TopicRank en domaines de spécialité : participation du LINA à DEFT 2016

Adrien Bougouin Florian Boudin Béatrice Daille
LINA – UMR CNRS 6241, 2 rue de la Houssinière, 44322 Nantes Cedex 3, France
<prenom.nom>@univ-nantes.fr

RÉSUMÉ

Cet article présente la participation de l'équipe TALN du LINA au défi fouille de textes (DEFT) 2016. Développé pour l'indexation de mots-clés, notre système reprend une méthode à base de graphe état de l'art dans le domaine (TopicRank) et l'étend afin de mieux répondre aux attentes d'une indexation professionnelle réalisée dans le cadre d'une bibliothèque scientifique numérique. Notre système s'est classé à la troisième place sur un total de cinq participants.

ABSTRACT

LINA at DEFT 2016

This article presents the participation of the TALN group at LINA to the défi fouille de textes (DEFT) 2016. Developed specifically for automatic keyphrase annotation, our system improves an existing method (TopicRank), mimicking professional indexers of Digital Libraries. Our system ranked third out of a total of five systems.

MOTS-CLÉS : DEFT 2016, extraction de mots-clés, assignement de mots-clés, méthode à base de graphe, domaine de spécialité.

KEYWORDS: DEFT 2016, keyphrase extraction, keyphrase assignment, graph-based method, specific domain.

1 Introduction

L'indexation automatique consiste à identifier un ensemble de mots-clés (e.g. mots, termes) qui décrit le contenu d'un document. Les mots-clés peuvent ensuite être utilisés, entre autres, pour faciliter la recherche d'information ou la navigation dans les collections de documents. L'édition 2016 du défi fouille de textes (DEFT) porte sur l'extraction automatique de mots-clés à partir d'articles scientifiques en français.

L'objectif du défi DEFT 2016 est de retrouver, à partir du contenu d'articles scientifiques, les mots-clés qui ont été attribués par des indexeurs professionnels. Les documents sont issus de quatre domaines de spécialité : la linguistique, les sciences de l'information, l'archéologie et la chimie. Cet article décrit le système que nous avons mis au point pour le défi.

Le reste de cet article est organisé comme suit. Nous commençons par présenter les données (cf Section 2), puis notre approche (cf Section 3) et enfin, nos résultats comparés à ceux des autres participants de DEFT 2016 (cf Section 4).

2 Données du défi DEFT 2016

Les données mises à disposition par les organisateurs du défi DEFT 2016 sont composées de quatre corpus traitant chacun d'un domaine de spécialité parmi la linguistique, les sciences de l'information, l'archéologie et la chimie. Chaque corpus consiste en un ensemble de notices bibliographiques (titre, résumé) aux formats TEI et TXT (texte pré-traité de la notice), et un thésaurus au format SKOS. Les corpus sont divisés en trois sous-ensembles : jeu d'apprentissage, de développement et de test. Pour les deux premiers jeux, nous disposons des mots-clés de référence assignés par des indexeurs professionnel de l'Inist.

3 Approche

Nous proposons une approche fondée sur celle de TopicRank (Bougouin & Boudin, 2014). TopicRank est une méthode à base de graphe pour l'extraction de mots-clés. Elle sélectionne d'abord des mots-clés candidats au sein du document à analyser, les groupes en sujets, projette les sujets dans un graphe et les ordonnent à la manière de TextRank (Mihalcea & Tarau, 2004). Contrairement à d'autres méthodes d'extraction de mots-clés, TopicRank est capable de réduire considérablement la redondance des mots-clés extraits.

Dans ce travail, nous reprenons TopicRank et en améliorons ses performances en tirant partie des éléments du domaine des collections de DEFT 2016. Notre approche utilise les mots-clés des notices d'entrainement comme éléments du domaine et n'utilise pas les thésaurus fournis par les organisateurs. De cette manière, son usage ne se restreint pas uniquement aux données similaires à celles présentées dans le cadre de DEFT 2016.

3.1 TopicRank

TopicRank repose sur cinq grandes étapes :

1. **Sélection des mots-clés candidats.** Suivant les travaux précédents (Wan & Xiao, 2008; Hasan & Ng, 2010), TopicRank sélectionne les plus longues séquences de noms et d'adjectifs en tant que mots-clés candidats :

$$\text{mots\_cles\_candidat} = (NOM|ADJ)+ \tag{1}$$

2. **Groupement en sujets.** TopicRank groupe les mots-clés candidats similaires en sujets. Deux candidats c_i et c_j sont jugés similaires lorsqu'ils partagent au moins un quart de leurs mots, racinisés d'après la méthode de (Porter, 1980) :

$$\text{sim}(c_i, c_j) = \frac{|\text{Porter}(c_i) \cap \text{Porter}(c_j)|}{|\text{Porter}(c_i) \cup \text{Porter}(c_j)|} \tag{2}$$

$$\forall c_i, c_j \in CANDIDATS, c_j \in \text{sujet}(c_i) \Rightarrow \text{sim}(c_i, c_j) \geq \frac{1}{4} \tag{3}$$

Le groupement est réalisé avec un algorithme de groupement hiérarchique agglomératif.

3. **Construction du graphe.** TopicRank représente le document par un graphe complet $G = (N, A \subseteq N \times N)$ où les nœuds N sont les sujets. Chaque sujet $n \in N$ est connecté aux autres par une arête pondérée $a \in A$ selon la force du lien sémantique entre les sujets :

$$\text{poids}(n_i, n_j) = \sum_{c_i \in n_i} \sum_{c_j \in n_j} \text{distance}(c_i, c_j) \tag{4}$$

$$\text{distance}(c_i, c_j) = \sum_{p_i \in \text{positions}(c_i)} \sum_{p_j \in \text{positions}(c_j)} \frac{1}{|p_i - p_j|} \tag{5}$$

   Plus faible est la distance entre les mots-clés candidats de deux sujets dans le document, plus élevé est le poids de l'arête entre les deux sujets.

4. **Ordonnancement des sujets.** À la manière de TextRank (Mihalcea & Tarau, 2004), TopicRank ordonne les sujets par importance selon le principe de recommandation. Plus un sujet est fortement connecté à un grand nombre de sujets, plus il gagne d'importance, et plus les sujets avec lesquels il est fortement connecté sont importants, plus l'importance qu'il gagne est forte :

$$\text{importance}(n_i) = (1 - \lambda) + \lambda \times \sum_{n_j \in A(n_i)} \frac{\text{poids}(n_i, n_j) \times \text{importance}(n_j)}{\sum_{n_k \in A(n_j)} \text{poids}(n_j, n_k)} \tag{6}$$

   Où λ est un facteur de lissage fixé à 0,85 par Brin & Page (1998).

5. **Extraction des mots-clés.** TopicRank extrait un unique mot-clé pour chacun des k plus importants sujets. Bougouin *et al.* (2013) ont choisi de sélectionner dans chaque sujet le mot-clé candidat qui apparaît en premier dans le document.

Notre approche modifie les étapes de construction du graphe, d'ordonnancement par importance et de sélection des mots-clés de TopicRank. La construction du graphe étend le graphe de sujet en l'unifiant à un graphe des mots-clés de référence du domaine. L'ordonnancement est désormais conjoint entre les sujets du document et les mots-clés du domaine. Enfin, la sélection des mots-clés ajoute la possibilité de puiser dans le graphe du domaine. De cette manière, notre méthode est capable de réaliser simultanément deux catégories d'indexation par mots-clés :

- **Extraction de mots-clés** : les mots-clés sont sélectionnés parmi les unités textuelles du document (e.g. TopicRank) ;
- **Assignement de mots-clés** : les mots-clés ne sont pas restreints au contenu du document et doivent faire partie d'un vocabulaire contrôlé construit pour cette tâche.

3.2 Extraction et assignement (M1)

Afin de réaliser simultanément extraction et assignement de mots-clés, nous unifions deux graphes : l'un représentant le document (graphe de sujets) et l'autre les mots-clés de référence de son domaine (graphe du domaine). Le premier graphe sert à l'extraction de mots-clés. Le second, construit à partir des mots-clés de référence de documents d'apprentissage, sert à l'assignement. Ainsi, nous faisons l'hypothèse que les mots-clés de référence des documents d'apprentissage constituent la terminologie du domaine et nous les utilisons comme substituts au vocabulaire contrôlé usuel en assignement de

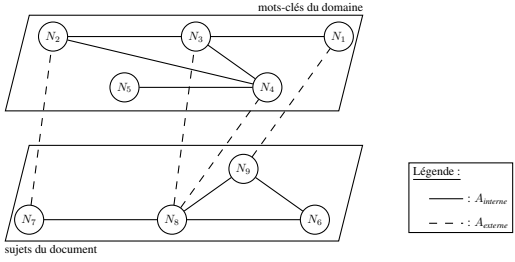

Figure 1: Illustration du graphe unifié que nous proposons

mots-clés. Contrairement aux mots-clés candidats sélectionnés dans le document, les mots-clés de référence ne sont pas redondants et ne sont donc pas groupés en sujets.

Soit le graphe unifié non orienté $G = (N, A = A_{interne} \cup A_{externe})$. N dénote indifféremment les sujets et les mots-clés du domaine. A regroupe les arêtes $A_{interne}$, qui connectent deux sujets ou deux mots-clés du domaine, et les arêtes $A_{externe}$, qui connectent un sujet à un mot-clé du domaine (voir la figure 1). Nous connectons deux sujets ou deux mots-clés du domaine lorsqu'ils apparaissent dans le même contexte et nous pondèrons leur arête par le nombre de fois que cela se produit. Lorsqu'il s'agit des sujets, le contexte est une phrase du document ; lorsqu'il s'agit des mots-clés du domaine, le contexte est l'ensemble des mots-clés du domaine d'un document d'apprentissage. Les contextes étant utilisés pour la création du graphe, le graphe de sujets n'est plus complet comme celui de TopicRank.

Le graphe de sujets et le graphe du domaine sont unifiés grâce aux arêtes $A_{externe}$. L'objectif des arêtes $A_{externe}$ est de connecter le document à son domaine par l'intermédiaire des concepts qu'ils partagent. Une arête $A_{externe}$ est donc créée entre un sujet et un mot-clé du domaine si ce dernier appartient au sujet, c'est-à-dire correspond à l'un de ses mots-clés candidats.

À partir du graphe unifié, nous ordonnons simultanément sujets $s \in N$ du document et mots-clés $m \in N$ du domaine par importance. Pour cela, nous reprenons le même principe que TopicRank et l'adaptons de sorte que sujets et mots-clés du domaine se transfèrent de l'importance. Nous proposons deux formes de recommandation : une recommandation interne $R_{interne}$ qui intervient entre deux nœuds du même type (sujets ou mots-clés du domaine) et une recommandation externe $R_{externe}$ qui intervient entre un sujet et un mot-clé du domaine.

$$\text{importance}(s_i) = (1 - \lambda_s)\ R_{externe}(s_i) + \lambda_s\ R_{interne}(s_i) \tag{7}$$

$$\text{importance}(m_i) = (1 - \lambda_m)\ R_{externe}(m_i) + \lambda_m\ R_{interne}(m_i) \tag{8}$$

$$R_{interne}(n_i) = \sum_{n_j \in A_{interne}(n_i)} \frac{\text{poids}(n_i, n_j)(n_j)}{\sum_{n_k \in A(n_j)} \text{poids}(n_j, n_k)} \tag{9}$$

$$R_{externe}(n_i) = \sum_{n_j \in A_{externe}(n_i)} \frac{\text{importance}(n_j)}{|A_{out}(n_j)|} \tag{10}$$

Où λ_s et λ_m sont deux facteurs de lissage définis empiriquement pour l'ordonnancement par importance des sujets et des mots-clés du domaine, respectivement.

Pour finir, indifféremment de leur type nous identifions dix mots-clés parmi les sujets et/ou mots-clés du domaine les plus importants.

3.3 Extraction seule (V1.1)

Nous proposons une variante de notre approche qui ne consiste qu'à identifier les mots-clés parmi les sujets. Le graphe unifié est toujours utilisé et les mots-clés du domaine sont tout de même utilisés pour l'ordonnancement.

3.4 Assignement seul (V1.2)

Nous proposons une seconde variante de notre approche qui ne consiste qu'à identifier les mots-clés parmi les mots-clés du domaine. Le graphe unifié est toujours utilisé et les sujets du document sont tout de même utilisés pour l'ordonnancement.

4 Résultats

Nous présentons dans cette section les résultats officiels de la campagne DEFT 2016. Nous avons soumis trois exécutions pour chaque collection : l'une pour notre approche (M1) et deux autres pour ses variantes (V1.1 et V1.2).

	Archéologie			**Chimie**			**Linguistique**			**Sciences de l'information**		
Méthode	Précision	Rappel	F-mesure	Précision	Rappel	F-mesure	Précision	Rappel	F-mesure	Précision	Rappel	F-mesure
M1	49,86	31,16	37,28	20,87	17,45	18,11	22,23	24,87	23,24	20,61	20,65	20,21
V1.1	43,63	26,63	32,17	15,77	13,10	13,60	13,77	15,56	14,47	15,67	15,87	15,39
V1.2	**53,77**	**33,46**	**40,11**	**21,15**	**17,54**	**18,28**	**23,16**	**25,85**	**24,19**	**21,93**	**21,83**	**21,45**

Table 1: Résultats de nos trois exécutions pour les collections d'archéologie, de chimie, de linguistique et de sciences de l'information

Le tableau 1 présente les résultats de nos trois exécutions pour les collections d'archéologie, de chimie, de linguistique et de sciences de l'information. Nous constatons que la variante V1.2 obtient les meilleurs résultats. Compte tenu de la nature des collections de données, cette observation semble normale. En effet, les notices sont indexées par des indexeurs professionnels utilisant principalement un vocabulaire controlé. Cela révèle toutefois que notre modèle ne fait pas suffisamment émerger les mots-clés du domaine pour M1.

Le tableau 2 présente, pour les collections d'archéologie, de chimie, de linguistique et de sciences de l'information, le classement des différentes équipes sur la base de la meilleure soumission. Notre soumission est classée au rang 3 sur 5.

5 Conclusion

Nous avons décrit la participation du LINA à DEFT 2016. Notre système est le résultat d'une analyse approfondie du mode de fonctionnement d'un indexeur professionnel de l'INIST. Il combine deux

	Archéologie		Chimie		Linguistique		Sciences de l'information	
Rang	Équipe	F-mesure	Équipe	F-mesure	Équipe	F-mesure	Équipe	F-mesure
1	EXENSA	45,59	EXENSA	21,46	EBSIUM	31,75	EBSIUM	28,98,
2	LIMSI	43,26	EBSIUM	21,07	EXENSA	26,30	EXENSA	23,86
3	**LINA**	**40,11**	**LINA**	**18,28**	**LINA**	**24,19**	**LINA**	**21,45**
4	EBSIUM	34,96	LIPN	15,31	LIPN	19,07	LIPN	15,34
5	LIPN	30,75	LIMSI	15,29	LIMSI	15,63	LIMSI	12,49

Table 2: Classement de DEFT 2016 sur la base des meilleures soumissions pour les collections d'archéologie, de chimie, de linguistique et de sciences de l'information. Notre classement est indiqué en gras.

graphes représentant le document à analyser et son domaine de spécialité, lui permettant d'extraire des mots-clés du document et d'en assigner à partir de son domaine. Ces derniers n'apparaissent pas nécessairement dans le document. Notre système s'est classé à la troisième place sur un total de cinq systèmes.

Parmi les trois variantes que nous avons proposées, celle qui ne réalise que l'assignement de mots-clés est la meilleure. Cela signifie que notre modèle hybride pour l'extraction et l'assignement simultanés de mots-clés n'est pas optimal. Nous envisageons d'étendre ce travail en affinant le schéma de connection des deux graphes afin de faciliter l'émergence des mots-clés à assigner.

Remerciements

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence (ANR-12-CORD-0029).

Références

BOUGOUIN A. & BOUDIN F. (2014). TopicRank : ordonnancement de sujets pour l'extraction automatique de termes-clés. *TAL*, **55**(1), 45–69.

BOUGOUIN A., BOUDIN F. & DAILLE B. (2013). Topicrank: Graph-Based Topic Ranking for Keyphrase Extraction. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP)*, p. 543–551, Nagoya, Japan: Asian Federation of Natural Language Processing.

BRIN S. & PAGE L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, **30**(1), 107–117.

HASAN K. S. & NG V. (2010). Conundrums in Unsupervised Keyphrase Extraction: Making Sense of the State-of-the-Art. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING)*, p. 365–373, Stroudsburg, PA, USA: Association for Computational Linguistics.

MIHALCEA R. & TARAU P. (2004). TextRank: Bringing Order Into Texts. In DEKANG LIN & DEKAI WU, Eds., *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 404–411, Barcelona, Spain: Association for Computational Linguistics.

PORTER M. F. (1980). An Algorithm for Suffix Stripping. *Program: Electronic Library and Information Systems*, **14**(3), 130–137.

WAN X. & XIAO J. (2008). Single Document Keyphrase Extraction Using Neighborhood Knowledge. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, p. 855–860: AAAI Press.