Expert Systems With Applications 85 (2017) 46–60

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

New types of computational perceptions: Linguistic descriptions in deforestation analysis

Patricia Conde-Clemente[a], Jose M. Alonso[b,*], Éldman O. Nunes[c], Angel Sanchez[d], Gracian Trivino[e]

[a] Departamento de Informática, Universidad de Oviedo, Edif. Dept. 1, Campus de Viesques s/n. 33204 Gijón, Asturias, Spain
[b] Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS), University of Santiago de Compostela, Campus Vida, E-15782, Santiago de Compostela, Spain
[c] Universidade Salvador, UNIFACS, 41770-235 Salvador, Bahia, Brazil
[d] Departamento de Informática y Estadística, Universidad Rey Juan Carlos, 28933 Móstoles, Madrid, Spain
[e] Phedes Lab, Asturias, Spain

A R T I C L E I N F O

Article history:
Received 11 February 2017
Revised 17 April 2017
Accepted 12 May 2017
Available online 13 May 2017

MSC:
03B52
68T30
68T37

Keywords:
Linguistic description of data
Fuzzy logic
Computational theory of perceptions
Granular linguistic description of phenomena
Linguistic summarization
Deforestation analysis

A B S T R A C T

Automatic linguistic description of the available data about complex phenomena is a challenging task that is receiving the attention of data scientists in recent years. As an evolution of previous research results, there is a need of creating new linguistic computational models that allow us dealing with more complex phenomena and more complex descriptions of a growing amount of heterogeneous and real-time data. This paper contributes to this field by presenting three new ways of describing added-value information automatically extracted from data. Also, we extend previous computational models by including a description of the reliability of the available input data. Namely, we face this challenge by using a new implementation of the concept of Z-number proposed by Zadeh. We demonstrate the possibilities of the proposed extension with a practical application. The application generates automatic linguistic reports about the deforestation evolution in the Amazon region, e.g., “The deforestation last month was high. Because of the cloudiness, the reliability of this information is moderate”. Additionally, we evaluate the quality of the generated linguistic descriptions through fuzzy rating scale-based questionnaires. Moreover, we have also made a comparative study between reports generated with and without the new contributions introduced in this paper. The results show that the new types of computational perceptions introduced in this paper are ready to help data scientists to automatically generate good quality reports.

1. Introduction

Humankind generates more and more data in a dynamic and highly technological world where everyone and everything is connected. We are living in the era of Data Science (Dhar, 2013), Big Data (Chen & Zhang, 2014), Social Networks (Scott, 2000), and so on. As result, researchers must face many hot challenges. Among them, we have focused this paper on the text generation challenge, i.e., on the development of computational systems that automatically process data with the aim of generating understandable information using natural language. This new generation of computational systems should provide linguistic reports that are, as informative as required, trusted, relevant, reliable and well organized (Grice, 1975).

This is a novel approach in Data Science. In general, the sequential steps to seek new knowledge in application domains include analysis, representation and interpretation of the extracted knowledge (Cios, Pedrycz, & Swiniarski, 2007). Data scientists commonly analyze these data by using statistics, machine learning or data mining techniques. Usually, they use visualization techniques such as tables, graphs and charts to represent the extracted knowledge. Note that in this type of visualization techniques, users often must make an important effort to interpret the provided information.

Text generation challenges arise with the aim of generating more human-friendly reports. We can see linguistic information as a complement to other forms of visualization. Moreover, linguistic messages provide better understanding for all type of users when they make use of the receiver’s everyday language (Ramos-Soto, Bugarin, & Barro, 2016a). Namely, most approaches for linguistic

* Corresponding author.
E-mail addresses: condepatricia@uniovi.es (P. Conde-Clemente), josemaria.alonso.moral@usc.es (J.M. Alonso), eldman.nunes@pro.unifacs.br (É.O. Nunes), angel.sanchez@urjc.es (A. Sanchez), gracian.trivino@phedes.com (G. Trivino).

http://dx.doi.org/10.1016/j.eswa.2017.05.031
0957-4174/

description of data fit into the standard protoform "*Q* of X are *A*", which is not enough to fulfill most real description needs. Currently, as evolution of the available technology, there is a need of creating new linguistic computational models that allow us dealing with more complex phenomena and more complex descriptions.

This paper contributes to this field by presenting three new ways of describing information about datasets. Also, with the extension of the concept of Computational Perception (CP) by adding a description of the reliability of the available information. We face this last challenge by using the concept of Z-number proposed by Zadeh who introduced Z-numbers as a general tool for describing the reliability of information (Zadeh, 2011). In addition, several authors have made proposals in this research line regarding both theory and applications (Aliev, Huseynov, Aliyev, & Alizadeh, 2015). Namely, a novel approach for decision-making (Aliev, Pedrycz, Kreinovich, & Huseynov, 2016; Lorkowski, Kreinovich, & Aliev, 2014), new applications in the field of natural language processing (Pal & Banerjee, 2013), or a new analytical hierarchy process method (Azadeh, Saberi, Atashbar, Chang, & Pazhoheshfar, 2013), have been recently developed. However, it is worthy to highlight that none of these are directly related to automatic generation of linguistic descriptions.

The main contributions in this paper are summarized as follows:

- Extension of the definition of CP with the aim of properly describing the reliability of the available input data, e.g., "With a high reliability Ana is the tallest woman in the country".
- Definition of three new types of CPs:
  1. **Emphatic CP**. It describes the subset of elements that best fulfill a specific feature, e.g., "The tallest students are Jose, Ana and Maria".
  2. **Superlative CP**. It describes the degree with which the superlative element (attending to a specific feature) is different from the rest of elements in the set, e.g., "The tallest student is Jose by a small difference with respect to the average of the rest of students".
  3. **Comparative CP**. It describes the difference between two elements in the set, paying attention to a specific feature, e.g., "Maria and Ana have a very similar height".
- Testing our proposal in a real-world problem. We present a practical application to generate automatic linguistic reports about the deforestation evolution in the Amazon region. We generate and analyze linguistic descriptions about several representative municipalities in the Amazon region between years 2000 and 2014. We evaluate the quality of these linguistic descriptions by using fuzzy rating scale-based questionnaires. However, it is noteworthy that automatically assessing the quality of linguistic descriptions is still an open issue (Bugarín, Marín, Sánchez, & Trivino, 2015) that we do not address deeply in this paper. Please note that, in this paper, we do not deal with the problem of how to obtain the best linguistic descriptions but just with exploring the possibilities of successfully applying the proposed new CPs to generate good quality human friendly reports.

The remainder of the paper is organized as follows. Section 2 presents some preliminary concepts that are required to understand the rest of the manuscript. Section 3 extends the definition of CP. It takes into account the reliability of perceptions. Section 4 presents the three new types of CPs. Section 5 describes the practical application under study. Section 6 goes in detail with the experimentation and discussion of results. Finally, Section 7 summarizes the main conclusions and sketches future work.

2. Preliminaries

2.1. Background on data-to-text

According to Ramos-Soto, Bugarín, and Barro (2016b), there are two main research lines in the specialized literature for text generation from data (Data-to-Text, in short): (1) Natural Language Generation (NLG); and (2) Linguistic Descriptions of Data (LDD).

On the one hand, NLG deals with generating texts that are indistinguishable from those produced by humans. The input can be numerical data, graphics or even other text. The first applications are dated back to 1980s (Kittredge, Polguére, & Goldberg, 1986). Applications of NLG include the generation of weather reports from meteorological data (Coch, 1998; Goldberg, Driedger, & Kittredge, 1994), the generation of reports about the state of neonatal babies from intensive care data (Portet et al., 2009), air quality reports (Busemann & Horacek, 1997), etc. Nowadays, NLG is an increasingly relevant challenge (Cambria & White, 2014) and several companies offer text generation from Big Data as a commercial product or service.

Reiter and Dale were the first to propose a generic methodology and architectural framework for designing NLG systems (Reiter & Dale, 2000). To face the complexity of the whole NLG system, they divided the text generation process in a pipeline with three core tasks: (1) Document Planner, (2) Micro Planner and (3) Surface Realizer.

Firstly, the Document Planner decides what information should be communicated in the output text. In addition, it must provide order and structure over the information to be conveyed. Then, the Micro Planner solves the problem of selecting the content words (nouns, verbs, adjectives, adverbs, etc.) for the generated text. It is also in charge of the linguistic aggregation, which involves the use of linguistic resources to build sentences that communicate several pieces of information at once. Finally, the Surface Realizer produces the final text. This task is called linguistic realization and is generally viewed as the problem of applying some characterization of the rules of grammar to some more abstract textual representation; thus, producing a text which is syntactically and morphologically correct.

It is noteworthy that even though the pipeline described above has served as inspiration for many NLG systems, the design of NLG systems remains as an open hot research challenge.

On the other hand, LDD deals with generating linguistic descriptions (or summaries) from numerical datasets. Concepts defined as fuzzy sets and partitions (Zadeh, 1965) are used to manage the uncertainty and vagueness inherent to the human language. The key concept of fuzzy linguistic summary was established in Yager (1982); Zadeh (1983) and a recent review on related methods is available in Boran, Akay, and Yager (2016). Typically, fuzzy linguistic summaries are based in the idea of quantified propositions which have the ability of summarizing a set of elements related to the specific domain of a phenomenon under consideration. For example, "*Most* of the students are *tall*" corresponds with the quantified sentence "*Q* of X are *A*", where *Most* is the quantifier *Q* and *tall* is the fuzzy predicate *A*, i.e., an attribute (feature) that characterizes the perception of the phenomenon *"the students' height"* (X).

In addition, Zadeh proposed the Computational Theory of Perceptions (Zadeh, 1999; 2002) as a framework to develop computational systems with the capacity of computing with the meaning of natural language expressions, i.e., with the capacity of computing with imprecise descriptions of the world in a similar way how humans do. In recent years, this concept has been extended in different ways and used for different applications, e.g., data mining (Yager, 1995), database queries (Castillo-Ortega, Marín, & Sánchez, 2011b; Kacprzyk & Zadrozny, 2010), description of temporal series