LIPN@DEFT2016 : Annotation de documents en utilisant l'Information Mutuelle

Davide Buscaldi Haïfa Zargayouna
Laboratoire d'Informatique de Paris Nord (LIPN, UMR 7030)
Université Paris 13 – Sorbonne Paris Cité & CNRS
{prenom.nom}@lipn.univ-paris13.fr

RÉSUMÉ

Cet article décrit le système proposé par le LIPN pour participer à l'édition 2016 pour le Défi Fouille de Textes. La tâche consiste à proposer des mots clés pour indexer des notices bibliographiques. Quatre domaines de spécialités ont été proposés : linguistique, sciences de l'information, archéologie et chimie. Nous avons proposé trois approches : une approche qui s'appuie sur le volet terminologique des thesaurus, une approche fondée sur l'information mutuelle et une approche qui fusionne les deux. Les mêmes approches ont été appliquées aux quatre domaines de spécialité.

ABSTRACT

LIPN@DEFT2016 : Document annotation based on pointwise Mutual Information

This paper describes the participation of the LIPN team at DEFT 2016. The task consisted on generating keywords in order to index bibliographic records. We proposed three approaches to solve the annotation problem in all the four domains offered (linguistics, information science, archeology and chemistry). The first approach is based on the terminology that is contained in domain ontologies or thesauri, the second one is based on pointwise mutual information, while the third on is a combination of the two approaches.

MOTS-CLÉS : Annotation Sémantique, Information Mutuelle, Fouille de Texte.

KEYWORDS: Semantic Annotation, Mutual Information, Text Mining.

1 Introduction

L'édition 2016 du défi a consisté à proposer des mots clés pour indexer une notice bibliographique. Ces mots clés peuvent être contrôlés ou libres. Des thesaurus sont fournis par domaine. Ils permettent de sélectionner les descripteurs censés caractériser au mieux le contenu des notices. Quatre domaines de spécialité sont proposés : linguistique, sciences de l'information, archéologie et chimie. Quand on dispose de thesaurus, la tâche consiste à proposer une indexation ou une annotation sémantique.

Dans nôtre participation à DEFT, nous avons identifié deux cas dans lesquels un processus d'annotation sémantique peut être mis en place :

1. l'étiquette du concept est dans le texte ;
2. l'étiquette n'est pas dans le texte mais il y a des indices repérables dans le texte.

Dans le premier cas, nous avons choisi de chercher directement dans le texte les étiquettes du thesaurus avec un moteur de recherche. Dans le deuxième cas, nous avons choisi d'exploiter le corpus

d'entraînement pour extraire les termes qui s'associent le plus souvent à certains descripteurs ou concepts. Finalement cela correspond à calculer l'information mutuelle entre les concepts et les termes des textes annotés dans le corpus d'entraînement. Nous ne nous sommes pas intéressés aux cas où les descripteurs ont été librement proposés par les documentalistes.

Dans la littérature, plusieurs approches d'annotation automatique existent. Parmi les plus efficaces, on peut citer une méthode de classification fondée sur les K plus proches voisins appliquée au domaine médical par exemple (Trieschnigg *et al.*, 2009), et dans le même domaine, la proposition de classificateurs bayesiens et machines à vecteurs de support (SVM) par (Jimeno Yepes *et al.*, 2012). L'information mutuelle a aussi été utilisée efficacement pour sélectionner des caractéristiques dans le contexte de l'annotation ou catégorisation de texte, individuellement (Lewis, 1992) ou en conjonction avec des algorithmes d'apprentissage supervisés (Dumais, 1998; Wang & Lochovsky, 2004). Dans notre proposition, l'information mutuelle est utilisée directement, sans passer par l'intermédiaire d'un algorithme d'apprentissage. Il est en effet, difficile voire impossible d'envisager de proposer un modèle par concept.

L'annotation fondée sur le volet terminologique d'une ontologie nécessite de mettre en place des techniques de reconnaissance de termes dans les textes comme par exemple dans (Jacquemin, 1997). L'approche que nous proposons repose sur un moteur de recherche. Elle ne nécessite aucune analyse fine des textes et permet de faire une reconnaissance approximée des termes. Nous décrivons dans la section suivante les algorithmes d'annotation que nous avons testés. Dans la section 3, nous présentons les résultats obtenus sur les différents corpus.

2 Algorithmes d'annotation

Nous présentons brièvement les deux techniques d'annotations proposées. La première technique s'appuie sur les étiquettes des concepts, la deuxième technique s'appuie sur la co-occurence entre les termes et les concepts.

2.1 Annotation fondée sur la terminologie

Cet algorithme d'annotation est utilisé par le moteur de recherche d'information sémantique YaSemIR (*Yet another Semantic Information Retrieval system*) (Buscaldi & Zargayouna, 2013) [1]. D'abord, la ressource sémantique (ontologie ou thesaurus) est analysée et indexée. L'index crée, relie chaque identifiant de concept à ses représentations lexicales ou étiquettes (principales ou autres) qui constituent le volet terminologique de la ressource. Les fragments textuels qui constituent les étiquettes peuvent ainsi être cherchés par le moteur de recherche.

En phase d'annotation, le texte du document à annoter d est analysé pour extraire les fragments à soumettre au moteur de recherche en tant que requêtes. Annoter un document revient à interroger la ressource indexée. Dans YaSemIR, seuls les syntagmes nominaux sont pris en compte. L'analyse syntaxique de tous les documents s'avérant très lente, pour DEFT2016 nous avons choisi d'extraire les trigrammes. Le moteur de recherche renvoie une liste ordonnée de résultats contenant les IDs des concepts avec leurs poids. Le meilleur résultat est ajouté à l'ensemble d'annotations du documents. La Figure 1 montre un exemple d'annotation en utilisant l'ontologie du domaine archéologique.

1. https://github.com/dbuscaldi/YaSemIR

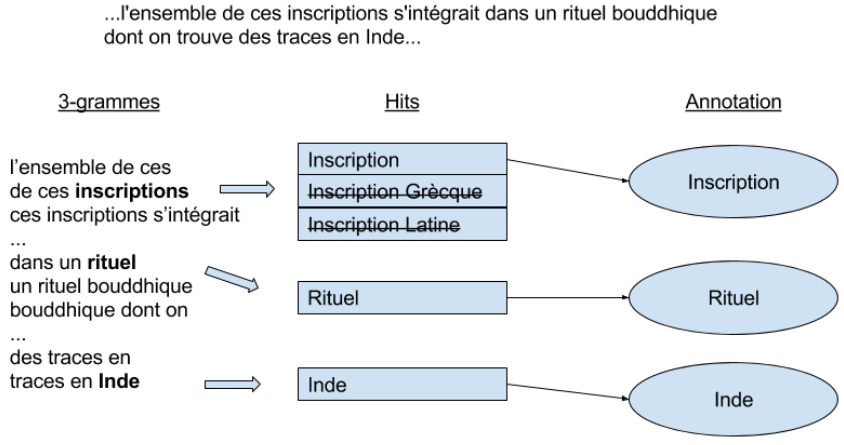

FIGURE 1 – Exemple de chaîne de traitement de l'annotateur YaSemIR

Cet algorithme d'annotation est utile quand les étiquettes sont présentes dans le texte même partiellement. Le bruit peut provenir des étiquettes partageant la même racine ainsi que des concepts retournés en premier mais avec un faible score.

2.2 Annotation fondée sur l'Information Mutuelle

Quand les concepts n'ont pas de représentations lexicales dans le texte, les algorithmes d'annotation fondés sur la terminologie échouent. Nous proposons un algorithme qui se fonde sur la co-occurrence entre le concept et les termes du documents. L'idée est que si, dans un corpus de textes annotés, un ou plusieurs termes co-occurrent souvent avec le même concept c, alors si on retrouve le même terme ou la même combinaison de termes dans un document sans annotation, on peut l'annoter avec le concept c.

Nous avons d'abord procédé à quelques filtrages. Nous avons éliminé tous les cas pour lesquels les co-occurrences sont dues au pur hasard. Pour cela, dans la phase d'entraînement nous avons pris en compte seulement les concepts qui ont servi pour annoter au moins 5 documents dans le corpus de référence. Nous n'avons pris en compte que les substantifs et adjectifs qui représentent habituellement les étiquettes des concepts.Au final, nous avons calculé l'information mutuelle uniquement si le terme et le concept co-occurrent dans au moins 3 documents.

L'information mutuelle ponctuelle est définie comme :

$$IP(t,c) = \log \frac{p(t,c)}{p(t)p(c)}$$

Où t est un terme (substantif ou adjectif dans notre cas) et c un concept ; $p(t,c)$ est la probabilité conjointe entre t et c, calculée comme $freq(t,c)/N$, $p(t) = freq(t)/N$ et $p(c) = freq(c)/N$, où N est le nombre de documents dans le corpus d'entraînement.

2.2.1 Construction du modèle

Nous construisons une matrice M avec $|C|$ lignes et $|T|$ colonnes, où C est l'ensemble des étiquettes dans le corpus d'entraînement avec fréquence > 5, et T est l'ensemble des mots du dictionnaire (donc tous les substantifs et adjectifs observés dans le corpus) avec fréquence > 3. Chaque élément $M_{i,j}$ est calculé comme :

$$M_{i,j} = \begin{cases} IP(t_i, c_j) & \text{if } IP(t_i, c_j) > 0 \\ 0 & \text{else} \end{cases}$$

Sur cette matrice, nous appliquons l'analyse sémantique latente (LSA - *Latent Semantic Analysis*), en décomposant M en valeurs singulières (algorithme SVD) $M = U\Sigma V^T$, et en approximant M avec $\hat{M} = U_k \Sigma_k V_k^T$, avec les meilleurs k valeurs singulières sélectionnés ($k = 100$).

Nous espérons améliorer la couverture en appliquant LSA sur la matrice. En effet, LSA permet de trouver des termes fortement associés avec d'autres termes qui sont très caractéristiques pour certaines étiquettes, même si dans le corpus d'entraînement l'étiquette n'apparaît pas avec ce mot. Par exemple, si « Paris » et « français » co-occurrent souvent, mais dans la collection on ne trouve que « France » associé au terme « français », on peut déduire que « France » est une annotation plausible si dans le texte on trouve « Paris », même si on ne les a jamais vus ensemble dans le corpus l'entraînement.

Finalement, nous appliquons un filtre sur la matrice $\hat{M}$ de la façon suivante :

$$\hat{M}^{\sigma}_{i,j} = \begin{cases} \hat{M}_{i,j} & \text{if } \hat{M}_{i,j} \geq \left(0.5 * \sigma'(\hat{M}_{*,j}) + \mu'(\hat{M}_{*,j})\right) \\ 0 & \text{else} \end{cases}$$

Où $\hat{M}_{*,j}$ est le $j-$ème vecteur colonne de la matrice $\hat{M}$, $\sigma'(\mathbf{x})$ est l'écart type des éléments non nuls du vecteur $\mathbf{x}$ et $\mu'(\mathbf{x})$ est la moyenne des éléments non nuls du vecteur $\mathbf{x}$. La matrice $\hat{M}^{\sigma}$ ainsi obtenue est utilisée comme *modèle* pour l'annotation des documents.

Nous avons éliminé du modèle les termes non discriminants afin d'améliorer la précision.

2.2.2 Annotation

Nous associons au document d à annoter un vecteur binaire $\mathbf{b}$ de taille $|T|$ (taille du vocabulaire) où chaque élément $\mathbf{b}_i$ est à 1 si le terme correspondant est dans le texte du document d, 0 autrement. Nous calculons pour chaque étiquette l_i le score $s(l_i) = \mathbf{b} \cdot \hat{M}^{\sigma}_{i,*}$. Le document est annoté finalement avec les 20 étiquettes avec les meilleurs scores > 0. S'il y a moins de 20 étiquettes avec des scores positifs, alors une annotation est générée avec toutes les étiquettes ayant un score positif. Le choix de 20 a été défini empiriquement après les tests sur le corpus de développement.

L'exemple en figure 2 montre un cas d'annotation, avec une bonne étiquette (« Recherche scientifique ») et une mauvaise (« système de recherche d'information »). Nous pouvons observer que le modèle ne prend pas en compte l'absence de caractéristiques importantes dans les vecteurs (le terme « système »). Un algorithme d'apprentissage pourrait tirer profit de ces informations.

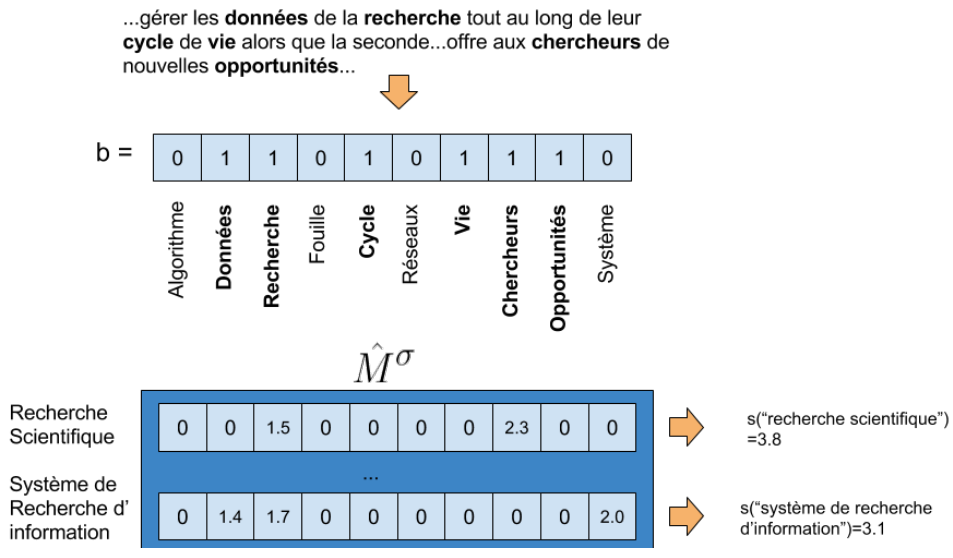

FIGURE 2 – Exemple d'annotation en utilisant l'IM sur un document du domaine *sciences de l'information*. Les termes en gras sont les termes du texte à annoter présents dans le thesaurus

3 Résultats

Les modèles utilisés pour la participation à DEFT2016 ont été les suivants : modèle M1 - annotateur YaSemIR ; modèle M2 - annotateur fondé sur l'information mutuelle ; modèle M3 - fusion des résultats M1 et M2. L'idée du troisième modèle est d'exploiter la complémentarité entre M1 et M2. Ci dessous les résultats obtenus sur le corpus d'entraînement et développement avec M1 et M2 respectivement.

Corpus	Prec	Rappel	F-Measure
archeo	19.61%	18.86%	19.23%
chimie	6.44%	9.61%	7.71%
linguistique	1.08%	2.26%	1.46%
sciences info	6.13%	14.16%	8.56%

TABLE 1 – Résultats sur le corpus d'entraînement avec YaSemIR, sur les différents corpus.

La comparaison des résultats (voir tableau 2) montre clairement l'avantage obtenu grâce à l'application de LSA sur la matrice M. Les résultats officiels (Tableau 3) montrent que la fusion a permis de profiter effectivement de la complémentarité des sorties pour améliorer le rappel, mais la faible précision des méthodes a empêché d'obtenir des améliorations par rapport à la F-mesure, sauf dans le cas du domaine de la chimie.

Corpus	LSA	Prec	Rappel	F-Measure
archeo	n	16.22%	20.50%	18.11%
	y	20.71%	26.67%	23.31%
chimie	n	11.62%	13.68%	12.56%
	y	12.53%	20.16%	15.45%
linguistique	n	2.95%	7.17%	4.18%
	y	5.75%	13.66%	8.09%
sciences info	n	12.60%	21.62%	15.92%
	y	13.07%	28.68%	17.95%

TABLE 2 – Résultats sur le corpus de développement avec IM - avec et sans LSA.

Corpus	Modèle	Prec	Rappel	F-Measure
archeo	M1	**33.93%**	31.25%	**30.75%**
	M2	19.67%	24.01%	20.95%
	M3	22.63%	**46.54%**	29.61%
chimie	M1	**13.87%**	16.00%	13.29%
	M2	11.13%	17.76%	13.08%
	M3	10.88%	**30.25%**	**15.31%**
linguistique	M1	10.04%	10.45%	12.97%
	M2	**13.98%**	30.81%	**19.07%**
	M3	10.59%	**45.23%**	16.98%
sciences info	M1	8.86%	18.65%	11.48%
	M2	**11.72%**	23.54%	**15.34%**
	M3	9.03%	**36.61%**	14.25%

TABLE 3 – Résultats officiels sur le corpus d'évaluation avec les modèles M1, M2 et la fusion des sorties (M3).

4 Conclusions

Les deux difficultés majeures de la tâche proposée à DEFT2016 était le grand nombre d'étiquettes et le fait que les étiquettes pouvaient provenir ou pas d'un thésaurus. Nous avons choisi d'utiliser un annotateur fondé sur la terminologie pour tirer profit au mieux des thesauri. Vu la difficulté d'entraîner un modèle pour des milliers d'étiquettes, nous avons choisi d'utiliser un algorithme d'annotation fondé sur l'information mutuelle. Ce dernier a été mis au point sur le corpus de développement, mais les différents paramètres qui ont été choisis (fréquence minimale des étiquettes, poids des termes dans les vecteurs, nombre d'étiquettes à retourner, en particulier) n'ont pas été testés de façon approfondie. Nous estimons donc qu'il reste une marge d'amélioration pour cette méthode, cela pourra être vérifié par des expériences plus étendues. Il est aussi à noter que pour bien estimer l'information mutuelle, nous avons besoin d'un corpus d'entraînement assez grand. Dans le corpus d'entraînement DEFT2016, nous estimons qu'il n'y a assez de données que pour 12.47% des étiquettes en moyenne ; par exemple pour les 2256 étiquettes trouvées dans le corpus apprentissage pour le domaine de l'archéologie, seulement 278 ont une fréquence supérieure à 5. Nous pourrons augmenter le seuil de fréquence pour améliorer la précision du modèle mais il faudrait compenser la perte de rappel. Finalement, cette expérience nous a permis d'observer l'utilité du LSA dans ce contexte.

Remerciements

Cette recherche s'insère dans le programme « Investissements d'Avenir » géré par l'Agence Nationale de la Recherche ANR-10-LABX-0083 (Labex EFL).

Références

BUSCALDI D. & ZARGAYOUNA H. (2013). YaSemIR : Yet Another Semantic Information Retrieval System. In *Proceedings of the Sixth International Workshop on Exploiting Semantic Annotations in Information Retrieval*, ESAIR '13, p. 13–16, New York, NY, USA : ACM.

DUMAIS S. (1998). Using SVMs for text categorization. *IEEE Intelligent Systems*, **13**(4), 21–23.

JACQUEMIN C. (1997). *Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus.* Habilitation à diriger des recherchesinformatique fondamentale, Université de Nantes.

JIMENO YEPES A., MORK J. G., WILKOWSKI B., DEMNER FUSHMAN D. & ARONSON A. R. (2012). MEDLINE MeSH Indexing : Lessons Learned from Machine Learning and Future Directions. In *Proceedings of the 2Nd ACM SIGHIT International Health Informatics Symposium*, IHI '12, p. 737–742, New York, NY, USA : ACM.

LEWIS D. D. (1992). Feature Selection and Feature Extraction for Text Categorization. In *Proceedings of the Workshop on Speech and Natural Language*, HLT '91, p. 212–217, Stroudsburg, PA, USA : Association for Computational Linguistics.

TRIESCHNIGG D., PEZIK P., LEE V., DE JONG F., KRAAIJ W. & REBHOLZ-SCHUHMANN D. (2009). MeSH Up : effective MeSH text classification for improved document retrieval. *Bioinformatics*, **25**(11), 1412–1418.

WANG G. & LOCHOVSKY F. H. (2004). Feature Selection with Conditional Mutual Information Maximin in Text Categorization. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, CIKM '04, p. 342–349, New York, NY, USA : ACM.