Information Extraction with Humans in the Loop

Anna Lisa Gentile
annalisa.gentile@ibm.com
IBM Research Almaden, CA, US

ABSTRACT

Information Extraction (IE) techniques enables us to distill Knowledge from the abundantly available unstructured content. Some of the basic IE methods include the automatic extraction of relevant entities from text (e.g. places, dates, people, ...), understanding relations among them, building semantic resources (dictionaries, ontologies) to inform the extraction tasks, connecting extraction results to standard classification resources. IE techniques cannot decouple from human input - at bare minimum some of the data needs to be manually annotated by a human so that automatic methods can learn patterns to recognize certain type of information. The human-in-the-loop paradigm applied to IE techniques focuses on how to better take advantage of human annotations (the recorded observations), how much interaction with the human is needed for each specific extraction task.

ACM Reference Format:
Anna Lisa Gentile. 2019. Information Extraction with Humans in the Loop. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3308560.3316478

TALK OUTLINE

In this talk I will describe various experiments of the human-in-the-loop model on various IE tasks, such as (i) building dictionaries from text corpora in various languages [1]; (ii) extracting mentions of adverse drug reaction from text and matching them to a reference ontology [2]; (iii) relation extractions, e.g. automatically identifying from the text which drug is causing which adverse drug reaction [3].

SPEAKER'S BIO

Dr Anna Lisa Gentile (https://w3id.org/people/annalisa) is a Research Staff Member at IBM Research Almaden. Her research is principally focused on studying methods and techniques for semantic annotating unstructured and semi-structured content. Her main Research Areas are Information Extraction (IE), Natural Language Processing (NLP) and Semantic Web. She obtained her PhD with a thesis on Named Entity Disambiguation at the University of Bari, Italy in 2010. She has published more than 50 peer-reviewed scientific publications including papers at major venues such as LREC, EMNLP, ESWC and ISWC. She has been serving as Organizing Committee member for conferences such as ISWC, ESWC, WWW amongst many others and organized workshop series such as *LD4IE* on Linked Data for Information Extraction (http://w3id.org/ld4ie) and *HumBL* on Augmenting Intelligence with Bias-Aware Humans-in-the-Loop (http://w3id.org/huml).

REFERENCES

[1] Alfredo Alba, Anni Coden, Anna Lisa Gentile, Daniel Gruhl, Petar Ristoski, and Steve Welch. 2017. Multi-lingual Concept Extraction with Linked Data and Human-in-the-Loop. In *K-CAP 2017*, Óscar Corcho, Krzysztof Janowicz, Giuseppe Rizzo, Ilaria Tiddi, and Daniel Garijo (Eds.). ACM, 24:1–24:8. https://doi.org/10.1145/3148011.3148021

[2] Kenneth Clarkson, Anna Lisa Gentile, Daniel Gruhl, Petar Ristoski, Joseph Terdiman, and Steve Welch. 2018. User-Centric Ontology Population. In *ESWC 2018 (Lecture Notes in Computer Science)*, Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam (Eds.), Vol. 10843. Springer, 112–127. https://doi.org/10.1007/978-3-319-93417-4_8

[3] Ismini Lourentzou, Alfredo Alba, Anni Coden, Anna Lisa Gentile, Daniel Gruhl, and Steve Welch. 2018. Mining Relations from Unstructured Content. In *PAKDD 2018*. 363–375. https://doi.org/10.1007/978-3-319-93037-4_29

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316478