Utilisation des ontologies et des bases de connaissances dans la désambiguïsation des entités - Sélection d'articles

BERNIER - ERABLE - LOPEZ - FUMAROLI

9 septembre 2020

Contents

1 Description

Contexte scientifique Ce projet de recherche se développera au sein des projets européens NewsEye et EMBEDDIA qui sont financés par la commission européenne dans le cadre du programme Horizon 2020, en collaboration avec d'autres universités comme celle de Montpellier ou des bibliothèques nationales. L'équipe sera incluse dans la partie "Images et contenus" dans le thématique "Documents et Contenus Numériques" au sein du Laboratoire L3i. L'objectif de ce projet est de permettre l'amélioration de systèmes d'annotation sémantique du laboratoire L3i dans les projets NewsEye et EMBEDDIA.

Problématique scientifique L'objectif du projet est de pouvoir mettre à disposition des articles anciens pour tous, et de pouvoir les mettres en relation entre eux. La difficulté du projet réside dans la qualité des ouvrages qu'il faudra numérisé.

2 Articles sélectionnés

Voici la liste de mots clés utilisés dans la recherche d'article : ontology, knowledge bases, disambiguation, semantic annotation, candidate identification, named entities, Wikipedia, populate ranking, linking, Semi-Structured Resources, Extraction, Record Linkage

2.1 Large-Scale Named Entity Disambiguation Based on Wikipedia Data[12]

Résumé Ce document sous forme de compte rendu met en évidence les phases de tests réalisées pour l'ontologie de mots sous wikipédia. L'objectif étant de démontrer sa fiabilité selon différentes méthodes de désambiguïsation. Ainsi, de nouvelles pistes s'ouvrent et démontrent également que ce système est adaptable à d'autres systèmes.

Description de la conférence Cette conférence a eu lieu à Pragues en juin 2007 lors de la Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.

Description des auteurs L'auteur de ce document, Silviu Cucerzan, est le directeur de recherche aux laboratoires Microsoft FUZE. Il cumule 17 brevets à son actif et de nombreuses distinctions ou prix comme le Microsoft Hackathon 2016 Grand Prize Winner.

2.2 Ontology-Based Information Extraction from Twitter[14]

Résumé Le document met en avant la quantité de messages disponibles sur twitter et le processus de détection des entités nommées pour les relier à une

base de connaissances via une désambiguïsation basée sur un score de popularité et la syntaxe similaire des mots.

Description de la conférence Compte rendu d'une réunion tenue le 9 décembre 2012, Mumbai, Inde. Tenue lors de la 24e Conférence internationale sur la linguistique informatique

Description des auteurs Kamel Nebhi est l'auteur de ce compte rendu travaillant au Laboratoire d'Analyses et de Technologie du Langage (LATL), département linguistique, à l'université de Genève en Suisse.

2.3 Towards ontology-based disambiguation of geographical identifiers[15]

Résumé Ce document met en lumière l'utilisation d'une nouvelle méthode désambiguïsation par système de rang à partir de mots géographiques basés sur une ou plusieurs ontologies. Cette nouvelle méthode pourrait être utilisée dans l'utilisation de mots géographiques sur le web et permettre de proposer à l'utilisateur une meilleure expérience.

Description de la revue Ce document a été publié dans le cadre d'une étude de recherche d'informations sur le thème de l'ontologie.

Description des auteurs Les auteurs de ce document font parties de l'institut AIFB à l'Université de Karlsruhe et au FZI Research Center for Information Technologies en Allemagne.

2.4 Evaluating Entity Linking with Wikipedia[13]

Résumé Cette revue fait état des méthodes de désambiguïsation existant pour la recherche d'identité nommée, effectue une comparaison entre elle-même pour permettre de différencier leurs forces et faiblesses ainsi que leurs efficacitées.

Description de la revue Cette revue fait partie d'un livre contenant de nombreuses revues sur l'intelligence artificielle, Wikipédia et les "semi-structures ressources".

Description des auteurs Ben Hachey est un chercheur chez Thomson Reuters Corporation aux Etats-Unis, Will Radford, James R. Curran et Joel Nothman sont des chercheurs à l'école d'information technologique de Sydney en Australie et Matthew Honnibal qui travaillent au département d'informatique à l'université Macquarie en Australie

2.5 Learning to Link Entities with Knowledge Base[5]

Résumé Ce document fait état d'expérimentation de différents algorithmes sur les méthodes de désambiguïsation existantes pour permettre de relier des entités trouvées dans un document ou un texte à une base de connaissances déjà existante selon les correspondances dans le but de faire des liens vers d'autres entités. Ainsi, un retour de ces expérimentations est présent dans cet article.

Description de la conférence Cet article est extrait de la conférence "Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL" qui a eu lieu en juin 2010 à Los Angeles en Californie. Cette conférence était composée de nombreux événements et dont de nombreux articles étaient regroupés dont celui-ci est extrait. Cette conférence est organisée par Association for Computational Linguistics (ACL), une société basée notamment sur le traitement automatique des langues.

Description des auteurs Zhicheng Zheng, Fangtao Li, Minlie Huang et Xiaoyan Zhu sont les auteurs de cet article travaillant dans le laboratoire Tsinghua national des sciences et technologies de l'information, lié au département informatique et de technologie de l'Université Tsinghua situé à Beijing en Chine.

2.6 Entity Disambiguation for Knowledge Base Population[4]

Résumé Cet article permet de mettre en évidence un moyen d'enlever les ambiguïtés des différentes entités pour permettre une intégration aux bases de connaissances déjà présentes. Ainsi, un système sur la désambiguïsation d'entité est présenté. De plus, une notion de gestion des ressources est présente pour un grand nombre d'entrée dans ce système.

Description de la conférence Ce document est extrait des articles de la conférence "Proceedings of the 23rd International Conference on Computational Linguistics" (Coling 2010). Cette conférence a eu lieu à Pékin du 23 au 27 aout 2010, gérée par "the International Committee on Computational Linguistics" (ICCL). La première conférence a eu lieu en 1965 et elle accueille de plus en plus de monde chaque année.

Description des auteurs Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber et Tim Finin sont les auteurs de cet article. Il a été écrit en collaboration entre deux universités. Mark Dredze, Paul McNamee, Delip Rao et Adam Gerber travaillent au Centre d'excellence en technologie du langage humain et Centre de traitement du langage et de la parole à l'Université Johns Hopkins,une université privée américaine. Tim Finin travaille lui à l'université du Maryland dans le comté de Baltimore.

2.7 Ontology-Driven Automatic Entity Disambiguation in Unstructured Text[1]

Résumé

Description de la conférence Cet article est extrait de la conférence "The 5th International Semantic Web Conference"(ISWC) qui a eu lieu à Athènes aux États-Unis du 5 au 9 novembre 2006. Le livre de la conférence contient plus de 52 articles dans le domaine du "Semantic Web".

Description des auteurs Les auteurs de ce document sont Joseph Hassell, Boanerges Aleman-Meza et I. Budak Arpinar appartenant au Laboratoire Large Scale Distributed Information Systems (LSDIS) dans le Département informatique à Université de Géorgie aux Etats-Unis.

2.8 Ontology-Driven Automatic Entity Disambiguation in Unstructured Text[2]

Résumé

Description de la conférence La conférence, dont cet article a été extrait est la conférence suivante: "Proceedings of the 18th ACM Conference on Information and Knowledge Management" qui s'est déroulée à Hong Kong en Chine entre le 2 et 6 novembre 2009. Elle a été organisée par "Association for Computing Machinery" (ACM), une association à but non lucratif dont le souhait est de soutenir la recherche informatique.

Description des auteurs Les auteurs de ce document Xianpei Han et Jun Zhao sont deux chercheurs chinois de l'Institut d'automatisation de l'Académie chinoise des sciences qui se situe dans le district de HaiDian à Pékin en Chine. Ces deux auteurs ont également rédigé d'autres articles dans ce domaine dont un qu'on retrouvera dans cette bibliographie également.

2.9 Entity Linking with Effective Acronym Expansion, Instance Selection and Topic Modeling[7]

Résumé Cet article parle de comment détecter un acronyme, pour ensuite le lier à son entité. Le document nous montre également une solution pour automatiser le processus de détection de l'acronyme mais aussi une méthode pour rendre l'acronyme le moins ambigu possible.

Description de la conférence / revue Ce document est paru lors de la conférence: IJCAI (International Joint Conference on Artificial Intelligence) du 16 au 22 Juillet 2011 à Barcelone en Espagne. Dû à la sélection très sélective de

ces articles, elle est considérée comme l'une des conférences les plus prestigieuses à propos L'IA.

Description des auteurs Les auteurs sont : les professeurs Wei Zhang, Chew Lim Tan de l'École d'informatique de l'Université nationale de Singapour et Yan Chuan Sim, Jian Su de l'Institut de Recherche en Infocomm.

2.10 NLPR_KBP in TAC 2009 KBP Track: A Two-Stage Method to Entity Linking[3]

Résumé Cet article parle ici d'une méthode pour trouver la meilleure correspondance à une entité. Il utilise pour cela un processus en deux étapes qu'ils décrivent, la première étant la détection des candidats possibles de l'entité dans une base de connaissances, et la seconde étape est un outil pour lier une entité aux candidats les plus pertinents.

Description de la conférence / revue Ce papier vient de la conférence TAC (Text Analysis Conference) de 2009. La TAC a pour but de pousser les chercheurs à travailler sur le domaine du traitement du langage.

Description des auteurs Les auteurs sont : Le professeur Han Xianpei et le chercheur Jun Zhao du Laboratoire national de reconnaissance des formes à l'Institut d'automatisation de l'Académie chinoise des sciences à Beijing en Chine.

2.11 Utilisation des relations d'une base de connaissances pour la désambiguïsation d'entités nommées[10]

Résumé Cet acte de conférence nous montre une solution pour permettre la désambiguïsation d'une entité à partir des liens de celle-ci dans une base de connaissances. Ils basent leurs tests sur plusieurs bases de connaissances pour avoir des résultats pertinents.

Description de la conférence / revue Cet acte provient du volume 2 : TALN (Traitement Automatique des Langues Naturelles) de la conférence conjointe JEP-TALN-RECITAL (Journées d'Études sur la Parole Traitement Automatique des Langues Naturelles Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues) qui a eu lieu en 2016.

Description des auteurs Les auteurs sont : Romaric Besançon, Hani Daher, Olivier Ferret et Hervé Le Borgne, ils sont chercheurs au Laboratoire Vision et Ingénierie des Contenus appartenant au Commissariat à l'énergie atomique et aux énergies alternatives dans la section LIST pour les outils digitaux.

2.12 TRank: Ranking Entity Types Using the Web of Data[11]

Résumé Ce document nous explique les travaux réalisés de comparaisons de différentes méthodes, pour classer les types d'une entité selon leurs pertinences, en prenant en compte son contexte (au sein d'un document, d'une phrase ou autres) et de ses différents types.

Description de la conférence / revue Issue de la 12ème Conférence internationale sur le web sémantique (ISWC), partie 1, qui s'est déroulée à Sydney en Australie du 21 au 25 octobre en 2013.

Description des auteurs Les auteurs sont : Alberto Tonon, Gianluca Demartini, Philippe Cudré-Mauroux provenant du laboratoire eXascale Infolab de l'université de Fribourg en Suisse ainsi que Michele Catasta et Karl Aberer venant de l'école Polytechnique Fédérale de Lausanne en Suisse également.

2.13 Exploring Entity Relations for Named Entity Disambiguation [8]

Résumé Ce travail cherche à présenter différentes méthodes de désambiguation de entités et d'annotation sémantique dérivées de la structure de graph de Wikipédia. De plus la problématique principale cherche d'une part à améliorer la désambiguation des entités avec la base de connaissance donnée mais en plus mettre en relation correctement une entité nommée qui ne se trouve pas dans la base de connaissance.

Description de la conférence / revue Ce document, il a été présenté lors de la conférence dénommée Proceedings of the ACL en 2011. ACL est l'acronyme pour Association for Computational Linguistics, l'organisateur de cette conference. Les conférences de ACL sont parmies les plus prestigieuses. Leur thématique principal est la résolution des problèmes liés à l'étude du langage d'un point de vue informatique. D'ailleurs on peut constater l'importance de cette conférence avec le classement CORE, avec une note 'A*'. Cette 49eme conférence était centrée sur les technologies du langage humain. Chaque année propose une conférence et la dernière a eu lieu a le 4 juillet 2020.

Description des auteurs L'auteur de ce document est Danuta PLOCH, associé scientifique au laboratoire 'Dai-Labor', annexe à l'université technique de Berlin.

2.14 Entity Disambiguation with Hierarchical Topic Models [6]

Résumé L'intelligence artificielle est une technique utilisée pour la désambiguation des entités. Dû à cela, ce document présente l'utilisation de la méthode appelée 'Wikipedia-based Pachinko Allocation' (WPAM) qui servira dans un premier temps pour donner un thème / sujet à chaque entité en question. La base de connaissance de cette méthode d'apprentissage semi-supervisé est Wikipédia. Tout au long de ce document on pourra voir les différentes problématiques lors de la désambiguation que cette technique présente et comment ce groupe de chercheurs on réussit à surmonter cela.

Description de la conférence / revue Ce document, il a été présenté lors de la conférence dénommée ACM International Conference on Knowledge Discovery and Data Mining (KDD) en 2011. Comme son nom l'indique, les conférences organisées par la KDD sont principalement dans la thématique de la fouille de données. Cette conférence fait partie des plus renommée. Elle a obtenu la note de 'A*' dans le classement de conférences internationales CORE. La conférence 2011 où ce document a été publié était la numéro 17. La KDD organise une conférence par an, généralement dans le mois d'aout. La prochaine aura lieu du 14 au 18 aout 2021.

Description des auteurs Les 5 auteurs qu'ont participé à la création de ce document sont : Saurabh S. Kataria doctorat de l'Université d'État de Pennsylvanie en 2011 et actuellement chercheur dans la division du centre de recherche Palo Alto. Parmi ces plusieurs sujets de recherches, l'extraction d'information de texte est toujours d'actualité. Krishnan S. Kumar, Rajeev Rastogi, Prithviraj Sen, Srinivasan H Sengamedu : groupe de chercheurs dans l'entreprise 'Yahoo! Labs'

2.15 To Link or Not to Link? A Study on End-to-End Tweet Entity Linking [9]

Résumé De nos jours les réseaux sociaux comme twitter ou Facebook créent énormément de 'data'. Par exemple twitter,seul,produit plus de 340 millions de tweets par jour. Ce document nous présente quels sont les différentes problématiques des micro-textes (comme les tweets) comme au moment de localiser les mentions et désambiguiser des entités. De plus, elle expose des techniques pour résoudre les problématiques énoncées et quelles sont les performances de ces expérimentations.

Description de la conférence / revue Ce document, il a été présenté lors de la conférence dénommée Proceedings of the ACL en 2013. ACL est l'acronyme pour Association for Computational Linguistics, l'organisateur de cette conférence. Les conférences de ACL sont parmies les plus prestigieuses.

Leur thématique principal est la résolution des problèmes liés à l'étude du langage d'un point de vue informatique. D'ailleurs on peut constater l'importance de cette conférence avec le classement CORE, avec une note 'A*'.

Description des auteurs Les 3 auteurs qu'ont participé à la création de ce document sont : Stephen Guo Doctorat chez Standford university. Ming-Wei Chang, Emre Kıcıman chercheur chez Microsoft.

3 Conclusion

La relation entre chaque document permet de mettre en évidence la complexité dans la recherche d'un algorithme de désambiguïation de part sa fiabilité, son efficacité mais aussi sa capacité. Ces articles permettent d'approcher chacun des algorithmes déjà existant pour pouvoir les tester et les comparer, et de possiblement en créer un nouveau. La diversification de ces articles permettent de voir plusieurs utilisatons d'algorithmes dans la désambiguïsation, que ce soit pour une utilisation purement Web comme l'identification de mots nommés sur Twitter ou Wikipédia, mais également pour des utilisations géographiques notemment. Ces documents permettent d'orienter nos recherches en comparant les systèmes existants et en determinant l'algorithme le plus adaptable à notre projet.

References

[1] *Ontology-Driven Automatic Entity Disambiguation in Unstructured Text*, volume 4273, 11 2006.

[2] *Named entity disambiguation by leveraging wikipedia semantic knowledge*, 01 2009.

[3] *NLPR_KBP in TAC 2009 KBP Track: A Two-Stage Method to Entity Linking*, 2009.

[4] *Entity Disambiguation for Knowledge Base Population*, Beijing, China, August 2010. Coling 2010 Organizing Committee.

[5] *Learning to Link Entities with Knowledge Base*, Los Angeles, California, June 2010. Association for Computational Linguistics.

[6] *Entity disambiguation with hierarchical topic models*. ACM, 2011.

[7] *Entity Linking with Effective Acronym Expansion, Instance Selection and Topic Modeling*, 01 2011.

[8] *Exploring Entity Relations for Named Entity Disambiguation*, Portland, OR, USA, June 2011. Association for Computational Linguistics.

[9] *To Link or Not to Link? A Study on End-to-End Tweet Entity Linking*, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[10] *Utilisation des relations d'une base de connaissances pour la désambiguïsation d'entités nommées (Using the Relations of a Knowledge Base to Improve Entity Linking)*, Paris, France, 7 2016. AFCP - ATALA.

[11] Harith Alani, Lalana Kagal, Achille Fokoue, Paul Groth, Chris Biemann, Josiane Xavier Parreira, Lora Aroyo, Natasha Noy, Chris Welty, and Krzysztof Janowicz, editors. *TRank: Ranking Entity Types Using the Web of Data*, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[12] Jason Eisner, editor. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[13] Ben Hacheya, Will Radfordb, Joel Nothmanb, Matthew Honnibald, and James R. Curranb. Evaluating entity linking with wikipedia. Banff, Canada, April 2012.

[14] Sriram Raghavan and Ganesh Ramakrishnan, editors. *Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data*, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

[15] Raphael Volz, Joachim Kleb, and Wolfgang Mueller. Towards ontology-based disambiguation of geographical identifiers. Banff, Canada, August 2007.