Approche d'annotation automatique des événements

Rim Faiz*, Aymen Elklifi**

* LARODEC, IHEC de Carthage, 2016 Carthage Présidence, Tunisie
Rim.Faiz@ihec.rnu.tn
**LARODEC, ISG de Tunis, 2000 Le Bardo, Tunisie
Aymen_elkhlifi@yahoo.fr

Résumé. Quotidiennement, plusieurs agences de presse publient des milliers d'articles contenant plusieurs événements de toutes sortes (*politiques, économiques, culturels, etc.*). Les preneurs de décision, se trouvent face à ce grand nombre d'événements dont seulement quelques uns les concernent. Le traitement automatique de tels événements devient de plus en plus nécessaires. Pour cela, nous proposons une approche, qui se base sur l'apprentissage automatique, et qui permet d'annoter les articles de presse pour générer un résumé automatique contenant les principaux événements. Nous avons validé notre approche par le développement du système "AnnotEv".

1 Introduction

Acquérir de la connaissance à partir de textes est une nécessité qui s'est accrue ces vingt dernières années, avec l'essor considérable de la masse de documents disponibles en format électronique, qu'il faut gérer afin d'extraire ou de filtrer les informations pertinentes parmi toutes celles contenues dans ces documents (Faiz, 2006). A titre d'exemple; les événements boursiers sont nombreux et diversifiés. Les experts de la bourse doivent analyser ces événements en un temps relativement raisonnable pour prendre des décisions importantes. Il s'agit, donc, d'annoter les documents présentant des événements pour pouvoir extraire ceux qui sont pertinents. C'est dans ce cadre que s'inscrit notre travail dont l'objectif est de développer une approche qui annote automatiquement ces articles de Presse.

La suite du document est organisée comme suit : nous commençons, dans la section 2, par décrire les principaux systèmes d'annotations existants. Au cours de la section 3, nous présentons notre approche d'annotation, qui a été validée par le système AnnotEv lequel sera présenté et évalué dans la section 4. Enfin, dans la section 5, nous présentons quelques perspectives de notre travail.

2 Présentation de quelques systèmes d'annotation

Plusieurs méthodes et techniques sont utilisées par les systèmes d'annotations dédiés au Web sémantique telles que l'Exploration Contextuelle (Desclés, 1997), les graphes conceptuels (Roussey et al, 2002), les méta-thésaurus (Khelif et al., 2004) et les indicateurs linguistiques (Muller et al., 2004). Nous pouvons citer :

Le système EXCOM (Djaoua et al., 2006) utilise un ensemble d'outils linguistiques qui visent à annoter un document par un ensemble de connaissances aussi bien internes

qu'externes. Le système prend en entrée des textes et procède à l'annotation sémantique et discursive de certains segments à partir des points de vue de fouille. L'annotation sémantique de ces segments fait appel à la technique linguistique et informatique d'exploration.

Annotea (Kahan et al., 2001) est un système client-serveur collaboratif pour l'annotation de documents. Les annotations, externes aux documents, sont écrites en RDF. Des utilisateurs, connectés à un serveur d'annotations, peuvent les ajouter, les modifier et les consulter. Cependant, nous constatons que l'affichage des annotations est séparé du document, ce qui devrait être résolu pour améliorer la compréhension et l'efficacité de telle annotation.

SyDoM (Roussey et al., 2002) est un système d'annotations sémantiques de pages Web. Il permet l'enrichissement de ces pages par le biais des connaissances pour pouvoir les retrouver sans tenir compte de leur langue d'écriture. Cependant, SyDoM ne peut effectuer des recherches que sur des pages Web qu'il a annoté, c'est-à-dire qu'il est incapable d'interroger des pages Web quand les annotations ont été créés à l'aide de thésaurus différents.

Nous constatons que tous ces travaux s'intéressent à l'annotation des documents en général (*documents Web, articles scientifiques, documents multimédias, services Web, etc.*), et peu de travaux se sont intéressés à l'annotation des informations temporelles. Nous pouvons citer quelques uns de ces travaux :

Muller et Tannier (2004) ont travaillé sur l'annotation automatique d'informations temporelles dans des textes (*dépêches d'agence*), particulièrement, sur les relations entre les événements introduits par les verbes dans chaque clause. La procédure d'annotation est décomposée en deux étapes : marquer une expression temporelle dans un document et identifier la valeur de temps indiqué par l'expression. Un déclencheur lexical génère les mots réservés.

Setzer et Gaizauskas (2000) ont proposé un système d'annotation qui se base sur la détermination des événements et les relations entre eux en se basant sur des connaissances linguistiques. Cependant, nous constatons, le système ne prend pas compte de l'état et des questions de type *Qui* et *Comment* pour extraire un événement, et d'autre part, il se base uniquement sur les marqueurs temporels pour déterminer les relations entre les événements, hypothèse qui n'est pas tout à fait validée étant donné qu'il existe des relations implicites inter-événements qui sont exprimées sans utiliser des marqueurs temporels.

3 Approche proposée pour l'annotation des événements

Nous constatons que les approches proposées pour l'annotation de l'information temporelle sont principalement linguistiques et se basent sur les indices temporels. Nous nous sommes intéressés plutôt à l'annotation des événements sous forme de métadonnées liées aux documents. Notre travail ne se limite pas à la détection des événements, mais permet aussi de regrouper les événements similaires pour faciliter un traitement ultérieur (indexation, remplissage des formulaires, etc.). Le processus d'annotation automatique des documents que nous présentons, s'effectue en quatre étapes :

1. **Prétraitement** : qui consiste d'une part, à détecter les frontières des phrases dans un texte et d'autre part, à nommer les entités.
2. **Annotation des événements** : qui utilise un classificateur jouant le rôle d'un filtre pour les phrases non événementielles.
3. **Clustering** : qui consiste à regrouper les phrases se référant au même événement. Nous avons proposé, pour cette étape, une nouvelle mesure de similarité entre les événements.

4. **Annotation du document** : qui prendra différentes formes (*phrase, formulaire, concept, etc.*) selon le domaine d'application.

3.1 Etape 1 : Prétraitement

Dans notre étude, le prétraitement est l'application de certains outils de Traitement Automatique des Langage Naturelles (TALN) au texte brut pour le segmenter en phrases et annoter les entités nommées.

Pour la tâche de segmentation, nous avons utilisé le système SegaTex développé par Mourad (2001) pour son aspect multilingue et sa disponibilité. Concernant la nomination des entités, nous avons utilisé le système GATE (Bontcheva et al., 2004).

3.2 Etape 2 : Annotation des événements

Un événement est un objet spécifique qui se produit à un instant spécifique et dans un endroit bien déterminé. Notre objectif est d'identifier tous les événements présents dans un document. Nous marquons par une balise chaque événement détecté. Pour cela, un modèle de classification est construit automatiquement, il permet de prédire si une phrase contient un événement ou non. Nous avons utilisé, dans un premier temps, les attributs qui se rapportent aux événements tels qu'ils sont définis par Naughton et al. (2006). Ces attributs sont les suivants : *Longueur de la phrase, Position de la phrase dans le document, Nbre de lettres capitales, Nbre de Stopwords, Nbre de noms de villes; Nbre de marques numériques.*

Il s'agit donc de classer une phrase comme étant événementielle ou non. Plusieurs techniques d'apprentissage automatique peuvent être utilisées tels que les réseaux de neurones, l'arbre de décision, les réseaux bayésiens, etc. Nous avons choisi l'arbre de décision puisque la construction de l'arbre est moins paramétrable, comparé aux autres techniques, ce qui permet la réduction de la complexité du système.

L'ensemble de l'apprentissage (*training set*) a été annoté par un expert. Pour chaque article de presse, les événements sont annotés comme suit : L'annotateur est amené à assigner des étiquettes à chaque phrase représentant un événement ; Si une phrase se rapporte à un événement, il lui assigne l'étiquette "Oui" sinon "Non".

Nous avons appliqué à ce même ensemble d'apprentissage, différents algorithmes de construction des arbres de décision. Puis, nous avons choisi le modèle qui a le plus grand PCC (Pourcentage de Classification Correcte). Le résultat de cette étape est l'ensemble des phrases se référant à des événements.

3.3 Etape 3 : Le Clustering

Au cours de cette troisième étape, nous regroupons les phrases se référant aux mêmes événements par l'application de l'algorithme « Hierarchical Clustering (HAC) », qui assigne initialement chaque objet à un cluster, puis fusionne, à plusieurs reprises, les clusters jusqu' à ce qu'un des critères d'arrêts soit satisfait (Manning et Schutze, 1999).

Dans ce cadre, nous avons proposé une nouvelle mesure de **similarité sémantique** entre les événements. Nous signalons qu'il existe plusieurs mesures de similarités entre les documents, telle que le Cosinus de Salton (1988), le Cosinus dans l'espace distributionnel et la distance de khi-deux (Lebart et Rajman, 2000). D'autres mesures, qui nous intéressent le

plus, portent sur la similarité entre les phrases, dont la plus récente est la mesure de Naughton (2006).

Ainsi, la nouvelle mesure de similarité que nous proposons, est inspirée de tf-idf (weight term frequency–inverse document frequency) et tient compte également de la position des clusters dans le document. Afin de pouvoir regrouper des phrases exprimant le même événement par deux lexiques différents, nous avons utilisé une base de synonymes permettent le remplacement des instances par leurs classes. Par exemple, soient les deux phrases événementielles suivantes, initialement considérées comme deux clusters C1 et C2:

C1 : *In Baquba, two separate shooting incidents Sunday afternoon left six dead and 15 wounded, officials said.*

C2 : *In other attacks reported by security and hospital officials, two car bombings in the northern city of Kirkuk killed 10 and wounded 32, and a blast in the southern city of Basra killed five and injured 15.*

Nous remarquons que les mots (*shooting incidents* et *car bombings*), (*dead* et *killed*) impliquent le même sens. D'où l'idée de remplacer ces mots par leurs classes afin d'augmenter la similarité entre les deux clusters. Nous avons défini alors **SIM** comme suit :

$$SIM(C_1, C_2) = \frac{\sum_{j=1}^{t} Ct_{1j} Ct_{2j}}{\sqrt{\sum_{j=1}^{t} Ct_{1j}^2 + \sum_{j=1}^{t} Ct_{2j}^2}}$$

Avec Ct_{ij} une classe de la base des synonymes $Ct_{ij} = tf(t_i, c) \times \log\left(N / df(t_i)\right)$

Par la suite nous présentons **FSIM** par la formule suivante :

$FSIM(C_1, C_2) = \alpha \times SIM(Ct_1, Ct_2) + (1 - \alpha) \times Cos(Ct_1, Ct_2)$ avec $\alpha \in [0,1]$ (cf. Elkhlifi et Faiz, 2006).

3.4 Etape 4 : Annotation du document

Nous continuons l'enrichissement du document par d'autres métadonnées qui seront très utiles pour d'autres applications (*RI, Résumé Automatique, Question-Réponse, Indexation*).

Une forme possible de métadonnée est le remplissage de formulaires, c'est-à-dire stocker les événements dans une base de données en répondant à des questions bien déterminées, par exemple : **Keyword**: Killed ; **Location**: Baghdad ; **Time/date**: 2 p.m ; **Person**: U.S. soldier.

Une autre forme de métadonnée est le résumé automatique, qui consiste à marquer, par des balises, les phrases qui forment le résumé d'un document. En général, le but d'un système de résumé automatique est de produire une représentation condensée du contenu dès son entrée, où les informations importantes du texte original sont préservées, il faut prendre en considération les besoins de l'utilisateur et de la tâche spécifiée (Minel et al., 2001).

Dans le cadre de notre étude, nous proposons un résumé informatif contenant les informations essentielles contenues dans l'article. Ce résumé est sélectif puisqu'il néglige les aspects généraux de l'article. Mais il est ciblé étant donné qu'il concerne les événements.

A partir de chaque cluster généré par la troisième étape, nous annotons l'article par les principaux événements qu'il contient. Nous utilisons une heuristique qui est : La phrase ayant les attributs maximaux est la meilleure pour annoter l'article.

4 Expérimentation et Résultat

Pour valider notre approche, nous avons développé le système **AnnotEv**. Il comporte les cinq modules suivants : la *segmentation du texte brut*, la *reconnaissance des entités nommées*, l'*annotation des événements*, le *regroupement des événements similaires* et le *résumé automatique*. Nous avons utilisé un corpus de 82 articles concernant la guerre au Proche Orient en 2006 et ce à partir de 5 agences de presse : *CNN* : 13 articles, *Reuters* : 17 articles, *BBC*: 14 articles, *Associated Press* : 22 articles et *AFP* : 16 articles. La longueur moyenne d'une phrase est de 12.23, avec une moyenne de 6.05 événements par article, pour un total d'environ 91 500 mots, 7479 phrases et 496 événements.

A travers une interface de gestion du corpus, un expert peut choisir un article et parcourir ses phrases pour marquer celles qui sont événementielles. Nous avons ensuite, généré automatiquement un fichier en format ARFF qui présentera l'entrée d'apprentissage sous Weka[1].

Nous avons utilisé les algorithmes suivants pour construire des arbres de décision : **RandomTree**, **J48** et **ADTree**. Nous avons mesuré la Précision et le Rappel tels qu'ils sont définis par Naughton (2006). Pour chacun des algorithmes, nous avons obtenu ces résultats :

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.706	0.313	0.706	0.706	0.706	yes
0.688	0.294	0.688	0.688	0.688	no

TAB. 1 – *Résultat avec Random Tree.*

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.824	0.25	0.778	0.824	0.8	yes
0.75	0.176	0.8	0.75	0.774	no

TAB. 2 – *Résultat avec J48.*

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.765	0.313	0.722	0.765	0.743	yes
0.688	0.235	0.733	0.688	0.71	no

TAB. 3 – *Résultat avec ADTree.*

5 Conclusion et perspectives

Nous avons présenté une approche d'annotation automatique des événements qui se base sur l'apprentissage automatique, accompagnée d'une exploitation de l'annotation qui constitue un résumé automatique. Nous avons, également, proposé une nouvelle mesure de similarité sémantique : FSIM entre les événements.

Notre approche se compose de quatre étapes : en commençant, dans une première étape, par le prétraitement qui consiste à appliquer des outils de TALN pour préparer les données. Dans une deuxième étape, un classifieur qui permet de filtrer les phrases événementielles. Au cours de la troisième étape, les phrases sont regroupées dans des clusters selon leur degré de similarité (FSIM). Dans la dernière étape, un résumé est généré automatiquement et portant sur les principaux événements constituant l'article. Nous avons validé notre approche sur un corpus d'articles de presse. Une des perspectives que nous proposons est d'adopter AnnotEv à la langue arabe.

[1] http://www.cs.waikato.ac.nz/ml/weka/

Références

Bontcheva K., V. Tablan et D. Maynard, H. Cunningham (2004). Evolving GATE to Meet New Challenges in Language Engineering. *Natural Lang. Engineering*. 10, pp. 349-373.

Desclés J.P. (1997). *Systèmes d'exploration contextuelle, Co-texte et calcul du sens*, Claude Guimier, Presses de l'université de Caen, pp. 215-232.

Djaoua B., J.G. Flores, A. Blais, J.P. Desclés, G. Gael, A. Jackiewicz, F. Le Priol, N.B. Leila et B. Sauzay (2006). EXCOM : an automatic annotation engine for semantic information. *FLAIRS 2006*.

Elkhlifi A. and Faiz R. (2007). Machine Learning Approach for the Automatic Annotation of Events. *FLAIRS 2007*, Florida.

Faiz R. (2006). Identifying relevant sentences in news articles for event information extraction. *International Journal of Computer Processing of Oriental Languages*, Vol. 19, No.1, pp. 1-19.

Kahan J. et M-R. Koivunen (2001). Annotea: an open RDF infrastructure for shared Web annotations. *Proceedings of the 10th international conference on World Wide Web*.

Khelif K. et R. Dieng-Kuntz (2004). Web sémantique et mémoire d'expériences sur les biopuces. *Second séminaire francophone du Web Sémantique Médical,* Rouen.

Lebart L., Rajman M. (2000). *Computing Similarities*. In Dale R., Moisl H., Somers H. editors : Handbook of Natural Language Processing, Marcel Dekker, pages 477-505, New York

Manning C. et H. Schutze (1999). *Foundations of Statistical Natural Language Processing.* MIT Press.

Minel J.L., J.P. Desclés, E. Cartier, G. Crispino, S. Ben hazez et A. Jackiewicz (2001). Résumé automatique par filtrage sémantique d'informations dans des textes. *Revue TSI*. Hermès.

Mourad Gh. (2001). *Analyse informatique des signes typographiques pour la segmentation de textes et l'extraction automatique de citations. Réalisation des Applications informatiques: SegATex et CitaRE*. Thèse de doctorat Université Paris-Sorbonne.

Muller P. et X. Tannier (2004). Annotating and measuring temporal relations in texts. *Proceedings of Coling*, volume I. Genève, Association for Computational Linguistics.

Naughton M., N. Kushmerick, et J. Carthy (2006). Event extraction from heterogeneous news sources. *Proc.Workshop Event Extraction and Synthesis, American Nat. Conf. AI*.

Roussy C., S. Calabretto et J.M Pinon, (2002). SyDoM : un outil d'annotation pour le Web sémantique. *Proceedings of Journées Scientifiques Web sémantique*.

Salton G. et C. Buckley (1988). Term-weighing approaches in automatic text retrieval. *Information Processing & Management*, 24(5), p.513-523.

Setzer A. et R. Gaizauskas (2000). Annotating Events and Temporal Information in Newswire Texts. *LREC 2000, workshop: Information Extraction Meets Corpus Linguistics*.

Summary

Daily, several news agencies publish thousands of articles concerning several events of all kinds (*political, economic, cultural, etc)*. The decision makers find themselves in front of a great number of events of which only some relate to them. The automatic treatment of such events becomes increasingly necessary. Thus, we propose an approach, based on the machine learning that allows annotating news articles to generate an automatic summary of the events. We validated our approach by the development of the "AnnotEv" system.