Relier automatiquement des entités textuelles à des concepts d'une ontologie par apprentissage avec (presque) aucune donnée

Arnaud Ferré[1,2], Mouhamadou Ba[2], Robert Bossy[2], Thomas Lavergne[1], Louise Deléger[2], Pierre Zweigenbaum[1], Claire Nédellec[2].

[1] LIMSI, CNRS, Université Paris-Saclay, 91405 Orsay, France
[2] MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France
arnaud.ferre@universite-paris-saclay.fr

Mots-clés : Ontologie, apprentissage automatique, traitement automatique des langues naturelles, extraction d'information, normalisation d'entité.

1 Introduction

Le monde de la recherche scientifique produit une quantité gigantesque de connaissances sous forme d'articles. Ce volume de données est si énorme qu'il est difficile, voire impossible, pour les scientifiques d'effectuer une veille manuelle exhaustive (Cohen et al., 2014). Pour pouvoir assister les chercheurs, une solution est d'utiliser une approche de traitement automatique du langage naturel appelée "extraction d'information" (EI) (Russell et al., 2016). L'EI est le processus d'acquisition de connaissances à partir de texte par la recherche d'occurrences d'une classe particulières d'objets ou de relations entre ces objets. Le domaine des sciences du vivant et de l'environnement a des besoins importants en EI (Bossy et al., 2012; Craven et al., 1999; Marsi, 2014; Weeber et al., 2005) et a l'avantage de déjà posséder de nombreuses bases de connaissances (Ashburner et al., 2000; Bossy et al., 2016; McCray, 1989).

2 Etat de l'art pour la tâche de normalisation

Parmi les différentes tâches d'EI, la normalisation d'entité consiste à relier automatiquement des expressions du texte désignant des entités à des références sémantiques précises, telles que les entités décrites dans une ontologie ou une base de connaissance (Morgan et al., 2008). Les concepts d'une ontologie sont fréquemment utilisés comme références sémantiques depuis la fin des années 90 (Faure et al., 1998; Hwang, 1999). La normalisation est une tâche centrale de l'EI, qui peut être vue comme un problème de classification : classer les mentions d'entités d'intérêt dans la ou les classes correctes représentant les concepts d'intérêt pour la tâche. Elle est souvent abordée comme la recherche du concept le plus similaire à une expression donnée. Deux grandes catégories d'approches sont rencontrées :

- Les approches fondées sur les similarités de forme, de vocabulaire ou de structure syntaxique entre les termes du texte désignant une entité et les termes représentant les étiquettes des concepts de référence (Aronson, 2001; Golik et al., 2011; Kang et al., 2013). Le concept dont l'étiquette est la plus similaire à une mention du texte à

normaliser est choisie. Les approches de ce type ont une limitation importante : les mentions qui ne présentent aucune similarité de forme avec des étiquettes de concept ne peuvent être normalisées. C'est le cas par exemple des termes nouveaux.

- Plus récemment, les approches basées sur des représentations distributionnelles des mentions et des étiquettes de concepts ont émergé qui répondent à cette limitation. Dans l'espace vectoriel manipulé, si deux expressions ont un sens proche, il est attendu que leurs représentations soient spatialement proches selon l'hypothèse que les vecteurs capturent l'usage, donc une partie du sens des expressions considérées. En exploitant la distribution des mots dans de grands corpus de textes (Harris, 1954), plusieurs méthodes (Mikolov et al., 2013; Pennington et al., 2014) construisent des vecteurs possédant cette propriété. Le concept dont le vecteur est le plus proche du vecteur d'une mention du texte est choisi comme pour normaliser cette mention. Le cosinus des deux vecteurs est classiquement choisi comme mesure de similarité.

L'efficacité de ces dernières méthodes dépend de la qualité des représentations sémantiques utilisées, du domaine et de la tâche. Pour répondre au besoin d'adaptation au domaine, l'approche par apprentissage supervisé permet d'apprendre une mesure de similarité adaptée par entraînement avec les données associées au domaine (Leaman et al., 2013; Limsopatham et al., 2016). Néanmoins, ces méthodes sont dépendantes de la quantité et de la qualité des données d'entraînement. Ces méthodes obtiennent de bons résultats lorsque la quantité de ces données est élevée (Sil et al., 2018). Néanmoins, produire ces données demande des efforts importants (Uschold et al., 1995), notamment dans les domaines de spécialité pour lesquels le niveau d'expertise requis pour les annotateurs est élevé et le nombre de classes (de concepts) considérées important (Delèger et al., 2016; Lipscomb, 2000). Le manque d'exemples d'apprentissage représente un défi important de l'apprentissage supervisé (Larochelle et al., 2008; Xian et al., 2017). C'est à ce défi que répond la méthode CONTES (CONcepts to TErms System) décrite ici.

3 Méthode : CONcepts to TErms System (CONTES)

CONTES (Ferré et al., 2017) est une méthode de normalisation par les concepts d'une ontologie, fondée sur des représentations distributionnelles et sur de l'apprentissage supervisé. La méthode ne souffre pas de la grande variabilité morphologique des mentions d'entités. Elle est fondée sur l'intégration des connaissances issues de l'ontologie du domaine définie pour la tâche, ce qui lui permet notamment de s'accommoder de l'insuffisance des données annotées. Les principales connaissances utilisées sont les relations d'ordre ("is_a") entre les concepts, qui permettent de construire directement des représentations vectorielles encodant la hiérarchie des concepts plutôt que la forme de leurs étiquettes. Ces représentations sont définies dans un espace distinct de celui des représentations de mentions. Une étape d'apprentissage permet alors d'apprendre une fonction de projection des représentations de mentions vers les représentations de concepts qui devraient les normaliser. La méthode représente aujourd'hui l'état de l'art sur la tâche de normalisation des habitats bactériens proposée par le challenge BioNLP Shared-Task 2016 (Delèger et al., 2016). La méthode est disponible et documentée en ligne et peut être utilisée grâce à la suite logicielle libre AlvisNLP/ML (Ba et al., 2016).

Remerciements

Ce travail est soutenu par le projet "IDI 2015" de l'IDEX Paris-Saclay, ANR-11-IDEX-0003-02, ainsi que par l'appel à projets émergents du département STIC de l'Université Paris-Saclay et le Réseau Francilien en Sciences Informatiques (RFSI).

Références

ARONSON A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium (p. 17). American Medical Informatics Association.*

ASHBURNER M., BALL C. A., BLAKE J. A., BOTSTEIN D., BUTLER H., CHERRY J. M. & SHERLOCK G. (2000). Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1), 25-29.

BA M. & BOSSY R. (2016). Interoperability of corpus processing workflow engines: the case of. AlvisNLP/ML in OpenMinTeD. In *Meeting of working Group Medicago sativa* (p. np).

BOSSY R., CHAIX E., DELEGER L., FERRE A., BA M., BESSIERES P. & NEDELLEC C. (2016). OntoBiotope: une ontologie pour croiser les habitats microbiens avec les analyses de génomes. In *Les journées Bioinformatique de l'Inra* (p. 1).

BOSSY R., JOURDE J., MANINE A.-P., VEBER P., ALPHONSE E., VAN DE GUCHTE M. & NEDELLEC C. (2012). Bionlp shared task-the bacteria track. In *BMC bioinformatics* (Vol. 13, p. S3). *BioMed Central*.

COHEN K. B. & DEMNER-FUSHMAN D. (2014). Biomedical natural language processing. *John Benjamins*.

CRAVEN M., KUMLIEN J., et al. (1999). Constructing biological knowledge bases by extracting information from text sources. In ISMB (Vol. 1999, p. 77–86).

DELEGER L., BOSSY R., CHAIX E., BA M., FERRE A., BESSIERES P. & NEDELLEC C. (2016). Overview of the Bacteria Biotope task at BioNLP Shared Task 2016. In *Proceedings of the 4th BioNLP Shared Task Workshop* (p. 12–22).

FARUQUI M., DODGE J., JAUHAR S. K., DYER C., HOVY E. & SMITH N. A. (2014). Retrofitting word vectors to semantic lexicons. arXiv preprint arXiv:1411.4166.

FARUQUI M., TSVETKOV Y., RASTOGI P. & DYER C. (2016). Problems With Evaluation of Word Embeddings Using Word Similarity Tasks. arXiv:1605.02276 [cs].

FAURE D. & NEDELLEC C. (1998). A Corpus-based Conceptual Clustering Method for Verb Frames and Ontology Acquisition. In *LREC workshop* on (p. 5–12).

FERRE A., ZWEIGENBAUM P. & NEDELLEC C. (2017). Representation of complex terms in a vector space structured by an ontology for a normalization task. BioNLP 2017, 99–106.

GOLIK W., WARNIER P. & NÉDELLEC C. (2011). Corpus-based extension of termino-ontology by linguistic analysis: a use case in biomedical event extraction. In *WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence* (p. 37–39).

HWANG C. H. (1999). Incompletely and Imprecisely Speaking : Using Dynamic Ontologies for Representing and Retrieving Information, 13.

KANG N., SINGH B., AFZAL Z., VAN MULLIGEN E. M. & KORS J. A. (2013). Using rule-based natural language processing to improve disease normalization in biomedical text. *Journal of the American Medical Informatics Association*, 20(5), 876-881.

LAROCHELLE H., ERHAN D. & BENGIO Y. (2008). Zero-data learning of new tasks. In AAAI (Vol. 1, p. 3).

LEAMAN R., ISLAMAJ DOGAN R. & LU Z. (2013). DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22), 2909-2917.

LIMSOPATHAM N. & COLLIER N. (2016). Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (p. 1014-1023).

LIPSCOMB C. E. (2000). Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association*, 88(3), 265-266.

MARSI E. (2014). Towards Text Mining in Climate Science: Extraction of Quantitative Variables and their Relations, 8.

MCCRAY A. T. (1989). The UMLS Semantic Network. In Proceedings/the... Annual Symposium on Computer Application [sic] in Medical Care. Symposium on Computer Applications in Medical Care (p. 503–507). *American Medical Informatics Association*.

MORGAN A. A., LU Z., WANG X., COHEN A. M., FLUCK J., RUCH P. & HIRSCHMAN L. (2008). Overview of BioCreative II gene normalization. *Genome Biology*, 9(Suppl 2), S3.

MRKSIC N., SEAGHDHA D. Ó., THOMSON B., GASIC M., ROJAS-BARAHONA L., SU P.-H. & YOUNG S. (2016). Counter-fitting Word Vectors to Linguistic Constraints. arXiv:1603.00892 [cs].

RUSSELL S. J. & NORVIG P. (2016). Artificial intelligence: a modern approach. Malaysia; *Pearson Education Limited*.

SIL A., KUNDU G., FLORIAN R. & HAMZA W. (2018). Neural cross-lingual entity linking. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

USCHOLD M. & KING M. (1995). Towards a Methodology for Building Ontologies, 15.

WEEBER M., KORS J. A. & MONS B. (2005). Online tools to support literature-based discovery in the life sciences. *Briefings in Bioinformatics*, 6(3), 277-286.

XIAN Y., LAMPERT C. H., SCHIELE B. & AKATA Z. (2017). Zero-Shot Learning - A Comprehensive Evaluation of the Good, the Bad and the Ugly. arXiv:1707.00600 [cs].

YU Z., C. WALLACE B., JOHNSON T. & COHEN T. (2017). Retrofitting Concept Vector Representations of Medical Concepts to Improve Estimates of Semantic Similarity and Relatedness.