Comparaison de méthodes sémantique et asémantique pour la catégorisation automatique de documents

Romaric BOLEY
Université de Montréal – École de Bibliothéconomie et des Sciences de l'Information
C.P. 6128, Succursale Centre-Ville, Montréal, Québec, Canada, H3C 3J7
romaric.boley@umontreal.ca

Résumé

Cet article présente les résultats obtenus pour les deux tâches proposées par le DEFT 2011. La première, variation diachronique, consistait à déterminer la date de parution d'un article de presse française. La deuxième tâche consistait à apparier un résumé avec l'article scientifique dont il a été extrait. L'objectif à travers ces études était de comparer, pour chacune des tâches, les démarches sémantique et asémantiques (extraction des longueurs des phrases et sélection de mots non porteurs de sens) pour la catégorisation automatique de documents.

Abstract

This paper presents the results for both tasks proposed by the DEFT2011. The first one, the diachronic variation, consisted of determining the publication date of French press articles. The second task consisted of matching a summary to the scientific paper from which it was extracted. The objective through these studies was to compare, for each of the tasks, the unsemantic (i.e. sentences length or stop words extraction) and semantic processes.

Mots-clés : catégorisation ; approche sémantique ; approche asémantique, fouille de texte
Keywords: classification ; semantic ; unsemantic ; text mining

1 Introduction

L'information de nos jours occupe une place centrale dans nos sociétés. L'ère industrielle a laissé place à l'ère de l'information. Du fait du volume croissant d'informations numériques non structurées, le domaine de la fouille de textes est en pleine expansion. Avec le développement d'Internet, on retrouve plusieurs applications issues de la recherche en fouille de textes, par exemple pour la gestion des courriels indésirables ou encore pour la détection de la cybercriminalité (Kontostathis A. 2009). Cette année, le DEFT 2011 propose deux thématiques. D'une part l'étude de la variation diachronique en corpus de presse, et d'autre part, l'appariement d'articles scientifiques avec leur résumé.

L'heure est à la numérisation du patrimoine culturel comme en témoigne l'appel du Ministère de la Culture de la Communication (France) pour rendre ce patrimoine largement accessible. Face à un projet de numérisation de grande ampleur, on peut craindre que certaines informations nécessaires au repérage du document risquent d'être absentes de l'original. À ce titre, les tâches proposées cette année par le comité d'organisation du DEFT 2011 peuvent rapidement trouver une application. En effet, l'année de publication d'un document est une donnée primordiale pour le repérage. La datation d'un document peut se faire manuellement par un expert, mais toutes les structures ne disposent pas des ressources et du temps nécessaires à une telle expertise. Les études menées par les différentes équipes du DEFT 2011 permettront certainement de développer des modèles et des classifieurs afin d'automatiser cette tâche. Toutefois ce n'est pas une tâche aisée car la numérisation à l'aide de la reconnaissance optique de caractères n'est pas exempte d'erreurs, rendant la reconnaissance de motifs difficile.

Le résumé de texte est une autre application de la fouille de texte. Cette année le défi ne consiste pas à résumer automatiquement un texte, mais à l'associer à l'article pour lequel il a été écrit. La réalisation de cette tâche offre plusieurs pistes d'application telles que la recherche d'articles à partir de résumés, ou l'analyse de résumé pour en isoler les principaux éléments afin d'améliorer la génération automatique de résumés. Mais cette tâche peut aussi permettre d'accélérer le processus de recherche d'information, grâce à une indexation automatique de l'article à partir de son résumé et non de l'article ou d'un livre. Sur des volumes importants, le gain de temps pourrait ne pas être négligeable.

2 Objectif général

À travers les deux tâches proposées pour le DEFT 2011, nous souhaitions observer si l'utilisation d'une stratégie asémantique donne des résultats proches ou meilleurs que ceux issus d'une stratégie sémantique. L'approche sémantique consistera à ne garder que les mots les plus significatifs pour chaque texte alors que l'approche asémantique se basera sur l'utilisation des mots non porteurs de sens, aussi appelés mots vides, tels que les particules, les mots de liaisons, les verbes auxiliaires etc. Une autre approche asémantique consistera à ne garder que les longueurs des phrases.

Un des avantages de l'approche asémantique est que l'on peut facilement faire abstraction de la langue (en utilisant la longueur des phrases par exemple), et qu'elle est moins coûteuse en terme de calcul (en utilisant uniquement les mots non porteurs de sens).

3 Tâche 1 : Variation diachronique

3.1 Objectif

Le but de cette tâche est de pouvoir attribuer la bonne année de publication à un extrait d'article de presse. Cette tâche est complexe car s'il est possible d'identifier des mots propres à une époque, notamment grâce aux réformes de l'orthographe (Albert *et al.* 2010), il est plus difficile d'associer un mot à une année précise ou du moins d'identifier un mot caractéristique d'une année précise.

3.2 Méthodologie

La démarche utilisée pour le traitement des documents repose sur un modèle vectoriel (Salton 1988, Memmi 2000, Forest 2009). L'analyse manuelle du corpus a fait ressortir beaucoup de problèmes liés au processus de numérisation avec reconnaissance optique de caractères. Certaines terminaisons de mots se voyaient tronquées, ou encore de nombreuses lettres « a » remplacées par des « o ». Malgré ces erreurs, nous avons fait le choix de ne pas intervenir sur le corpus original, de peur d'ajouter des erreurs en appliquant par exemple une orthographe actuelle à un mot orthographié différemment à son époque.

Les corpus de documents ayant été fournis par l'équipe organisatrice du DEFT 2011, nous avons pu commencer immédiatement par l'extraction et le filtrage du lexique. Trois types d'extraction ont été mis en place suivant la stratégie utilisée, sémantique ou asémantique. Pour la stratégie consistant à ne garder que les mots significatifs, un premier filtrage est appliqué, à partir d'une liste, pour enlever les mots non porteurs de sens. Dans le cas des stratégies asémantiques, un filtrage est effectué pour ne garder que les mots non porteurs de sens. Aucun filtrage n'est effectué pour la stratégie portant sur l'extraction des longueurs des phrases. Toujours dans un souci de garder les spécificité des mots par rapport à leur époque, aucun algorithme de lemmatisation n'a été appliqué.

Une fois le lexique extrait, il reste à représenter chaque document sous forme vectorielle. Ainsi chaque document sera représenté par la présence d'un terme pondérée par le *tf*idf*.

$$tf * idf = \mathrm{tf} \cdot \log_2\left(\frac{n}{df}\right)$$

où n est le nombre de documents

tf est la fréquence du terme (*term frequency*) dans le document

df est la fréquence de document ou apparaît le terme (*document frequency*)

Afin de pouvoir appliquer l'attribution automatique de l'année, nous avons recouru à un classifieur utilisant l'algorithme des bayesiens naïfs (Manning et Schütze, 1999), classifieur entraîné sur le corpus d'apprentissage à l'aide d'une validation de type *leave one out*.

3.3 Corpus

Le corpus est composé de textes numérisés avec reconnaissance optique de caractères (OCR). Ceci implique de nombreuses erreurs. On peut ainsi lire dans l'un des documents : « ne songe à tirer aucun parti des folles qui se icltCBt à>3 tète il le* trompe sans awtre ». Cette exemple montre la difficulté d'identifier certains mots originaux, même placés dans leur contexte. C'est pourquoi aucun traitement ne sera fait pour corriger le corpus des erreurs liées à l'OCRisation pour cette tâche.

Le corpus d'apprentissage est constitué de 3596 extraits de six journaux français (300 ou 500 mots) parus entre 1801 et 1944. Le corpus a été traité différemment selon les trois orientations choisies.

Les mots non porteurs de sens ont été supprimés pour la stratégie consistant à ne garder que les mots qui ont du sens. Seuls les mots non porteurs de sens ont été gardés pour la stratégie asémantique basée sur les mots non signifiants. En ce qui concerne la stratégie asémantique portant sur la longueur des phrases, chaque phrase a été remplacée par sa longueur[1] en nombre de tokens. Ainsi, seule la longueur des phrases à été retenue.

[1] La phrase « trahir cette mission, ni de la déserter » est remplacée par « Lg7 »

3.4 Résultats : phase d'apprentissage

3.4.1 Phase d'apprentissage

L'ensemble des tests a été effectué avec les paramètres suivants :

- Pondération des attributs : Chi2 maximum
- Méthode de validation : Retrait d'un cas
- Algorithme : Bayésien naifs
- Méthode de représentation des traits : Pourcentage de mots clés

Stratégie	Nombre d'attributs	Score
Longueur des phrases	154	7,23%
Mots non porteurs de sens	505	9,75%
Mots porteurs de sens	103897	39,44%

Tableau 1 : Apprentissage pour le corpus contenant des extraits de 500 mots

Stratégie	Nombre d'attributs	Score
Longueur des phrases	137	7,09%
Mots non porteurs de sens	495	8,40%
Mots porteurs de sens	75116	32,78%

Tableau 2 : Apprentissage pour le corpus contenant des extraits de 300 mots

Le score présenté dans les deux tableaux est calculé suivant la formule suivante :

$$S(p)=\frac{1}{N}\sum_{i=1}^{N} e^{-\frac{\pi}{10^2}\left(d_p(a_i)-d_r(a_i)\right)^2}$$

où $d_p(a_i)$ est la date prédit pour le fragment de texte a_i

$d_r(a_i)$ est la date référence pour le fragment de texte a_i

N est le nombre total de fragments de texte

Les résultats des deux tableaux montrent que la stratégie sémantique, basée sur les mots porteurs de sens, est nettement meilleure. Ceci s'explique notamment par le nombre conséquent d'attributs, ce qui permet un apprentissage plus important et donc d'être plus précis dans l'attribution des catégories. Toutefois, il serait intéressant, dans de futurs travaux, de comparer l'approche sémantique avec une combinaison d'approches asémantiques.

3.4.2 Phase de test

Les tests ont été réalisés grâces aux classifieurs modélisés lors de la phase d'apprentissage.

Test	Score
1-Longueur des phrases	6,2%
2-Mots non porteurs de sens	7,3%
3-Mots porteurs de sens	6,1%

Tableau 3 : Résultats pour la variation diachronique sur les extraits de 500 mots

Bien que les résultats soient faibles, on peut constater que la stratégie qui était la plus prometteuse lors de la phase d'apprentissage s'est avérée être la plus faible lors de la phase de test. En revanche la stratégie asémantique portant sur les mots vides (non porteurs de sens) a donné les meilleurs résultats. Ceci pourrait s'expliquer par le fait que ces mots ont moins été touchés par les problèmes de l'OCRisation.

4 Tâche 2 : Appariement résumé / article

4.1 Objectif

Pour résoudre cette tâche il faut associer un résumé à l'article scientifique dont il a été extrait. Pour réaliser l'appariement deux piste étaient proposées. La piste 1 consistait à associer le résumé à un article complet alors que la piste 2 consistait à l'associer à un article privé de son introduction et de sa conclusion.

4.2 Méthodologie

Pour trouver l'article associé à un résumé, nous nous sommes basés sur les travaux de Salton (présentés dans Ibekwe-SanJuan 2007) sur la recherche automatique d'information (Salton, 1988), notamment la formule de similarité entre une requête (Q) et un document (D).

$$similarité\,(Q,D) = \frac{\sum_{k=1}^{t} w_{qk} \cdot w_{dk}}{\sqrt{\sum_{k=1}^{t} (w_{qk})^2} \cdot \sqrt{\sum_{k=1}^{t} (w_{dk})^2}}$$

où w_{qk}, w_{dk} sont respectivement les poids des mots dans la requête et dans le document

La première étape consiste à représenter chaque résumé et chaque document par un vecteur de mots. Par la suite, une pondération *tf*idf* (cf. 3.2) est appliquée sur chaque composante du vecteur, étape suivie d'une normalisation pour avoir des poids compris entre 0 et 1. Ceci afin de minimiser l'importance des termes trop

fréquents dans le corpus qui deviennent alors peu discriminants. Plus le résultat du calcul de similarité, suivant la formule ci-dessus, est proche de 1, plus les vecteurs représentant le résumé et l'article sont proches.

Ainsi, pour chaque résumé, on calcule la similarité entre celui-ci et chacun des articles, ce qui correspond au calcul du cosinus entre les deux vecteurs. L'article qui lui sera associé sera celui qui aura permis d'obtenir le score le plus élevé lors du calcul de similarité.

4.3 Corpus

L'ensemble du corpus d'apprentissage est constitué de 198 articles et résumés. La tâche se subdivise en deux pistes. L'une, contient les articles dans leur intégralité. Une deuxième piste contient les articles desquels ont été supprimées introduction et conclusion.

Un filtrage du lexique a ensuite été effectué pour les différentes stratégies : ne garder que les mots vides , supprimer les mots vides, ou extraire la longueur des phrases.

4.4 Résultats

4.4.1 Phase d'apprentissage

Les paramètres suivants s'appliquent pour chaque test :

- Variation du nombre de traits discriminants : entre 100 et 1000
- Sélection des termes discriminants : fréquence pondérée par IDF

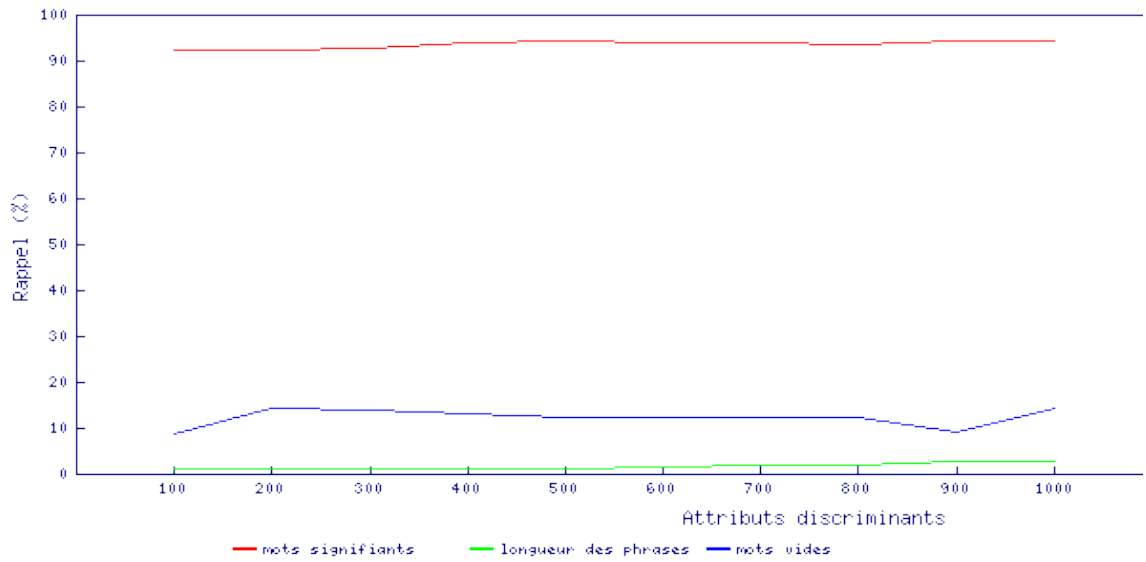

Figure 1: Performances de l'appariement, résumé/article sans introduction ni conclusion, suivant les stratégies adoptées

Les résultats montrent clairement que la stratégie utilisant les mots porteurs de sens est de loin la meilleure. Toutefois, cette dernière n'est pas parfaite, son score maximal est de 94% pour les articles privés de leur

introduction et conclusion et de 96.7% pour les articles complets. Toutefois il est intéressant de noter qu'elle varie seulement entre 92% et 94% pour les articles privés de l'introduction et de la conclusion et entre 95% et 96,7% pour les articles complets.

À l'issu de la phase d'apprentissage, on peut conclure que les approches asémantiques ne sont pas efficaces quand on les compare à l'approche sémantique. Donc, seule la stratégie basée sur les mots porteurs de sens sera retenue pour la phase de test.

4.4.2 Phase de test

Lors de la phase de test, seule l'approche sémantique a été retenue. Pour chaque test, les attributs sélectionnés sont pondérés par le *tf*idf*.

Test	Nombre d'attributs	Score global
1	600	98%
2	800	98,5%
3	1000	98%

Tableau 4 : Résultats pour l'appariement résumé / article complet

Test	Nombre d'attributs	Score global
1	800	95,4%
2	1000	95,5%

Tableau 5 : Résultats pour l'appariement résumé / article sans introduction ni conclusion

Le score global est la moyenne du nombre de résumés correctement appariés sur le nombre total de résumés. Ainsi le score global est obtenu suivant la formule :

$$S(p) = \frac{1}{N} \sum_{i=1}^{n} s\left(a_p(r_i), a_r(r_i)\right)$$

tel que $$s\left(a_p(r_i), a_r(r_i)\right) = \begin{cases} 1 & \text{si } a_p(r_i) = a_r(r_i) \\ 0 & \text{sinon} \end{cases}$$

où $a_p(r_i)$ est l'article prédit pour le résumé r_i

$a_r(r_i)$ est l'article de référence pour le résumé r_i

Les résultats présentés par les deux derniers tableaux montrent l'importance de l'introduction et de la conclusion qui permettent d'améliorer globalement l'appariement entre un résumé et son article. De plus nous pouvons noter qu'augmenter le nombre d'attributs à partir de 600 n'a pas vraiment d'influence sur les résultats.

5 Discussion

Les résultats obtenus pour la tâche d'appariement résumé / article sont relativement bons : 98,5% des résumés ont été correctement associés à l'article complet dont ils ont été extraits, et 95,4% pour l'appariement avec des articles privés d'introduction et de conclusion. L'avantage de la méthode utilisée, le calcul de similarité, est qu'elle n'est pas soumise au risque de surapprentissage. Par contre cette méthode est sensible à la longueur du résumé. Il est difficile de faire l'association d'un résumé très court : sur une ou deux phrase, il y a très peu de termes discriminants pouvant servir à la représentation vectorielle du résumé. Ainsi l'appariement devient incertain. Afin d'atteindre un appariement parfait, une solution pourrait être de mettre en place une lemmatisation des termes retenus afin de regrouper les mots ayant la même racine ou encore de ramener tous les verbes conjugués à leur infinitif. Une autre piste, en s'inspirant de des travaux de Lin, C. (Lin, 1995) serait de travailler avec des concepts plutôt que des mots. En effet, lorsque l'on écrit un texte nous avons tendance à vouloir éviter la répétition pour rendre la lecture plus fluide et agréable. Ainsi des mots comme « table », « armoire », « buffet » seraient regroupés sous le même concept de « meuble ». Toutefois, il y a toujours le risque de perdre la spécificité de certains termes discriminants. De plus, il pourrait être intéressant de donner un poids plus important aux termes extraits des introduction et conclusion afin de vérifier si l'appariement est meilleur.

Il est important de noter que les résultats obtenus pendant la période d'apprentissage n'ont pas été concluants pour les stratégies asémantiques. Il paraît évident que ce qui est associé au style d'écriture (longueur des phrases, utilisation des mots vides), ne permet pas d'identifier précisément le résumé d'un article. Ces résultats sont encourageants pour la génération automatique de résumés. En effet, cela veut dire que le style employé pour rédiger le résumé ne rentre pas en compte dans l'appariement. Quand bien même le résumé serait rédigé par une tierce personne ou automatiquement, nous serions donc encore capable de déterminer quel résumé appartient à quel article. Le générateur automatique de résumé doit par conséquent porter toute son attention sur les mots ou concepts caractérisant l'article.

Les résultats obtenus pour l'étude de la variation diachronique sont particulièrement décevants. Ceci s'explique notamment par un surapprentissage du classifieur. Les résultats obtenus pendant la phase d'apprentissage était de 39,44%, pour le corpus dont les extraits étaient de 500 mots, alors que lors de la phase de test, le meilleur résultat obtenu fut de 7,3%. Du fait de problèmes de ressources matérielles, nous avons été contraint de choisir une validation de type *leave one out*, nécessitant moins de ressources que la validation croisée. Choisir ce type de validation a été une erreur étant donné que cette méthode est caractérisée par un biais faible accompagné d'une grande variance (Cios, 2007). En d'autres termes, cette méthode garantit de meilleurs résultats que celle de la validation croisée, mais sa forte variance risque, lorsque les résultats seront faibles, d'être très éloignée de la meilleure solution.

Ces résultats sont aussi dus aux erreurs générées par le processus d'OCRisation. De nombreuses erreurs sur les mots ont mis à mal la stratégie asémantique, et la méthode asémantique basée sur la longueur des phrases n'a pu être optimale car de nombreux caractères de ponctuation apparaissaient aléatoirement au beau milieu d'une phrase. De plus la longueur des textes (300 ou 500 mots) était probablement insuffisante pour de telles stratégies. Toutefois il est intéressant de noter que la stratégie basée sur les mots non porteurs de sens à donné des résultats légèrement meilleurs. Cette stratégie était bien moins coûteuse en temps que celle basée sur le sens des mots. Si une stratégie asémantique devrait être testée de nouveau, il faudrait s'assurer d'avoir des textes plus longs et de corriger les erreurs du corpus. Par ailleurs, il faudrait définir de manière plus exhaustive ce qui caractérise le style d'écriture et combiner plusieurs facteurs.

Remerciements

Je tiens à remercier particulièrement le comité d'organisation du DEFT 2011. Je tiens aussi à remercier tous les participants des défis antérieurs. Grâce à la qualité des travaux effectués les années précédentes, j'ai pu me familiariser, en français, avec la fouille de textes sur des thématiques bien spécifiques. Mes remerciements vont aussi à Dominic Forest qui m'a fait découvrir le domaine de la fouille de texte et qui m'a encouragé à participer au DEFT2011.

Références

ALBERT P., BADIN F., DELORME M., DEVOS N., PAPAZOGLOU S., SIMARD J. (2010). Décennie d'un article de journal par analyse statistique et lexicale. Acte de *TALN2010*.

CIOS K. J. (2007). D*ata Mining : a Knowledge Discover approach*. New York: Springer.

BATHIA V. (1993). *Analysing Genre. Language Use in Professionnal Setting*. Boston: Addison Wesley.

FOREST D., HOEYDONCK VAN A., LÉTOURNEAU D., BÉLANGER M. (2009). Impacts sur la variation du nombre de trait discriminants sur la catégorisation des documents. Acte de *TALN2009*.

IBEKWE-SANJUAN F. (2007). *Fouilles de textes : méthodes, outils et applications*. Paris: Lavoisier.

KONTOSTATHIS A., EDWARDS L., LEATHERMAN A. (2009). Text mining and Cybercrime In *Text Mining : Application and Theory*. Chichester, U.K. : Wiley.

LIN C. (1995). Knowledge-Based Automatic Topic Identification. Actes de 33RD *Annual Meeting of the Association for Computational Linguistics*, 308-3010.

MANNING, C. D., SCHÜTZE, H.(1999). Foundations of statistical natural language processing. Cambridge (Mass.) : MIT Press.

MEMMI D. (2000). *Le modèle vectoriel pour le traitement de documents*. Grenoble: Cahiers Leibniz, no 2000-14.

ROWLEY J. (1982). *Abstracting and Indexing*. London: Clive Bingley.

SALTON G., BUCKLEY C. (1988). Term-weighting approches. *Automatique texte retrieval in information processing and management* 24(5), 456-465.