE</term>
          <term cert="0.20272618300842637" key="refCode : 001E" level="3">TERRE, OCEAN, ESPACE</term>
          <term cert="0.31605416143215126" key="refCode : 001E02" level="4">GEOPHYSIQUE EXTERNE</term>
        </keywords>
      </annotationBlock>
    </listAnnotation>
  </standOff>
```

FIG. 6 – *Illustration du format "standOff" TEI utilisé pour les enrichissements des données ISTEX*

8 Conclusion et perspectives

Cette présentation a pour but de faire connaître le réservoir ISTEX en tant qu'archive pour la recherche documentaire, mais aussi réservoir de documents en texte intégral dans toutes les disciplines scientifiques pour des développements d'applications/outils de TDM (Text and Data Mining). Si les traitements présentés sont assez standards, leur implémentation et exploitation sur de gros volumes de données hétérogènes (en termes de types de documents et de domaine scientifique) est un défi relevé. Nous avons illustré quelques développements en interne qui demandent encore, pour certaines méthodes, à être consolidés. A travers quatre

15. Un prototype exécutable est disponible sur `https://sites.google.com/site/pascalcuxac/outils`
16. `http://www.tei-c.org`

axes de travail (structuration des documents ; indexation automatique ; reconnaissance d'entités nommées ; catégorisation des documents) nous avons répondu aux trois principaux challenges rencontrés, c'est à dire :

— Mise au point et intégration des outils ;

  entraînement, configuration, adaptation, mise en production

— Passage à l'échelle ;

  20 millions de documents à traiter à terme

— Reversement des données ;

  modélisation des données, ré-intégration, mise à disposition

Dans un avenir proche, nous envisageons de mettre en place une plateforme dédiée à la fouille de textes connectée à ISTEX, accessible par Internet et qui permettra d'interroger le réservoir ISTEX afin d'en extraire un corpus de documents, puis de sélectionner un certain nombre d'outils qui s'exécuteront sur ce corpus. Ces outils pourront être par exemple des outils de visualisation, d'extraction terminologique, de classification, de statistiques, de graphes...Ils pourront être développés par les équipes ISTEX mais également en collaboration avec tout laboratoire désireux de faire partager une application.

9 Remerciements

Ces travaux ont été financés par le projet ISTEX avec le soutien de l'Agence Nationale pour la Recherche dans le cadre du programme d'Investissements pour le Futur de référence ANR-10-IDEX-0004-12. Je remercie les équipes ISTEX de l'INIST qui œuvrent au bon fonctionnement de la plateforme ISTEX et à son amélioration quotidienne, ainsi qu'aux différents partenaires cités dans cet article.

Références

Archambault, E., O. Beauchesne, et J. Caruso (2011). Towards a multilingual, comprehensive, and open scientific journal ontology. In *13th International Conference on Scientometrics and Informetrics held in Durban, Republic of South-Africa*.

Joulin, A., E. Grave, P. Bojanowski, et T. Mikolov (2016). Bag of tricks for efficient text classification. *arXiv:1607.01759 [cs]*. arXiv: 1607.01759.

Lamirel, J., M. Ghribi, et P. Cuxac (2010). Unsupervised recall and precision measures: a step towards new efficient clustering quality indexes. In *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010, Paris, France)*.

Le, Q. V. et T. Mikolov (2014). Distributed representations of sentences and documents. *arXiv:1405.4053 [cs]*. arXiv: 1405.4053.

Lopez, P. (2009). Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. *Research and Advanced Technology for Digital Libraries*, 473–474.

Lopez, P. et L. Romary (2010). Humb: Automatic key term extraction from scientific articles in grobid. In *SemEval 2010 Workshop, Jul 2010, Uppsala, Sweden*.

Pierrel, J. (2016). Séminaire chantiers d'usage. Communication orale.

Summary

Through this article we wish to make known the ISTEX resource for the purposes of *Text Mining* but also the treatments applied to a large digital library. Several challenges have to be met, including the scaling up with more than 18 million documents, the integration of different tools in a single processing chain, and the modeling of the data obtained to make them available.

Index

B

Bars, Rémi 3
Boullé, Marc 23
Bouraoui, Jean-Leon 13

C

Clérot, Fabrice 23
Cuxac, Pascal 43

G

Gay, Dominique 23
Guerraz, Aleksandra 3
Guigourès, Romain 23
Gupta, A. 31

L

Landomiel, F. 31
Legay, Nathalie 3
Lemaire, Vincent 13

M

Maurel, D. 31
Morbieu, Stanislas 8
Muhlenbach, Fabrice 1

N

Nadif, Mohamed 8

P

Poupon, A. 31

R

Role, François 8

T

Thouvenin, Nicolas 43