Generating Linguistic Descriptions of Data Using Fuzzy Set Theory[1]

Nicolás Marín and Daniel Sánchez

Department of Computer Science and A.I.,
University of Granada, 18071, Granada, Spain
nicm@decsai.ugr.es,daniel@decsai.ugr.es

In this extended abstract we give our view, as introduced in [1], of the problem of *generating linguistic descriptions of data* (GLiDD). This task is performed by *data-to-text* systems, which have their origin and more advanced development in the field of natural language generation (NLG). These systems aim at generating texts, we call *linguistic descriptions of data*, from non-linguistic input data, expressing knowledge extracted from the latter as humans would do, with the objective of satisfying specific user's needs. As different authors have shown, using a brief text is a feasible, and sometimes the most effective, mean for data description. This is particularly the case when the devices employed are based on written or spoken natural language, when the user is visually impaired, when the understanding of visual or mathematical means requires expert knowledge or complex cognitive tasks, etc. [1]

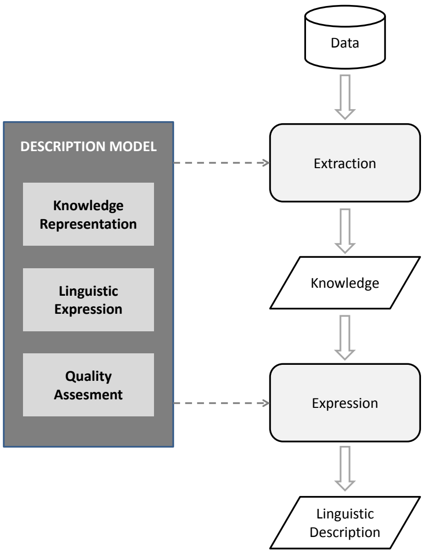

Fig. 1. Components and tasks of GLiDD.

In our view, illustrated in Figure 1, GLiDD consists of two main tasks: a *knowledge extraction* process which, in a broad sense, can be considered as a KDD (knowledge discovery in databases) process, and a *linguistic expression process*, which enhances the understandability and usefulness of the obtained knowledge by appropriately expressing it using natural language [1]. Both tasks are equally complex and important.

The extraction and expression tasks rely on a description model. As shown in Figure 1, this description model is comprised of three components. First, a *knowledge representation formalism* for representing the semantics of the *messages* (information content of the final text) extracted from the data. Second, a *linguistic expression model*, which determines the most appropriate text for transmitting the extracted information on the basis of both preferences and pragmatic

[1] This work has been supported in part by the Spanish Ministry of Economy and the European Regional Development Fund (FEDER) under project TIN2014-58227-P.

competence of the user, the final objective of the system, and other contextual aspects. Finally, a very important component is the *quality assessment model*, which is employed both for guiding the GLiDD process and for validating the results. As for the GLiDD process, the quality model comprises many *quality dimensions* employed in the extraction process (brevity, accuracy, relevance, data coverage, etc.), and in the expression process (user preferences, semantics/pragmatics issues, etc.). In validating the final results, dimensions include similarity with results provided by humans, satisfaction expressed by the user, novelty, usefulness, impact on the user's decisions, and beliefs, etc.

In parallel to the efforts of researchers in the NLG area, but mostly unaware of the relation to data-to-text systems, the generation of linguistic descriptions of data has been also dealt with by researchers in the fuzzy sets (FS) community, mostly under the name of *linguistic summarization.* In recent years, the research in this area is more and more in confluence with the view and objectives of the NLG community [2, 3]. Techniques developed within FS put the focus mainly on the extraction, whilst NLG techniques focus specially on the expression task. We consider in [1] that the extraction task (called *content determination* in the NLG view of GLiDD) can be addressed from the point of view of KDD, using algorithms, resources and concepts from that area. On its turn, the expression task is addressed by NLG using architectures and a large amount of techniques developed specifically for data-to-text systems within this field.

Both in the areas of KDD and NLG it is widely recognized the potential contributions of uncertainty representation techniques. In KDD, fuzzy sets improve understandability of the extracted knowledge, allowing to represent patterns by means of soft restrictions. By gradually diminishing the precision in the representation, we can express knowledge about data using higher levels of abstraction, diminishing the complexity of the description and increasing its usefulness.

In [1] we have shown by means of a deep review of the proposals in the literature, that formalisms employed in data-to-text systems can be seen as structured collections of protoforms within Zadeh's Generalized Theory of Uncertainty. This view allows us to deal with different kinds of uncertainty in GLiDD, filling the semantic gap between data and linguistic terms and expressions, a necessity widely acknowledged by the authors working on this area, including those in the NLG community. We have enumerated a rather long but non-exhaustive list of potential contributions of fuzzy sets for the GLiDD problem.

References

1. Marín, N., Sánchez, D.: On generating linguistic descriptions of time series. Fuzzy Sets and Systems **285** (2016) 6–30
2. Kacprzyk, J., Zadrozny, S.: Computing with words is an implementable paradigm: Fuzzy queries, linguistic data summaries, and natural-language generation. IEEE Trans. Fuzzy Systems **18**(3) (2010) 461–472
3. Ramos-Soto, A., Bugarín, A., Barro, S.: On the role of linguistic descriptions of data in the building of natural language generation systems. Fuzzy Sets and Systems **285** (2016) 31–51