Journal of Automata, Languages and Combinatorics **26** (2021) 3–4, 281 – 332

F-TRANSDUCERS FOR CONTEXTUAL TEXT REWRITING

Stoyan Mihov[(A)] Klaus U. Schulz[(B)]

[(A)] *Institute of Information and Communication Technologies*
Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Block 2, 1113 Sofia, Bulgaria
stoyan@lml.bas.bg

[(B)] *Centrum für Informations- und Sprachverarbeitung (CIS)*
Ludwig-Maximilians-Universität München
Oettingenstr. 67, 80538 München, Germany
schulz@cis.uni-muenchen.de

ABSTRACT

Rule-based text rewriting is a form of language processing where a given input text/string is rewritten to a new output form using some form of knowledge base plus contextual rules. Contextual rewrite dictionaries, to be defined below, represent a general input format for such knowledge bases with contextual rules. We present an algorithm for translating a given contextual rewrite dictionary into an f-transducer. This transducer can be applied to input texts, realizing the intended form of text rewriting. F-transducers are deterministic transducers that use failure transitions in order to reduce the size. A second algorithm is given for composing f-transducers. Using composition, cascaded forms of text rewriting can be realized in a single run over the input text. Put together, a general system for rule-based text rewriting is obtained. The strength of the approach relies on four points. First, the translation algorithm is highly optimized: using the algorithm even huge knowledge bases with contextual conditions can be efficiently converted into f-transducers. Second, since f-tranducers act in a deterministic way a very efficient form of text rewriting is obtained. Third, f-transducers can also be directly applied to text streams. Fourth, composition and the general format of contextual rewrite dictionaries guarantee high flexibility as to the forms of text (stream) rewriting that can be realized.

Keywords: natural language processing, text rewriting, finite-state transducers, failure transitions, contextual conditions

---

Supported by the Bulgarian Ministry of Education and Science via grant DO1-200/2018 "Electronic healthcare in Bulgaria". We acknowledge the provided access to the e-infrastructure of the Centre for Advanced Computing and Data Processing, with the financial support by the Grant No BG05M2OP001-1.001-0003, financed by the Science and Education for Smart Growth Operational Program (2014-2020) and co-financed by the European Union through the European structural and Investment funds.