Extraire et classifier pour évaluer, comprendre et communiquer

Bénédicte Grailles, Touria Ait El Mekki

Extraire et classifier pour évaluer, comprendre et communiquer

Bénédicte Grailles, maîtresse de conférences en archivistique,
Université d'Angers, laboratoire Temos
Touria Aït El Mekki, maîtresse de conférences en informatique,
Université d'Angers, laboratoire Leria

Le programme Pêle-mél (Plateforme d'exploration, de livraison et d'évaluation des méls) a été l'occasion de tester des approches de traitement automatique de la langue naturelle reposant sur de l'extraction de termes, de relations sémantiques et des techniques d'apprentissage artificiel. Quels enseignements peuvent-ils en être tirés ?

Dans le cadre de ce programme, nous avons utilisé la terminologie pour accéder au contenu d'un corpus de méls provenant du cabinet de la ministre de la Santé entre 2007 et 2011, comprendre les sujets abordés et classifier ces messages en les reliant à des thématiques.

L'apport de la terminologie

La terminologie computationnelle vise automatiser des étapes de travail habituellement effectuées à la main, comme l'identification de mots-clés d'un domaine spécifique dans un texte. Dans notre cas, il s'agissait d'extraire des termes – une unité lexicale d'un ou plusieurs mots représentant un concept (par exemple, « durée d'utilité administrative ») – et des entités nommées – une personne, un organisme, un lieu, un événement – à partir des méls, des pièces jointes aux méls et des nommages. Ce repérage nécessite de pré-traiter le corpus et de convertir les fichiers de format divers en format texte. Ensuite, on utilise un « étiqueteur », un programme qui identifie, pour chaque mot de la phrase, les catégories grammaticales (déterminant, verbe, adjectif, adverbe etc.) produit une analyse morpho-syntaxique et donne les informations de lemmatisation (pour un verbe son infinitif, pour un substantif son singulier, pour un adjectif son masculin-singulier). Dans le cadre de notre projet, l'interface d'extraction produite (fig. 1) permet de choisir de lemmatiser, le nombre minimum et maximum de mots constituant le terme, et la méthode de *scoring* : fréquence (on compte simplement le nombre d'occurrences), TF-IDF (*term frequency-inverse document frequency*, une méthode de pondération qui permet d'évaluer l'importance et la pertinence d'un terme). Les entités nommées sont parallèlement extraites dans l'objectif de constituer une liste de noms de personnes, d'organismes et de leurs abréviations. À cette étape, il y a forcément du bruit. Une phase de validation, qui peut partiellement être automatisée, est indispensable.

La création de nuages regroupant des termes de sens proche

La seconde étape cherche à établir des relations sémantiques entre des termes ou des termes et des entités nommées. On peut, pour ce faire, s'appuyer sur des règles linguistiques et/ou sur de l'apprentissage automatique. Ce dernier peut être supervisé – il faut alors disposer de données d'entraînement préalablement étiquetées à la main – ou non supervisé. Nous avons combiné deux

approches, l'approche par patron lexico-syntaxique (automatisation de l'extraction de relations grâce à des schémas, comme par exemple, le schéma « Terme 1 + être + déterminant + Terme 2 » grâce auquel la relation hypernymique entre SIDA et maladie incurable est identifiable dans la phrase « Le SIDA [terme 1] est [être] une [déterminant] maladie incurable [terme 2] ») et l'approche symbolique non supervisée pour laquelle nous avons utilisé Word2Vec.

Word2Vec est une méthode de plongement lexical et un réseau de neurones artificiels à deux couches entraînées pour reconstruire le contexte linguistique des mots. C'est un modèle prédictif qui permet de prendre en compte le contexte dans lequel un mot a été trouvé. Chaque mot est représenté par un vecteur de nombres réels. Les mots utilisés dans des contextes similaires, supposés avoir des significations proches, sont représentés dans l'espace vectoriel par des vecteurs proches. Nous avons utilisé Word2Vec sur les entités et termes validés à l'étape précédente, en utilisant un modèle générique déjà pré-entrainé sur de larges corpus en français (fig. 2).

Ces deux méthodes permettent de relier différents termes et entités à un terme ou différents termes et entités à une entité. L'objectif est de faire émerger un nuage de termes et d'entités qui constitue l'aura sémantique d'un terme ou d'une entité. Il faut injecter en entrée des termes et en sortie on obtient un ensemble de termes et/ou d'entités dont on peut d'ailleurs varier la profondeur. Nous avons demandé aux archivistes expertes du domaine[1] de nous proposer une liste de termes correspondant aux missions, attributions et actions du ministère puis nous avons associé à chacun des 70 termes proposés le nuage correspondant.

Regrouper les messages par famille

Pour classer les messages, nous avons utilisé une méthode semblable à Word2Vec dite de plongement de documents. Cette fois, chaque message (le message + les pièces jointes + les nommages) est représenté par un vecteur. L'enjeu est ensuite de créer des relations entre ces vecteurs et les nuages de termes réalisés précédemment. Un même message peut être relié à différents thèmes. De cette manière, nous avons pu créer des clusters de messages associés à une thématique. On peut ensuite construire des graphes permettant une approche quantitative des thèmes des échanges et éditer la liste des messages pour lesquels la méthode prédit un lien avec la thématique. Il est possible de choisir le niveau de granularité de la classification : une seule boîte mél ou plusieurs boîtes.

Cette exploration de méthodes appuyées sur des réseaux de neurones artificiels et une démarche de traitement automatique de la langue adapté au français a permis de valider la pertinence d'une approche par plongement lexical et plongement de documents pour organiser de grandes masses de données archivées ou à archiver et augmenter la pertinence de la recherche. À partir des résultats obtenus, il est envisageable d'améliorer sensiblement non seulement l'accès par mots clés, puisque la recherche ne porte pas sur l'identification d'un mot précis mais sur ce mot et les termes dont le sens est proche, mais aussi la connaissance du contenu réel des messageries, des thèmes des conversations, de leur évolution dans le temps et donc de produire des descriptions plus pertinentes et précises. Il est également possible de s'appuyer sur les résultats pour effectuer des choix entre boîtes méls et du tri interne à chaque boîte. Ces méthodes peuvent donc être mobilisées à différentes étapes de la chaîne archivistique : évaluation archivistique, sélection, description.

1 Ce projet, soutenu par le ministère de la Culture, a été mené avec l'aide de la mission Archives du ministère de la Santé (Anne Lambert et Chloé Moser) et de l'École nationale des Chartes (Édouard Vasseur).

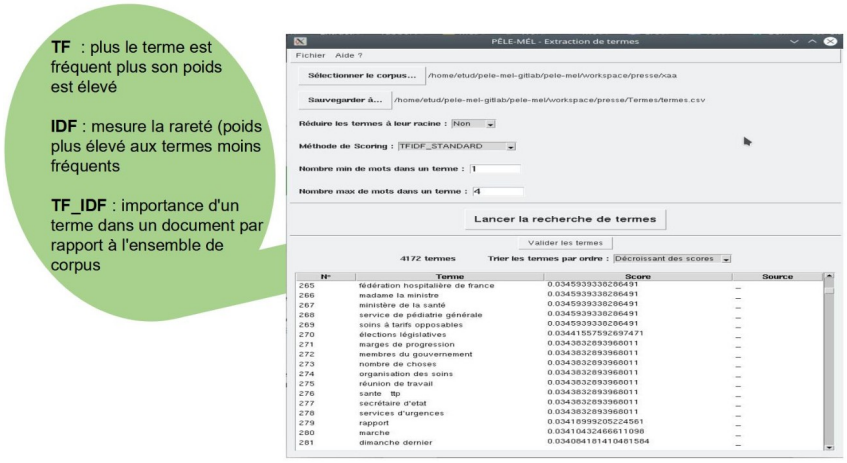

Fig. 1 Interface d'extraction des termes. Elle intervient après étiquetage morpho-syntaxique et lemmatisation, et permet de paramétrer la manière de mesurer l'importance d'un terme, ce qui aura une incidence sur sa vectorisation.

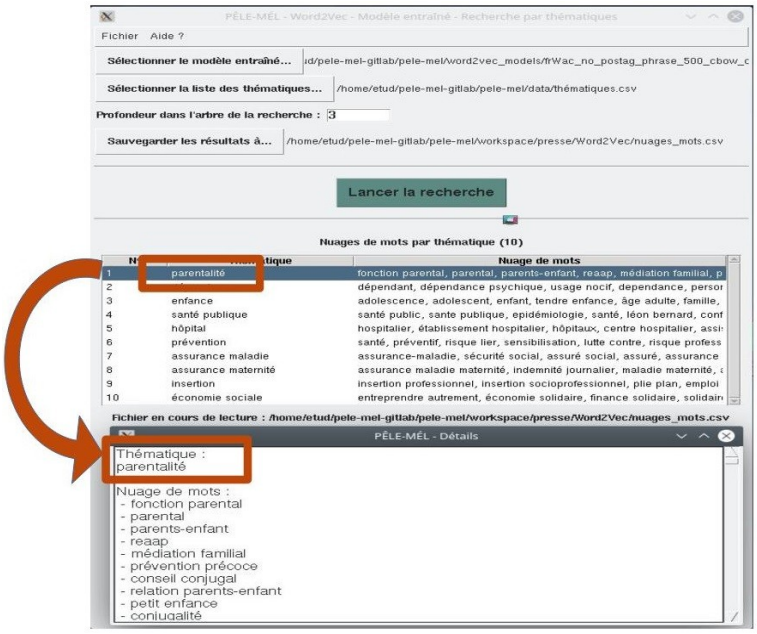

Fig. 2 Interface de création des nuages de termes par plongement lexical. À l'aide de Word2Vec et en sélectionnant un modèle pré-entraîné, des relations sont établies entre les termes extraits et lemmatisés (entités nommées incluses) et des thèmes pour constituer des nuages de mots. Ces nuages serviront à classer les messages, après vectorisation et par calcul de similarité.