Case Retrieval Nets for Heuristic Lexicalization in Natural Language Generation

Raquel Hervás and Pablo Gervás

Departamento de Sistemas Informáticos y Programación,
Universidad Complutense de Madrid, Spain
raquelhb@fdi.ucm.es, pgervas@sip.ucm.es

Abstract. In this paper we discuss the use of Case Retrieval Nets, a particular memory model for implementing case-base reasoning solutions, for implementing a heuristic lexicalisation module within a natural language generation application. We describe a text generator for fairy tales implemented using a generic architecture, and we present examples of how the Case Retrieval Net solves the Lexicalization task.

1 Introduction

Natural Language Generation (NLG) is divided into various specific tasks [1], each one of them operating at a different level of linguistic representation (discourse, semantics, lexical,...). NLG can be applied in domains where communication goals and features of generated texts are diverse, from transcription into natural language of numerical contents [2] to literary texts generation [3].

Each kind of NLG application may need a different division into modules [4]. Given a specific organization (or *architecture*) of the system, it may occur that diverse classes of application require different solutions when facing each of the specific tasks involved in the generation process. For a particular task, in processes where a quick answer is required (for instance, in interactive communication between user and machine in real time) it can be useful to use simple solutions based on heuristics, that provide quick answers even if the achieved quality is poor. On the other hand, in situations where long texts of high quality are needed with no constraints on response time it would be better to draw on knowledge-based techniques that exhaustively consider more possibilities.

The present paper proposes a case-based approach to decide which words should be used to pick out or describe particular domain concepts or entities in the generated text. The idea is that people do not create new words each time they need to express an idea not used before, but rather they appeal to the lexicon they have acquired throughout time looking for the best way to express the new idea, always taking into account existing relations between the elements of the lexicon they already know.

The paper starts with a revision of the Case-Based Reasoning and Lexicalization fields. Then we expose the fairy tale text generator where the work presented in this paper is implemented, and we consider the performance of the CBR module. Finally, the obtained results and future research lines are discussed.

C. Bento, A. Cardoso, and G. Dias (Eds.): EPIA 2005, LNAI 3808, pp. 55–66, 2005.

2 Lexicalization and Case Based Reasoning

Lexicalization is the process of deciding which specific words and phrases should be chosen to express the domain concepts and relations which appear in the messages [1]. The most common model of lexicalisation is one where the lexicalisation module converts an input graph whose primitives are domain concepts and relations into an output graph whose primitives are words and syntactic relations. Lexicalization researchers have developed powerful graph-rewriting algorithms which use general "dictionaries" that relate domain primitives and linguistic primitives. Graph-rewriting lexical choice is most useful in multilingual generation, when the same conceptual content must be expressed in different languages. The technique handles quite naturally some kinds of lexical divergences between languages. This scheme can be valid for most applications where the domain is restricted enough in order that direct correspondence between the content and the words to express it is not a disadvantage. In general, thinking on more expressive and versatile generators, this model requires some improvement. Cahill [5] differentiates between "lexicalization" and "lexical choice". The first term is taken to indicate a broader meaning of the conversion of something to lexical items, while the second is used in a narrower sense to mean deciding between lexical alternatives representing the same propositional content. Stede [6] proposes a more flexible way of attaching lexical items to configurations of concepts and roles, using a lexical option finder that determines the set of content words that cover pieces of the message to be expressed. These items may vary in semantic specificity and in connotation, also including synonyms and nearsynonyms. From this set, the subsequent steps of the generation process can select the most preferred subset for expressing the message.

Machine learning techniques have been shown to significantly reduce the knowledge engineering effort for building large scale natural language processing (NLP) systems: they offer an automatic means for acquiring robust solutions for a host of lexical and structural disambiguation problems. A good review of how they have been applied in the past to specifical NLP tasks is available in [7]. Among the learning algorithms that have been used successfully for natural language learning tasks are case based learning methods where the natural language system processes a text by retrieving stored examples that describe how similar texts were handled in the past. Case based approaches have been applied to stress acquisition [8], word sense disambiguation [9] and concept extraction [10], among others.

Case-based Reasoning (CBR) [11] is a problem solving paradigm that uses the specific knowledge of previously experienced problem situations. Each problem is considered as a domain case, and a new problem is solved by retrieving the most similar case or cases, reusing the information and knowledge in that cases to solve the problem, revising the proposed solution, and retaining the parts of this experience likely to be useful for future problem solving. General knowledge, understood as general domain-dependent knowledge, usually plays a part in this cycle by supporting the CBR processes.