Classification-Based Generation of Route Directions

Doctoral Thesis

Tomasz Marcinak
EML Research gGmbH
Schloss-Wolfsbrunnenweg 33
69118 Heidelberg, Germany
http://www.eml-research.de/english/homes/marciniak

1 Abstract

This dissertation spans two areas of research relevant to the field of Natural Language Generation: domain modeling and tactical generation. The first objective is to design an ontological model of the domain of route directions, focusing on those elements of the process of route following which find a direct manifestation in the linguistic description. Hence, the primary motivation for the identification of the relevant entities in the domain of route direction comes from the analysis of the relevant texts. The ontology should serve as a formal specification of the content of route directions, and as such define the input to the tactical generator. While being necessarily domain-specific, the ontology should be designed to anticipate extensions to other instructional domains.

The second goal pursued in this thesis is development of a data-driven tactical generator capable of realizing the linguistic form of route directions from the underlying conceptual specification sanctioned by the ontology. The task of linguistic realization can formally be defined as a one-to-many mapping between the conceptual content and grammatical form of a linguistic expression. It has long been recognized as a knowledge-intensive process, involving a range of linguistic decisions applying to different levels of the linguistic organization: discourse, clause and phrase levels. To handle these decisions an NLG system requires substantial amount of linguistic knowledge of how both syntactic and lexical constructions

can be used to encode the intended meaning.

Two major considerations involved in designing an NLG system are: the *architecture* of the system and the *knowledge source*. In this dissertation we propose a model of generation which offers a novel perspective on both issues. Firstly, we subscribe to the principle of *lexicon and syntax continuity* proposed by Langacker (1988), and draw no dichotomy between lexical and syntactic decisions, regarding them all as forming a single category of generation tasks. Each such task is further considered as constituting a *classification* problem. The generation process is then modeled as a series of classifications integrated within a *discrete optimization* model. Secondly, the linguistic knowledge required for generation is not specified in an *explicit* way. Instead, in the model of generation proposed here, an annotated corpus is used as a source of linguistic data, from which *machine learning* algorithms can learn how to realize the individual tasks.