UMR IRISA
INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTÈMES ALÉATOIRES
Campus universitaire de Beaulieu, entrée Bât 12 F ● Avenue du Général Leclerc
35042 Rennes Cedex, France ● Tél. : +33 (0)2 99 84 71 00 ● Fax : +33 (0)2 99 84 71 71
www.irisa.fr

Construction incrémentale dynamique de graphes de connaissance par fouille de contenus

Le groupe Sipa Ouest-France dispose d'une base de contenus couvrant de nombreuses publications du groupe, dont les multiples éditions de Ouest-France, et regroupant actuellement plus de 35 millions de documents sur des sujets divers. Cette base de contenus contient des informations et connaissances précieuses dont une partie peut être extraite de manière automatique. Par exemple, la détection d'entités (*e.g.*, noms de personnalités) et des relations entre entités permet de faire des liens vers une base de connaissance de référence (*entity linking*) et, le cas échéant, de mettre à jour cette dernière en fonction de l'analyse des contenus. Le groupe Sipa Ouest-France maintient ainsi plusieurs bases de connaissance correspondant aux référentiels métiers du groupe. Aujourd'hui, ces référentiels sont construits par les acteurs du journalisme (documentalistes, journalistes), en ne s'appuyant que marginalement sur l'extraction de connaissance à partir des contenus, et ne sont pas utilisés pour guider l'extraction d'information dans les contenus. L'idée principale de la thèse est donc d'**étudier des mécanismes permettant de construire de manière incrémentale et dynamique les graphes de connaissance** correspondant à ces référentiels, **en s'appuyant conjointement sur la base de contenus du groupe et sur les référentiels existants**. En d'autres termes, on souhaite envisager l'enrichissement de la base de connaissance et l'annotation de documents comme des démarches complémentaires qui participent interactivement à la supervision d'un modèle unifié.

L'extraction de données structurées à partir de contenus textuels aboutit à présent à des résultats prometteurs (Ji *et al.*, 2011; Surdeanu 2013) mais soulève encore de nombreux problèmes. En particulier, les entités et relations extraites automatiquement des contenus textuels sont souvent très nombreuses et bruitées, rendant particulièrement complexes leur compréhension et leur validation par un service de documentation ou un expert du domaine. Il est dès lors crucial de mettre en place des techniques permettant aux concepteurs de bases de connaissance d'explorer les résultats de l'analyse automatique et d'éliminer le bruit. De plus, dans le contexte du journalisme, il est important de tenir compte, dans un référentiel, de la temporalité des connaissances manipulées (*e.g.*, sur quelle période Mr. X a-t-il été maire de Y ?), ce qui reste difficile avec les techniques classiques d'extraction. Ce dernier point requiert des méthodes itératives et dynamiques de construction d'un référentiel en évolution permanente qui n'ont que peu été étudiées à ce jour.

Dans ce contexte, l'objectif général de la thèse est de **proposer des approches permettant d'exploiter des graphes de connaissance existants pour faciliter l'extraction de données, ainsi que leur interprétation et leur validation par les acteurs du domaine**, **de manière à compléter et faire évoluer le modèle de connaissance**. En retour, cette évolution doit permettre une meilleure exploitation des contenus pour enrichir les graphes, mais également pour avoir une meilleure vue de ce que ces contenus contiennent globalement.

Plusieurs problématiques découlent de cet objectif général :

- Comment mettre en place et maintenir un système d'apprentissage bidirectionnel supervisé (par un service documentation, par un groupe d'utilisateurs experts du domaine, par l'usage implicite des parcours de lecture) où les contenus viennent compléter et étendre le graphe de connaissance, et où la précision du graphe de connaissance permet de réduire le bruit, le silence, et les ambiguïtés du système d'extraction ?
- Quel « *bootstrap* » ? La question qui se pose est celle d'un modèle de connaissance initial permettant de débuter le processus de construction incrémentale et dynamique du référentiel. Que faut-il décrire et modéliser des domaines traités, du plus générique au plus spécialisé, pour être en capacité de commencer le travail d'enrichissement des contenus ? Peut-on s'appuyer sur une analyse automatique des contenus pour commencer à décrire une base de connaissance ?
- Comment visualiser les connaissances modélisées et les contenus dont elles sont extraites de manière à permettre une validation rapide des données structurées obtenues automatiquement ? Peut-on créer

des parcours de navigation dans la base de contenus permettant une telle validation rapide des connaissances extraites ?

Sur le plan scientifique, nous identifions deux grands axes de travail pour apporter des éléments de réponses à ces problématiques. Au regard des méthodes distributionnelles développées récemment (Le and Mikolov, 2014), une première voie possible consiste à examiner des approches capables de raisonner à la fois sur les unités lexicales et documents issus d'un corpus de textes (*word and document embedding*) et sur les entités relationnelles d'un graphe de connaissance (*knowledge graph embedding*), à l'instar de Wang *et al*. (2014). On pourra notamment vérifier si l'intégration conjointe d'entités et de mots dans le même espace vectoriel continu peut préserver les relations entre les entités dans le graphe de connaissance et les similarités entre les mots observés dans un corpus. Cette intégration pourrait aider à compléter un graphe de connaissance en prédisant de nouvelles relations/faits après observation dans un corpus. Inversement, elle pourrait permettre d'expliquer à l'aide des connaissances existantes des relations statistiques entre mots observées sur le corpus. Par ailleurs, on cherchera à étendre le modèle et les techniques d'extraction, ainsi que l'interaction entre ces deux éléments, pour représenter différentes versions d'un même objet (document, entité ou événement). Il est très courant d'observer plusieurs interprétations d'un même objet qui diffèrent selon le point de vue du journaliste mais également au cours du temps (Wang *et al*. 2012; Popescu and Strapparava, 2014, 2015; Hamilton *et al*., 2016). À partir d'une même source, plusieurs auteurs peuvent aborder l'information sous des angles différents. Ces angles peuvent être considérés comme de multiples versions/embranchements qui s'entrecroisent et évoluent au cours du temps. Dans l'optique de construire une représentation combinant de multiples vues, on pourra s'inspirer du modèle de branches de développement popularisé par le logiciel de gestion de versions Git. On étudiera également l'impact de relations versionnées sur le processus interactif d'extraction d'informations et de construction des référentiels qui aura été mis en place.

La thèse se déroulera à Rennes, dans le cadre d'une collaboration étroite entre l'équipe SIB du groupe Sipa Ouest-France et l'équipe Linkmedia de l'IRISA (UMR 6074), laboratoire de recherche en informatique du Grand Ouest. En particulier, les deux équipes participent au projet intitulé *Knowledge-mediated Content and Data Interactive Analytics*, associant des acteurs des médias et des équipes de recherche Inria. La thèse, sous convention CIFRE, sera menée dans le cadre de ce projet, permettant des échanges significatifs avec les partenaires du projet et un partage des approches et résultats développés.

Références

W.L. Hamilton, J. Leskovec & D. Jurafsky (2016). Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

H. Ji, R. Grishman & H. Trang Dang (2011). Overview of the TAC2011 Knowledge Base Population Track. In 2011 Text Analysis Conference.

Q. Le & T. Mikolov (2014). Distributed representations of sentences and documents. In Intl. Conf. on Machine Learning, pp. 1188-1196.

O. Popescu & C. Strapparava (2014). Time corpora: Epochs, opinions and changes. Knowledge-Based Systems, 69, pp. 3-13.

O. Popescu & C. Strapparava (2015). SemEval-2015 Task 7: Diachronic text evaluation. Proceedings of SemEval.

M. Surdeanu (2013). Overview of the TAC2013 Knowledge Base Population Evaluation: English Slot Filling and Temporal Slot Filling. In 2013 Text Analysis Conference.

C. Wang, D. Blei & D. Heckerman (2012). Continuous time dynamic topic models. *arXiv preprint arXiv:1206.3298*.

Z. Wang, J. Zhang, J. Feng & Z. Chen (2014). Knowledge Graph and Text Jointly Embedding. In Empirical Methods on Natural Language Processing, pp. 1591-1601.

Contacts

Pascale Sébillot	pascale.sebillot@irisa.fr
Guillaume Gravier	guillaume.gravier@irisa.fr
Michel Le Nouy	michel.lenouy@ouest-france.fr