Construction automatisée d'une base de connaissances

Olivier Mesnard[1, 2] Yoann Dupont[3] Jérémy Guillemot[1] Rashedur Rahman[1,4]
(1) IRT SystemX, 8 avenue de la Vauve BP 30012, 92120 PALAISEAU, France
(2) CEA LIST Nanno Innov av de la Vauve, 92120 PALAISEAU, France
(3) Expert System France Tour Mattei, 207 rue de Bercy, 75012 Paris, France
(4) LIMSI-CNRS, rue John von Neumann, 91403 Orsay
olivier.mesnard@irt-systemx.fr, ydupont@expertsystem.com,
jeremy.guillemot@irt-systemx.fr, rashedur.rahman@irt-systemx.fr

RÉSUMÉ

Le système présenté permet la construction automatisée d'une base de connaissances sur des personnes et des organisations à partir d'une collection de documents. Il s'appuie sur de l'apprentissage distant pour l'extraction d'hypothèses de relations entre mentions d'entités qu'il consolide avec des informations orientées graphe.

ABSTRACT

Automated Building a Knowledge Base

We present a system to build automatically a knowledge base on organisations and persons from a collection of documents. The chain combines named entity extraction, distant learning to generate relation hypothesis which are consolidated with graph-oriented information.

MOTS-CLÉS : plate-forme de veille, extraction de relation, constitution de base de connaissances.

KEYWORDS: intelligence tool, relation extraction, knowledge base construction.

Le projet IMM de l'ITRT SystemX, qui a démarré il y a trois ans, se propose d'assembler les outils de différents partenaires pour construire une plateforme de veille.
Un environnement d'intégration a été développé pour accueillir les différents modules : traduction, extraction d'information, recherche d'information, analyse de réseaux sociaux... et ainsi prototyper des applications innovantes dans le domaine de l'analyse des données peu ou non structurées. L'adaptation au domaine, le multilinguisme, le passage à l'échelle, la gestion des entités nommées (EN) sont les principales problématiques du projet.

Cet environnement offre un ensemble de services que l'on peut résumer ainsi :

- une plateforme d'intégration qui privilégie la communication asynchrone entre composants ;
- un service de déploiement qui permet l'instanciation automatique de machines virtuelles et l'installation automatisée des composants dans le cloud ;
- un service d'intégration continue qui contrôle la non régressions sur les chaînes de traitement ;
- une interface d'administration pour créer l'environnement d'exécution d'une expérimentation, sélectionner les composants et créer les chaînes de traitement.

1 Description du système

La construction de la base connaissances repose sur deux grandes étapes : l'extraction d'hypothèses de relations entre des mentions d'entités à partir de textes suivie d'une consolidation en entités et relations pour alimenter la base de connaissances.

Extraction d'hypothèses de relations La première étape consiste à repérer les mentions d'entités nommées (EN) dans les textes. Elle porte essentiellement sur 6 types d'entités : personne, organisation, entité géo-politique, date, montant. Les entités comme les personnes ou les organisations possèdent des propriétés comme par exemple le prénom, le titre etc... L'extraction des EN s'appuie sur plusieurs systèmes : Luxid d'Expert System(Luxid, 1), Lima du CEA(Lima, 1), et Stanford NER (Standford, 1) qui ont été adaptés au modèle du projet et sont utilisés conjointement pour améliorer le rappel.

À partir des mentions d'entités reconnues, le système propose des hypothèses de relations binaires entre celles-ci. Par exemple, une personne et une organisation appartenant à un même fragment de texte peut donner lieu à l'hypothèse d'une relation *fondée_par*. Nous nous sommes placés dans le cadre d'un apprentissage distant basé sur Wikidata pour produire un corpus d'apprentissage. Cela évite d'annoter manuellement un corpus : on repère des fragments de texte qui contiennent les deux entités parties prenantes d'un fait extrait de Wikidata. On fait l'hypothèse que ces fragments peuvent exprimer le fait. Plusieurs paramètres (distance maximale entre entités, nature et taille des fragments à gauche et à droite...) permettent de contrôler cette production d'exemples.

Pour l'apprentissage, nous avons utilisé MultiR (Hoffmann *et al.*, 2011) avec le jeu de traits défini par Zhou *et al.* (2005). Les hypothèses de relations générées par MultiR permettent d'alimenter une base orientée graphe.

Construction de la base de connaissance Une même entité peut être mentionnée différemment dans les textes, par exemple par le prénom, le nom, le nom complet, etc. A l'inverse, deux entités différentes peuvent être homonymes. Il s'agit donc de regrouper les mentions présentes dans les textes lorsqu'elles désignent la même entité. Luxid effectue le suivi des entités au niveau du document et nous exploitons ce lien pour le regroupement. Le composant Stanford NER n'offre pas cette facilité et nous effectuons donc nous même ce suivi pour cet outil. Nous identifions les entités au niveau de la collection en deux étapes : 1) en regroupant les mentions par similarité selon leurs composants (prénom, nom, etc. pour les personnes) et 2) en validant les clusters obtenus par une distance cosinus.

Pour une entité et une relation données, nous requêtons la base pour obtenir les différentes hypothèses possibles, i.e. les différentes entités possibles. Nous consolidons la valeur de la relation pour une entité en sélectionnant la plus probable selon plusieurs traits dont le score de confiance calculé par MultiR, la fréquence des occurrences, et des traits calculés sur le graphe formé par les entités voisines dans les textes.

2 Futurs développements

La chaîne actuelle fonctionne en anglais mais la généralisation de l'approche à plusieurs langues est envisagée dans un futur proche. Enfin nous évaluerons nos résultats en participant à la tâche Slot Filling Cold Start de KBP. Par ailleurs, les données extraites d'une collection de documents ou du web vont pouvoir être rapprochées des graphes issus de l'analyse des réseaux sociaux qui sont fabriqués avec d'autres composants de la plate-forme.

Références

HOFFMANN R., ZHANG C., LING X., ZETTLEMOYER L. & WELD D. S. (2011). Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies - Volume 1*, HLT '11, p. 541–550, Stroudsburg, PA, USA : Association for Computational Linguistics.

LIMA (1). https ://github.com/aymara/lima.

LUXID (1). http ://www.expertsystem.com/products/luxid-annotation-server/.

STANDFORD (1). http :nlp.stanford.edusoftwarecrf-ner.shtml.

ZHOU G., SU J., ZHANG J. & ZHANG M. (2005). Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, p. 427–434, Ann Arbor, Michigan : Association for Computational Linguistics.