Fraunhofer Series in
Information and Communication Technology

Band 1/2011

Roman Klinger

Conditional Random Fields for Named Entity Recognition

Feature Selection and Optimization in Biology and Chemistry

D 290 (Diss. Technische Universität Dortmund)

Shaker Verlag
Aachen 2011

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Zugl.: Dortmund, Technische Univ., Diss., 2011

Printed in Germany.

ISBN 978-3-8440-0213-3
ISSN 1612-4863

Shaker Verlag GmbH • P.O. BOX 101818 • D-52018 Aachen
Phone: 0049/2407/9596-0 • Telefax: 0049/2407/9596-9
Internet: www.shaker.de • e-mail: info@shaker.de

Roman Klinger, ***Conditional Random Fields for Named Entity Recognition: Feature Selection and Optimization in Biology and Chemistry***. Shaker Verlag, 2011, ISBN 978-3-8440-0213-3.

Most knowledge is stored and communicated in the form of natural language text. Databases including abstracts of journal articles or proceeding contributions are freely available. To make this knowledge available in a structured form, allowing for deeper analysis and combination with existing databases, technologies from the field of information extraction are necessary. A fundament for most methods like relation extraction or semantic search is named entity recognition. Conditional random fields are an established probabilistic method for labeling sequences. Nevertheless, the adaption to novel domains or entity classes of interest requires manual effort.

This dissertation presents such adaptions for entity classes from the biological and chemical domain. Workflows for the detection of gene and protein names, mentions of mutations of genes, and chemical names following the nomenclature of the International Union of Pure and Applied Chemistry. For these classes, training corpora are discussed and built. Questions addressed include how to use knowledge from multiple annotators, how stable a model is on data from different time ranges, or how to normalize found entities.

The presented use cases exemplify the need for feature design and selection. Different methods for choosing a meaningful feature subset decreasing the run time and number of features clearly are developed and evaluated. To extend the applicability of conditional random fields, a training method based on multicriterial optimization is introduced allowing the user to choose between different precision-recall weightings without increase of runtime. Additionally, it is analysed if automatically selected structures going beyond the common linear structure of conditional random fields can be beneficial for named entity recognition.

These methods and analyses support the generation of workflows to build novel named entity recognition tools with less user intervention.