University of Fribourg / Faculty of Science and Medicine / Department of Informatics

Neural Machine Reading for Domain-Specific Text Resources

Sebastian Arnold

The vision of Machine Reading is to automatically understand written text and transform the contained information into machine-readable representations. This thesis approaches this challenge in particular in the context of commercial organizations. Here, an abundance of domain-specific knowledge is frequently stored in unstructured text resources. Existing meth- ods often fail in this scenario, because they cannot handle heterogeneous document structure, idiosyncratic language, spelling variations and noise. Specialized methods can hardly over- come these issues and often suffer from recall loss. Moreover, they are expensive to develop and often require large amounts of task-specific labeled training examples.

Our goal is to support the human information-seeking process with generalized language understanding methods. These methods need to eliminate expensive adaptation steps and must provide high error tolerance. Our central research question focuses on capturing domain- specific information from multiple levels of abstraction, such as named entities, latent topics, long-range discourse trajectory and document structure. We address this problem in three central information-seeking tasks: Named Entity Linking, Topic Modeling and Answer Pas- sage Retrieval. We propose a collection of Neural Machine Reading models for these tasks. Our models are based on the paradigm of artificial neural networks and utilize deep recurrent architectures and end-to-end sequence learning methods.

We show that automatic language understanding requires a contextualized document rep-resentation that embeds the semantics and skeletal structure of a text. We further identify key components that allow for robust word representations and efficient learning from sparse data. We conduct large-scale experiments in English and German language to show that Neural Ma- chine Reading can adapt with high accuracy to various vertical domains, such as geopolitics, automotive, clinical healthcare and biomedicine. This thesis is the first comprehensive research approach to extend distributed language models with complementary structure information from long-range document discourse. It closes the gap between symbolic Information Extrac- tion and Information Retrieval by transforming both problems into continuous vector space representations and solving them jointly using probabilistic methods. Our models can fulfill task-specific information needs on large domain-specific text resources with low latency. This opens up possibilities for interactive applications to further evolve Machine Reading with hu- man feedback.

Jury:
Prof. Dr. Philippe Cudré-Mauroux, University of Fribourg (Switzerland), thesis supervisor.
Prof. Dr. Alexander Löser, Beuth University of Applied Sciences (Germany), thesis supervisor.
Prof. Dr Laura Dietz, University of New Hampshire (USA), external examiner.
Prof. Dr. Ulrich Ultes-Nitsche, University of Fribourg (Switzerland), president of the jury.