Learning Information Extraction Rules with Sequential Patterns: a Quantitative and Qualitative Evaluation

Laura Handojo

Erasmus Mundus Master Program
in Language and Communication Technologies

June 18, 2012

Université de Lorraine
UFR - Mathématiques et Informatique
Supervisor: Yannick Toussaint
Academic year: 2011-2012

Rijksuniversiteit Groningen
Faculty of Arts
Supervisor: Dr. Gosse Bouma

Abstract

This thesis describes an automatic approach to learning information extraction rules, or more precisely relation extraction rules. It uses sequential pattern mining to learn patterns from research abstracts on a rare disease called *fibromuscular dysplasia*. Information extraction can be useful in this field for creating review articles on rare diseases, thereby making information on them more accessible.

Preprocessing includes lemmatization, POS tagging and semantic tagging. Different settings with regard to the semantic tags as well as what to use as input sequences are tried out. The patterns are postprocessed in order to obtain those which relate entities in a treatment-relation. After a manual analysis, the patterns are filtered and gap constraints are implemented to reduce noise. Evaluated on a test set, the best F_1-measure reaches 50%.