Data linguistica 23

Multilingual text generation from structured formal representations

av Dana Dannélls

Akademisk avhandling för filosofie doktorsexamen i språkvetenskaplig databehandling,
som enligt beslut av humanistiska fakultetsnämnden vid
Göteborgs universitet kommer att försvaras offentligt tisdagen den
5 februari 2013 kl. 10.15 i Lilla hörsalen, Humanisten.

GÖTEBORGS UNIVERSITET
HUMANISTISKA FAKULTETEN

Göteborg 2012

TITLE: Multilingual text generation from structured formal representations
LANGUAGE: English
AUTHOR: Dana Dannélls

Abstract

This thesis aims to identify the optimal ways in which natural language generation techniques can be brought to bear upon the problem of processing a structured body of information in order to devise a coherent presentation of text content in multiple languages.

We investigate how chains of referential expressions are realized in English, Swedish and Hebrew, and suggest several coreference strategies that can be used to generate coherent descriptions about paintings. The suggested strategies focus on the need to produce paragraph-sized written natural language descriptions from formal structured representations presented in the Semantic Web.

We account for principles of coreference by introducing a new modularized approach to automatically generate chains of referential expressions from ontologies. We demonstrate the feasibility of the approach by implementing a system where a Semantic Web domain ontology serves as the background knowledge representation and where the language-specific coreference strategies are incorporated. The system uses both the principles of discourse structures and coreference strategies to guide the generation process. We show how the system successfully generates coherent, well-formed descriptions in multiple languages.

KEYWORDS: Coherence, computational linguistics, coreference, corpus linguistics discourse structure, knowledge representation, language technology, lexical semantics, linked open data, multilingual natural language generation, natural language processing, ontology, semantic web.

DISTRIBUTION:
Department of Swedish
University of Gothenburg
Box 200
SE-405 30 Gothenburg
Sweden

Data linguistica 23
ISSN 0347-948X
ISBN 978-91-87850-48-6

PRINTED in Sweden by Ineko AB Göteborg 2012