Linguistic summaries as descriptive analytics tool

Carla Wrede*, Anna Wilbik* and Mark H.M. Winands*
* Department of Advanced Computing Sciences,
Maastricht University, Maastricht, The Netherlands
{c.wrede, a.wilbik, m.winands}@maastrichtuniversity.nl
† Department of Computer Science and Artificial Intelligence,
Universidad de Granada, Granada, Spain
{nicm, daniel}@decsai.ugr.es

***Abstract*—This paper aims to give an introduction to linguistic summaries as a tool to capture descriptive data analytics in form of natural language. We introduce the concepts behind it, like fuzzy sets, linguistic variables and quantifiers, and create a linguistic summary together. This presented paper acts as a summary of the full paper, which can be found online in IEEE Xplore. The full version features interactive elements to help understand the material in a practical way.**

***Index Terms*—Linguistic summaries, fuzzy logic**

I. Introduction

Recent developments in ICT technology enable more and more data to be collected. This rise in data leads to an increased potential to be used to support decision-making processes. In turn, also more data mining and knowledge discovery methods are deployed. Both the data used and results coming from these methods need to be described and possibly explained. While there are many ways to achieve this task, this paper focuses on linguistic summaries. Linguistic summaries are template-based, semi-natural language-like sentences that capture the underlying characteristics of the data and results. In the following, we describe the necessary background knowledge of these summaries, how they can be created and outline how they are used in real-world applications.

II. Linguistic Variable

Important key concepts of linguistic summaries are linguistic variables and linguistic values. In the words of Zadeh [1], a linguistic variable is a variable whose values are words or sentences in a language (natural or artificial). Each of these values is called a term and is associated with a semantics that is nothing more than a fuzzy restriction defined on a base universal set.

According to [1], a linguistic variable is defined as a quintuple (v, T, X, g, m), where v is the name of the variable, T is the set of linguistic terms of v, X is the universal set over which fuzzy restrictions are defined, g is a syntactic rule for generating linguistic terms (a so-called grammar), and m is a semantic rule that assigns to each linguistic term its meaning, which is a fuzzy restriction on X. In most cases, simple grammars g are used, in which the terms and their associated restrictions are provided directly. However, one could use a more complex, general grammar g, which allows to construct terms of the variable from base terms, connectives, and modifiers, for instance, generate *very large* from *large*, and *quite small* from *small*.

As an example, let us consider the iris data set and the linguistic variable sepal length. In this case, the set of linguistic terms may be $T = small, medium, large$. X is the set [4.3, 7.9], as this is the range of the sepal length. Each of the linguistic terms is mapped to a fuzzy set. Figure 1 illustrates the structure of the linguistic variable sepal length.

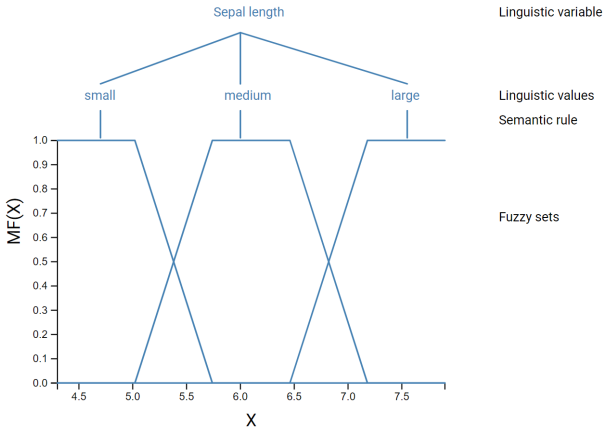

Fig. 1. Representation of the linguistic variable sepal length

III. Linguistic Quantifier

A special kind of linguistic variable, essential in linguistic summaries is a (linguistic) quantifier. Linguistic quantifier describe a quantity in linguistic terms. Quantifiers are divided into two main classes: absolute quantifiers and relative quantifiers. Absolute quantifiers are defined over numbers, $\mathbb{R}$, and represent absolute quantities, such as "around 5", "at least 10". The relative quantifiers, exemplified by "at least a half", "most", are defined on interval [0,1].

IV. Linguistic Summary

The notion of linguistic summary was introduced by R.R. Yager [2] as a fuzzy logic-based approach to the summarization of data. Yager's approach considers linguistic summaries

We thank Nicolás Marín Ruiz and Daniel Sánchez Fernández from the Department of Computer Science and Artificial Intelligence of Universidad de Granada for their help and input. This research has been funded by the Rijksdienst voor Ondernemend Nederland (RVO) in the framework of the project Green Transport Delta - Elektrificatie.

described by means of quantified statements of the form *Q [R] y's are P*, where:

- Q is a linguistic quantifier (e.g. *most*).
- Y's are the objects (e.g. tuples in a database) over which we want to provide the summary.
- P is called the *summarizer*, and corresponds to a fuzzy subset of Y. It is usual that P is induced by some linguistic label defined on the domain of one variable defining some aspect of objects in Y. For instance, if we consider Y to be a set of flowers and the variable *sepal length*, and we consider the linguistic label *small* defined on the domain of possible sepal lengths of flowers, then P can be defined as *the set of small flowers*, the membership of each person y to P being the membership degree of the sepal length of y to the label *small*.
- R is an optional argument called *qualifier*, corresponding to a fuzzy subset of Y. It allows to restrict the summary to those objects in Y that satisfy R. As with P, it can be induced by linguistic labels on the domain of variables of objects. Note that, in practice, avoiding the use of a qualifier is equivalent to considering that $R = Y$.

As an example we can consider the quantified statements *Most of the flowers are small*, where Q is the quantifier *most*, Y is a set of flowers (the specific set for which we want to perform the summary), and *small* stands for the fuzzy subset of *small flowers* defined on Y. In this case, it is $R = Y$ or, if you prefer it, we are avoiding the use of a qualifier R.

On the basis of these quantified statements, a linguistic summary is defined as a pair comprised of:

- A quantified statement of the form *Q [R] y's are P*, and
- A truth value T, corresponding to the accomplishment degree of the quantified statement on the data set, representing to which degree the quantified statement is accurate for the summarized data.

The truth value T can be computed in different ways, and differs whether the quantifier Q is absolute or relative. One of the most widely used approaches is proposed by Zadeh and employed in the original studies by Yager [2]. However, adaptations and improvements have been proposed as well, one of them the GD method as introduced in [3].

V. Applications

Linguistic summaries have the advantage that they can be adapted to summarize different types of data. Previous research has seen them applied on numeric data from databases, time series, standardized texts, description of images, videos, sensor data, web logs and event logs. In practice, several real-world applications have benefited from the use of linguistic summaries.

Strykowski et al. [4] described a use case in commerce, where linguistic summaries were used to identify and communicate the patterns in selling computer parts and weather in a B2C setting. Based on those patterns, the shop was predicting their sales and required personnel.

Linguistic summaries are also used to monitor the status of elderly residents in an eldercare facility, TigerPlace in Columbia, Missouri, USA [5]. Motion sensors were placed in the apartments of elderly residents. The sensors measure the activity level of their residents, that later is turned into linguistic description of the activity of the residents. The linguistic summaries were proven to have sufficient sensitivity to capture the change in behavior or adverse health events [6].

Linguistic summaries are also used in the agri-food industry, to monitor the climate inside a pigsty [7]. The evaluations with the farmers proved that linguistic summaries can be easily interpreted and understood by non-technically trained people.

Last, but not least, linguistic summaries seem to be good mechanisms for explanations. Initial works showed that the summaries can be used to find an explanation of an anomaly in the process data [8]. Moreover, linguistic summaries score highly in users' evaluations in terms of explanation understandability and usefulness [9].

VI. Conclusion

This paper introduced the concept of linguistic summaries as descriptive analytics tool. It is one technique for capturing data in form of natural language, which helps people with a non-technical background to understand the result from data analysis tools. The concept behind linguistic summaries, fuzzy sets, has been introduced, followed by the key concepts of linguistic variables and quantifiers. We have constructed a linguistic summary together and calculated its truth value and have seen some applications where linguistic summaries have been deployed. While this is just a brief introduction, there are many more aspects to linguistic summaries and we can recommend more interested readers a look into the cited references. For the full experience, we refer to the corresponding interactive article.

References

[1] G. J. Klir and B. Yuan, Eds., *Fuzzy Sets and Fuzzy Logic, Theory and Applications*. Prentice Hall, 1995.

[2] R. R. Yager, "A new approach to the summarization of data," *Information Sciences*, vol. 28, pp. 69–86, 1982.

[3] M. Delgado, M. J. Martín-Bautista, D. Sánchez, and M. A. Vila, "Fuzzy cardinality based evaluation of quantified sentences," *International Journal of Approximate Reasoning*, vol. 23, pp. 23–66, 2000.

[4] J. Kacprzyk and P. Strykowski, "Linguistic summaries of sales data at a computer retailer via fuzzy logic and a genetic algorithm," in *Proceedings of the 1999 Congress on Evolutionary Computation*, vol. 2. IEEE, 1999, pp. 937–943.

[5] A. Wilbik, J. M. Keller, and G. L. Alexander, "Linguistic summarization of sensor data for eldercare," in *2011 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2011, pp. 2595–2599.

[6] A. Wilbik and J. M. Keller, "Anomaly detection from linguistic summaries," in *2013 IEEE International Conference on Fuzzy Systems*. IEEE, 2013, pp. 1–7.

[7] A. Wilbik, D. Barreto, and G. Backus, "On relevance of linguistic summaries - A case study from the agro-food domain," ser. Communications in Computer and Information Science, vol. 1237. Springer, 2020, pp. 289–300.

[8] S. Chouhan, A. Wilbik, and R. Dijkman, "Explanation of anomalies in business process event logs with linguistic summaries," in *Proceedings of WCCI 2022*, 2022.

[9] C. Wrede, M. H. M. Winands, and A. Wilbik, "Linguistic summaries as explanation mechanism for classification problems," *The 34th Benelux Conference on Artificial Intelligence and the 31st Belgian Dutch Conference on Machine Learning*, 2021.