Generating Coherent References to Multiple Entities

Albert Gatt

October 12, 2007

What constitutes an adequate reference to a set of objects? Despite intensive research on the Generation of Referring Expressions (GRE), many GRE algorithms either lack empirical backing, or are motivated by concerns which arguably shift their focus away from the crucial problem, which is to generate *natural* descriptions, much as a person would generate them in a comparable situation. This problem becomes much more pronounced in the case of plural reference, where even psycholinguistic research is lacking.

This thesis focuses on the generation of plurals, with particular attention to the semantic heart of the problem, that is, *content determination*. The empirical and computational work addresses two hypotheses. First, descriptions of sets or groups of entities are more adequate if they maximise the similarity between elements of a set. Second, the form and content of referring expressions are strongly determined by the way entities are categorised, that is, what ontological category they belong to.

The first three chapters set the stage with an in-depth theoretical and empirical evaluation of the state of the art in GRE. Here, three main contributions are made. The first is the construction of a *semantically transparent* corpus of singular and plural descriptions. Second, an empirical investigation into reference by human authors in this corpus sheds further light on the content determination problem. Third, an evaluation study is conducted on various existing algorithms. This study is unique in that it is the first to directly address the semantic issues while attempting to abstract away from linguistic realisation. Moreover, it focuses not only on singular, but also on plural descriptions.

The second part of the thesis focuses directly on plurals. It begins (Chap-

ter 5) with a test of the similarity hypothesis on corpus data, leading to the development of a new algorithm which addresses the issues of similarity and conceptual categorisation. The algorithm is also extended with a form of aggregation, again motivated by a new corpus study. This work is generalised to pluralities in discourse in Chapter 6, which starts from the hypothesis that pluralities should be *conceptually coherent*, that is, should conceptualise entities from the same perspective. This hypothesis is investigated in a series of five psycholinguistic experiments. Finally, Chapter 7 uses the results of the previous two chapters to build an integrated framework for content determination in GRE. Among the contributions of this second part of the thesis are (a) the use of an experimental psycholinguistic methodology to test hypothesis that are relevant to generation; (b) the proposal of a novel approach to generation that seeks to satisfy conceptual coherence through the use of corpus-derived similarity metrics.