Enhancing Domain-Specific Entity Linking in DH

Federico Nanni
federico@informatik.uni-mannheim.de
University of Mannheim, Germany

Yang Zhao
yzhao@mail.uni-mannheim.de
University of Mannheim, Germany

Simone Paolo Ponzetto
simone@informatik.uni-mannheim.de
University of Mannheim, Germany

Laura Dietz
dietz@cs.unh.edu
University of New Hampshire, United States of America

For the purpose of information retrieval and text exploration, digital humanities (DH) scholars have examined the potential of methods such as keyphrase extraction (Hasan and Ng, 2014) and named entity recognition (Nadeau and Sekine, 2007). However, these solutions still face challenges in the presence of polysemy and synonymy (e.g. distinguish between "Paris" the capital of France or the city in Ontario or recognize that "POTUS" and "Barack Obama" might refer to the same person).

Entity Linking

In the last decade, advances in natural language processing (NLP) gave rise to word-sense disambiguation and entity linking techniques (Cornolti et al., 2013), which automatically disambiguate entities and concepts in context and link them to knowledge bases such as Wikipedia, DBpedia (Auer et al., 2007) or Babelnet (Navigli and Ponzetto, 2012). Among them, TagMe (Ferragina and Scaiella, 2010) has been often adopted in NLP, thanks to its decent performance on different datasets and to its easy-to-use API.

Current Limitations for DH research

TagMe also highlights a few common limitations of current standard entity linking systems that reduce their applicability within most scenarios found in the heterogeneous spectrum of Digital Humanities research.

- **Black Box and reproducibility.** As Hasibi et al. (2016) recently remarked, the TagMe RESTful API remains a black box, as it is impossible to check whether it corresponds to the system described in the original paper. Not knowing the reliability of the system limits its use for distant reading analyses, i.e. quantitative studies that go beyond text exploration.
- **Language Versions.** Currently, TagMe is only available in English, German and Italian but does not support other widespread languages such as Chinese, Arabic, Spanish, and French, which are essential for enhancing its use in the DH community.
- **Infrequent Updates.** TagMe has been initially created on the English 2009 version of Wikipedia and it has been updated only twice (summer 2012, summer 2016). Imagine a setting where a scholar intends to analyze a collection of mainstream news on the Middle East: before the most recent update the system was not able to detect mentions of "Al-Nursa Front", the former Syrian branch of al-Qaeda.
- **Wikipedia as Knowledge Base.** TagMe, as well as other entity-linking solutions, relies on the assumption that the entries and structure of Wikipedia provide us with a comprehensive and accurate knowledge base. While this is mostly true for standard NLP and IR approaches, when it comes to humanities research this assumption shows all its limitations. As a matter of fact, linking to Wikipedia is not ideal for example when dealing with historical documents, simply because entities and concepts relevant in the corpus may be missing from such a general-purpose knowledge resource (as remarked in Lauscher et al., 2016).

Specific contribution

While we are currently working on the implementation and optimization of a domain-adaptable entity linking pipeline, at the conference we intend to present a solution for generating, in an automatic fashion, domain-specific knowledge bases from an user-created Wiki. As the creation of a complete Wiki is too time-consuming, these domain-specific wikis are used

in combination with general world knowledge available on Wikipedia. In particular, we will describe how our system can make use of the following input:

The XML Dump of any language version of Wikipedia and rapidly create the indexes that compose the knowledge base. This permits to have a knowledge base for each language version of Wikipedia and to update it on the spot whenever needed.

Any MediaWiki website dump, such as Wikia (although it is important to consider the copyright license when downloading and using this data), to be merged into the same index. In the table we report a few examples from different Wikia sites. It is important

This solution gives the scholar the possibility of creating (or improving an already existent) domain specific Wikia (a practice common in DH education, see Farabaugh, 2007 and Giglio & Venecek, 2009) on the topic she/he intends to study and identifying mentions of domain-specific and general-purpose concepts in large text collections.

Inlinks (From Italian Wikipedia)	Entities and their mentions (From Star Trek Wikia)	Entities in Text (From Harry Potter Wikia)
Università di Bologna: "Emilia Romagna", "Umberto Eco", "Ulisse Aldrovandi", etc. Romano Prodi: "Massimo d'Alema", "Silvio Berlusconi", "Unione Europea", etc. Anniversario della liberazione d'Italia: "Seconda guerra mondiale", "Resistenza italiana", "Repubblica Italiana", etc.	James T. Kirk: "James Kirk","Kirk","James T. Kirk","James Tiberius Kirk", "James", "Admiral Kirk", etc. Jean-Luc Picard: "Picard","Jean-Luc Picard","Jean-Luc","Cpt. Picard", "Captain Jean-Luc Picard", etc. William T. Riker: "William Riker", "William T. Riker","Riker","Will Riker","Thomas Riker", "William Thomas Riker", etc.	"Sirius was sent to Azkaban, and after twelve years became the only known person to escape the prison unassisted" "After leaving school, Charlie went to Romania to study dragons." "Meanwhile, it was not long before Hermione realised that the Ministry of Magic had decided to interfere at Hogwarts."

Bibliography

Auer, S., et al. (2007) "Dbpedia: A nucleus for a web of open data." *The semantic web*. Springer Berlin Heidelberg, 2007. 722-735.

Cornolti, M., Ferragina, P., and Ciaramita, M. (2013) "A framework for benchmarking entity-annotation systems." *Proceedings of the 22nd international conference on World Wide Web*. ACM.

Farabaugh, R. (2007) "'The isle is full of noises': Using wiki software to establish a discourse community in a Shakespeare classroom." *Language Awareness* 16.1: 41-56.

Ferragina, P., and Scaiella, U. (2010) "Tagme: on-the-fly annotation of short text fragments (by wikipedia entities)." *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM.

Giglio, K., and Venecek, J. (2009) "The Radical Historicity of Everything: Exploring Shakespearean Identity with Web 2.0." *Digital Humanities Quarterly* 3.3.

Hasan, K. S., and Ng, V. (2014) "Automatic Keyphrase Extraction: A Survey of the State of the Art." *ACL (1)*.

Hasibi, F., Balog, K, and Bratsberg, S. E. (2016) "On the Reproducibility of the TAGME Entity Linking System." *European Conference on Information Retrieval*. Springer International Publishing, 2016.

Lauscher, A., et al. (2016) "Entities as topic labels: combining entity linking and labeled LDA to improve topic interpretability and evaluability." IJCol-Italian journal of computational linguistics 2.2 (2016): 67-88.

Nadeau, D., and Satoshi S. (2007) "A survey of named entity recognition and classification." *Lingvisticae Investigationes* 30.1: 3-26.

Navigli, R., and Ponzetto, S. P. (2012) "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network." *Artificial Intelligence* 193 (2012): 217-250.