Analyse en rôles sémantiques pour le résumé automatique

Elyase Lassouli*, Yasmine Mesbahi*
Camille Pradel* Damien Sileo*,**

*Synapse Développement
5 rue du Moulin Bayard, 31 000 Toulouse
http://www.synapse-developpement.fr/
**IRIT, Université Toulouse 3
118 Route de Narbonne, 31 062 Toulouse
https://www.irit.fr

Résumé. Cet article présente une approche visant à extraire les informations exprimées dans un corpus de textes et en produire un résumé. Plusieurs variantes de méthodes extractives de résumé de texte ont été implémentées et évaluées. Leur principale originalité réside dans l'exploitation de structures appelées CDS (pour *Clause Description Structure*) issues d'un composant d'annotation en rôles sémantiques et non directement des phrases composant les textes. Le résumé obtenu est un sous-ensemble des CDS issus du corpus d'origine ; ce format permettra dans la suite la détection d'incohérences textuelles. Dans ce travail, nous retransformons les CDS résumés en texte pour permettre la comparaison de notre approche avec celles de la littérature. Les premiers résultats sont très encourageants : les variantes que nous proposons obtiennent généralement de meilleurs scores que des implémentations de méthodes de référence.

1 Introduction

Sur le Web, l'utilisateur est démuni face à de très grands volumes de documents de qualité assez inégale et d'une fiabilité parfois douteuse. Le résumé automatique peut permettre aux humains de mieux appréhender ces données surabondantes Notre objectif est d'utiliser le résumer automatique de corpus pour construire une base de connaissances fiable permettant l'identification d'incohérences dans un nouveau texte. Cette approche est illustrée dans la figure 1 : elle consiste à extraire d'une grande quantité de textes les faits redondants en se basant sur l'idée qu'ils représentent des connaissances consensuelles et donc aptes à permettre de façon fiable l'identification d'incohérences dans un nouveau texte. En plus d'un début de modèle de détection d'incohérence, ce travail peut être vu comme une contribution à la tâche de résumé automatique de corpus de textes ; dans cet article, nous le présentons et l'évaluons comme tel.

La principale originalité de l'approche réside dans l'exploitation de structures appelées CDS (pour *Clause Description Structure*) issues d'un composant propriétaire produisant des

Ce travail a été financé dans le cadre du projet DGA-RAPID *Détection d'Incohérences Textuelles* n162906126

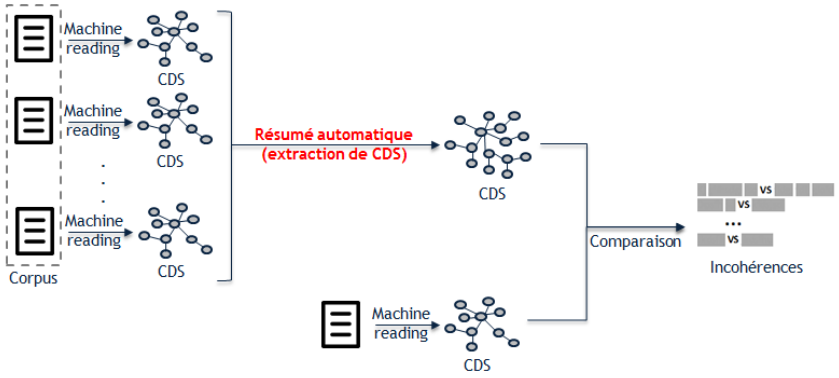

FIG. 1: Résumé automatique de corpus pour la détection d'incohérences dans un texte

résultats très proches sur la forme d'une analyse en rôles sémantiques (*Semantic Role Labeling*). Le format des CDS nous permet d'adapter directement des méthodes extractives de la littérature (il s'agit dans notre cas de sélectionner des CDS et non plus des phrases) et d'en proposer de nouvelles. Les CDS permettent également de représenter le contenu d'un texte avec un certain niveau d'abstraction (entités nommées et anaphores résolues, formes actives et passives normalisées, éléments du texte rattachés à une ressource lexicale généraliste), ce qui permet de concentrer les efforts sur le contenu et non sur la forme. Les premières expérimentations menées sur un jeu de données réduit montrent des résultats très encourageants : les variantes que nous proposons obtiennent généralement de meilleurs scores (rappel, précision et F-mesure calculés avec ROUGE) que des implémentations de méthodes de référence.

Nous donnons d'abord un aperçu des méthodes existantes en section 2. Puis nous décrivons le composant de génération des CDS en section 3 et les différentes variantes de notre approche en section 4. Enfin, nous présentons et discutons l'évaluation de l'approche en section 5 ;

2 Etat de l'art

Le résumé automatique de textes est une tâche populaire en Traitement Automatique des Langues et a fait l'objet de nombreuses recherches Saggion et Poibeau (2013); Lloret et Palomar (2012). Les systèmes existants exploitent principalement deux approches :

— L'approche extractive : consiste à extraire au sein d'un texte les phrases les plus "importantes" (ou les phrases clé). Les travaux de cette approche exploitent des techniques d'apprentissage automatique comme les modèles de Markov cachés Conroy et O'leary (2001), les méthodes Bayésiennes Aone et al. (1998), les réseaux de neurones Krysta M. Svore et Burges (2007).

— L'approche abstractive : consiste à reformuler les phrases du texte. Ces phrases peuvent alors être différentes du texte d'origine Mani (2001). Cette approche peut être divisée en deux tâches : la représentation sémantique de la phrase et la génération de textes. Les travaux liés à cette approche sont prometteurs mais plus exploratoires et ne sont que rarement mis en œuvre en contexte industriel.

Assez peu de travaux ont à notre connaissance été menés pour exploiter des représentations du texte de plus haut niveau, comme une analyse en rôles sémantiques Saggion et Poibeau (2013). Quelques études préliminaires ont pourtant montré des résultats encourageants : Trandabăţ (2011) exploite les rôles sémantiques, la résolution d'anaphores et les relations de discours, Salim et al. (2010) combine l'analyse en rôles sémantiques à des méthodes statistiques pour déterminer les phrases à sélectionner dans une approche extractive, Khan et al. (2015) étend ce travail à une approche abstractive.

3 Machine Reading

Dans l'approche que nous présentons, les textes considérés sont traités à l'aide de la technologie de *Machine Reading* (MR), conçue par Synapse. Les sorties produites par le MR sont proches de celles d'un composant d'annotation en rôles sémantiques Gildea et Jurafsky (2002) (*Semantic Role Labeling*) et sont décrites dans (Laurent et al., 2015). En résumé, ce traitement appliqué sur un texte produit un ensemble de structures appelées CDS (pour *Clause Description Structure*) décrivant chacune une clause (une unité lexicale portant sur une formule actancielle) identifiée dans le texte analysé sous forme de prédicat appliqué à des arguments (sujet, objet, compléments). Le composant de *Machine Reading* a été développé pour le français et l'anglais. Les expérimentations décrites plus bas portent exclusivement sur la langue française.

4 Méthodes mises en place

Nous avons développé trois méthodes différentes capables d'extraire un sous-ensemble d'un groupe de CDS dans le but de produire un résumé de leur contenu : une méthode très simple constituera une *baseline*, une autre, *SumBasic*, est une adaptation au format des CDS d'une méthode de référence en résumé automatique de texte, et la troisième exploite plus directement la structure des CDS et des représentations vectorielles latentes construites à partir de représentations des termes qui les composent.

4.1 Baseline

Cette première méthode est une application directe de l'hypothèse selon laquelle les informations à retenir sont les informations redondantes. Dans cette approche, nous sélectionnons simplement les CDS qui sont générées au moins n fois à partir du corpus pour construire la base de connaissances résumée de ce corpus. Dans les expérimentations menées dont les résultats sont décrits en section 5, nous avons considéré que deux CDS sont équivalentes si elles ont les mêmes sujet, verbe, objet, lieu et temps. Nous avons également fixé la valeur de n à 2 ; cette valeur devrait être plus grande pour traiter des corpus plus importants.

4.2 SumBasic

La méthode *SumBasic* introduite dans (Nenkova et Vanderwende, 2005) est une approche extractive pour le résumé d'un corpus de documents exploitant la fréquence relative des mots non nuls. Dans cette méthode, chaque phrase S se voit assigné un score représentant à quel

point la phrase est constituée de mots fréquents : $Weight(S) = \sum_{w \in S} \frac{p(w)}{|S|}$ où $p(w)$ est le poids de chaque mot ; à l'étape initiale, $p(w)$ reflète la fréquence relative du mot w (rapport entre le nombre d'apparition de w dans le corpus et le nombre total de mots dans le corpus).

Le résumé est construit progressivement en extrayant à chaque étape la phrase présentant le plus grand score jusqu'à obtenir un texte de la taille souhaitée. Pour limiter la redondance du résumé généré, après chaque sélection d'une phrase, le poids de chaque mot de cette phrase est mis à jour : $p_{new}(w) = p_{old}(w)^2$.

L'adaptation de la méthode *SumBasic* au format des CDS est directe : la CDS est l'équivalent d'une phrase dans la méthode originale, et les mots composant chacun des éléments qui constituent cette CDS sont traités comme les mots d'une phrase.

4.3 Embeddings et K-means

Cette méthode consiste à projeter sur un espace vectoriel les CDS d'un corpus de textes puis à appliquer l'algorithme *K-means* sur l'ensemble des vecteurs de ces CDS.

Chaque CDS est transformée en un vecteur qui est le résultat de la concaténation de chacun des vecteurs des éléments qui la constituent. Un élément de CDS (sujet, action, objet ou complément) peut être composé de un ou plusieurs mots. Le vecteur d'un élément est obtenu en sommant les vecteurs de chaque mot qui le compose (la prise en compte ou non des mots vides n'a exposé aucune différence sur les résultats des expérimentations). Les vecteurs de chaque mot (*word embeddings*) ont été pré-entraînés [1] avec FastText Bojanowski et al. (2016). Dans les expérimentation, nous avons considéré dans chaque CDS les éléments suivants : sujet, verbe, objet, lieu et temps. Les *embeddings* de mots étant représentés sur 300 dimensions, le vecteur représentant une CDS obtenu après concaténation a pour dimension 1500.

Nous appliquons ensuite l'algorithme de clustering K-means dans cet espace, l'hypothèse étant que les clusters obtenus réuniront des CDS de sens proches ou identiques. Pour chaque cluster, la CDS la plus proche du centroïde est sélectionnée pour construire le résumé.

Le paramètre k définit donc le nombre de CDS que comportera le résumé généré. On peut ainsi contrôler la taille du résumé et s'assurer que les faits peu appuyés (non redondants) du corpus n'apparaîtront pas dans ce résumé ; en effet, les CDS correspondant à ces faits seront rattachées à des clusters mais elles seront suffisamment éloignées du centroïde pour ne pas être extraites. Pour les expérimentations, nous avons déterminé la valeur de k pour que les résumés générés soient du même ordre de longueur que les résumés manuels.

5 Evaluation

Nous rappelons que l'objectif de ce travail est d'extraire les informations consensuelles des textes d'un corpus afin d'identifier d'éventuelles contradictions dans un nouveau texte. Des évaluations extrinsèques seront effectuées au fil des développements futurs. Nous présentons dans la suite une évaluation intrinsèque. Les CDS extraites par le composant de résumé sont utilisées pour regénérer un résumé textuel qui peut ensuite être comparé à des résumés effectués manuellement via la métrique ROUGE (Lin, 2004) (*Recall-Oriented Understudy for Gisting Evaluation*), traditionnellement utilisée pour évaluer la tâche de résumé automatique.

1. https ://fasttext.cc/docs/en/pretrained-vectors.html

Pour évaluer notre approche, nous avons utilisé le corpus de résumé multi-documents [2] issu du projet RPM2 (Résumé Plurimédia Multi-documents et Multi-opinions) (De Loupy et al., 2010). Ce corpus contient 400 article de journaux (datant de 2009) répartis dans 20 thématiques. Chaque thématique est constituée de 2 clusters, chaque cluster contenant 10 documents. Pour chaque catégorie, les articles du second cluster ont été publiés 1 mois après les articles du premier. Le corpus comporte également des résumés produits à la main. Chaque cluster de 10 documents a été résumé manuellement par 4 annotateurs différents. Il y a donc au total 160 résumés manuels, 4 pour chacun des 40 clusters du corpus.

Afin de pouvoir comparer les performances des méthodes présentées plus haut avec celles d'implémentations de méthodes de références, nous avons utilisé la librairie *sumy* [3] qui fournit des implémentations pour les méthodes de résumé automatique les plus populaires. Le tableau ci-dessous montre les scores des trois variantes présentées plus haut (préfixées de la mention 'CDS_') ainsi que les méthodes LSA, Lex-Rank, KL-Sum et Text-Rank.

	Rappel	Précision	F-score
LSA	0,533	0,243	0,331
Lex-Rank	0,610	0,226	0,325
KL-Sum	0,575	0,203	0,297
Text-Rank	**0,642**	0,160	0,255
CDS_Baseline	0,337	**0,314**	0,293
CDS_Sumbasic	0,591	0,244	0,344
CDS_Kmeans	0,535	0,259	**0,345**

Les approches exploitant les CDS surpassent en terme de précision et de F-score les méthodes de référence pour le jeu de données considéré. La méthode Baseline obtient la meilleure précision, ce qui n'est pas étonnant car celle-ci ne prend pas de risques en ne sélectionnant que du contenu redondant. La méthode K-means obtient le meilleur F-score, ce qui est encourageant car c'est selon nous la méthode qui exploite le mieux les informations portées par les CDS. Il est cependant important de garder à l'esprit que cette méthode d'évaluation présente l'avantage de permettre une comparaison directe aux travaux de la littérature mais n'est pas complètement représentative de l'usage final qui sera fait du composant de résumé automatique. Elle introduit en effet un biais important en traduisant les CDS en texte.

Nous prévoyons pour la suite d'exploiter ces résumés de corpus sous forme de CDS pour mettre en œuvre des mécanismes de détection d'incohérences sémantiques.

Références

Aone, C., M. E. Okurowski, et J. Gorlinsky (1998). Trainable, scalable summarization using robust nlp and machine learning. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pp. 62–66. Association for Computational Linguistics.

Bojanowski, P., E. Grave, A. Joulin, et T. Mikolov (2016). Enriching word vectors with subword information. *arXiv preprint arXiv :1607.04606*.

2. `http://rpm2.org/outils\_et\_ressources.html` - disponible sur demande
3. `https://github.com/miso-belica/sumy`

Conroy, J. M. et D. P. O'leary (2001). Text summarization via hidden markov models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 406–407. ACM.

De Loupy, C., M. Guégan, C. Ayache, et S. Seng (2010). A french human reference corpus for multi-document summarization and sentence compression. In *LREC. 2010.* http ://www.lrec-conf.org/proceedings/lrec2010/pdf/919_Paper.pdf.

Gildea, D. et D. Jurafsky (2002). Automatic labeling of semantic roles. *Computational linguistics 28*(3), 245–288.

Khan, A., N. Salim, et Y. J. Kumar (2015). A framework for multi-document abstractive summarization based on semantic role labelling. *Applied Soft Computing 30*, 737–747.

Krysta M. Svore, L. V. et C. J. Burges (2007). Enhancing single-document summarization by combining ranknet and third-party sources. In *Proceedings of the joint conference on empirical methods in natural language processing and computational natural language learning*.

Laurent, D., B. Chardon, S. Nègre, C. Pradel, et P. Séguéla (2015). Reading comprehension at entrance exams 2015.

Lin, C.-Y. (2004). Rouge : A package for automatic evaluation of summaries. In *Proceedings of the ACL-04 workshop. 2004*. http ://www.aclweb.org/anthology/W/W04/W04-1013.pdf.

Lloret, E. et M. Palomar (2012). Text summarisation in progress : a literature review. In *Journal Artificial Intelligence Review archive Volume 37 Issue 1, January 2012 Pages 1-41* .

Mani, I. (2001). *Automatic summarization*, Volume 3. John Benjamins Publishing.

Nenkova, A. et L. Vanderwende (2005). The impact of frequency on summarization. https ://pdfs.semanticscholar.org/676b/1549adae511164c1b5343f10260fd42035b4.pdf.

Saggion, H. et T. Poibeau (2013). Automatic text summarization : Past, present and future. In *Multi-source, multilingual information extraction and summarization. Springer Berlin Heidelberg, 2013. p. 3-21.* https ://hal.archives-ouvertes.fr/hal-00782442/document.

Salim, N., L. Suanmali, et M. Binwahlan (2010). Srl-gsm : a hybrid approach based on semantic role labeling and general statistic method for text summarization.

Trandabăţ, D. (2011). Using semantic roles to improve summaries. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pp. 164–169. Association for Computational Linguistics.

Summary

This article presents an approach to extracting the information expressed in a corpus of texts and to produce a summary. Several variants of extractive methods of text summarization have been implemented and evaluated. Their main originality lies in the exploitation of structures called CDS (which stands for *Clause Description Structure*) derived from an semantic role labeling component and not directly from the sentences composing the texts. The summary obtained is a subset of the CDSs from the original corpus; this format will allow the detection of textual inconsistencies. In this work, we re-transform the summarized CDS into text to allow comparison of our approach with those of the literature. First results are very encouraging: the proposed methods generally outperform implementations of reference methods.