Available online at www.sciencedirect.com

ScienceDirect

Fuzzy Sets and Systems 285 (2016) 31–51

FUZZY sets and systems

www.elsevier.com/locate/fss

On the role of linguistic descriptions of data in the building of natural language generation systems

A. Ramos-Soto *, A. Bugarín, S. Barro

Research Center on Information Technologies (CiTIUS), University of Santiago de Compostela, Spain

Received 23 August 2014; received in revised form 10 June 2015; accepted 12 June 2015

Available online 13 July 2015

Abstract

This paper explores the current state of the task of generating easily understandable information from data for people using natural language, which is currently addressed by two independent research fields: the natural language generation field — and, more specifically, the data-to-text sub-field — and the linguistic descriptions of data field. Both approaches are explained in a detailed description which includes: i) a methodological revision of both fields including basic concepts and definitions, models and evaluation procedures; ii) the most relevant systems, use cases and real applications described in the literature. Some reflections about the current state and future trends of each field are also provided, followed by several remarks that conclude by hinting at some potential points of mutual interest and convergence between both fields.

Keywords: Linguistic descriptions of data; Natural language generation; State of the art; Evaluation

1. Introduction

Nowadays, the task of generating easily understandable information for people using natural language is being addressed by two fields which, independently until now, have researched the processes this task involves from different perspectives: the natural language generation (NLG) field and the linguistic descriptions of data (LDD) field.

This paper covers these two research fields which, despite having different origins, are currently on a path which may (and should) lead to their convergence. The natural language generation field consists in the creation of texts which provide information contained in other kind of sources (numerical data, graphics or even other texts), with the aim of making such texts indistinguishable, as far as possible, from those created by humans. On the other hand, the linguistic descriptions of data field, which arises as one of the many applications born from the fuzzy sets theory, provides summaries or descriptions from data sets using linguistic concepts defined as fuzzy sets and partitions, which deal with the imprecision and ambiguity of human language.

* Corresponding author.
E-mail addresses: alejandro.ramos@usc.es (A. Ramos-Soto), alberto.bugarin.diz@usc.es (A. Bugarín), senen.barro@usc.es (S. Barro).

http://dx.doi.org/10.1016/j.fss.2015.06.019
0165-0114/

The NLG field has been in development since the 1980s (although there are systems which date from even before this period, e.g. [1]), when the first applications which translated data into legible texts appeared (e.g., [2,3]). Since then, the complexity of the developed systems has increased notably and there are several techniques and methodologies which guide the building of these solutions [4–6]. Even so, this research field is still open in many respects and there is no unique and well defined approach to address NLG problems.

The linguistic descriptions (or summaries) of data aim to obtain informative, brief and concise descriptions from numeric datasets and cover a group of soft computing-based techniques, such as linguistic variables or fuzzy quantifiers and operators. It is a young field when compared to the NLG domain, whose solutions provide information in the form of linguistic terms. Specifically, although preliminary ideas appeared early in the 1980s [7,8], it started to develop in the second half of the 1990s, when the advances in the field of fuzzy sets (namely computing with words [9] and the computational theory of perceptions [10,11]) provided new potential applications in the descriptive side of data mining. Due to its short career and its formal background, many approaches in this field are on the theoretical side, although in some cases practical examples and real life based problems are given. More recently, the use of hybrid approaches which employ LDD techniques together with NLG systems to provide solutions to real life problems has emerged [12].

This paper is organized in three main sections. Section 2 provides a thorough review on the NLG field, with a special focus on its data-to-text specialty, which deals with the generation of text from raw (usually numeric) data. This includes an overview about the motivations and objectives of this field, followed by an explanation of the most popular architectures and general models (Section 2.1), a review of some of the most relevant NLG systems (Section 2.2), a discussion on NLG evaluation methodologies (Section 2.3) and some general reflections about this research field (Section 2.4). The second part of this paper, Section 3, follows a similar structure as Section 2, where we provide an overview on the LDD field and introduce in Section 3.1 its basic concepts and elements. Section 3.2 reviews both theoretical and applied LDD approaches and Section 3.3 includes some considerations about the current state of LDD. Finally, in Section 4 we share some insights on potential points of interest and convergence for both fields.

2. Natural language generation

Natural language generation (NLG) is described by John Bateman in [13] as the branch of natural language processing which deals with the problem of how texts in human natural language can be automatically created by a machine. This may be seen as the inverse of the problems addressed by natural language understanding but, actually, the NLG field emerges from a very different set of motives and objectives, both theoretical and practical. In this sense, on the theoretical side it explores how language is grounded in non-linguistic information and how it is produced. From a practical point of view, NLG tries to provide solutions for text generation problems in real life application contexts.

The demand of natural language texts which provide all kinds of information is currently increasing. Thus, it is likely that NLG will be a key information technology in the future (a good indicator of this is the considerable number of NLG companies which have emerged in recent years). As a consequence, many NLG systems have found a practical use, while the demand of real life applications is having a growing impact in the approaches and questions contemplated in the NLG field. Examples of well established NLG applications include the generation of weather reports from meteorological data in several languages [14,15], the creation of custom letters which answer customers' questions [16], the generation of reports about the state of neonatal babies from intensive care data [17], and the generation of project management [18] and air quality reports [19].

Bateman also states that, usually, it is hard for a casual user to distinguish between hand made texts, texts built using simple techniques or a complete natural language generation using NLG technology. This is, in fact, what any NLG solution should achieve in order to be considered successful. It should be simply a perfect text production which ideally fulfills the necessities and the knowledge of the reader/listener. This duality directly translates into two quite different research issues within NLG: i) producing texts which are humanlike, and ii) producing comprehensible texts to fulfill certain needs.

The fact that a user is incapable of distinguishing between texts however they are produced is also a problem for the research and development of NLG in the sense that it implies that the required effort to build a successful NLG system is hard to be perceived by users. Since users are not frequently aware of it until something goes wrong, there is little appreciation of the possibilities and complexities of a full natural language generation. In fact, users and application