On a Paradox of Extended Linguistic Summaries

Citation of final publication:

Published in its final form at https://doi.org/10.1109/FUZZ48607.2020.9177598

This is the accepted manuscript.

©2020, IEEE

Downloaded from DRO:
http://hdl.handle.net/10536/DRO/DU:30141077
On a Paradox of Extended Linguistic Summaries

Anna Wilbik
Information Systems
Eindhoven University of Technology
Eindhoven, The Netherlands
a.m.wilbik@tue.nl

Timothy C. Havens
College of Computing
Michigan Technological University
Houghton, MI, USA
thavens@mtu.edu

Tim Wilkin
School of Information Technology
Deakin University
Geelong, Australia
tim.wilkin@deakin.edu.au

Abstract—Continued developments in information technologies allows for increasingly more data to be collected for decision making purposes. While statistical summaries and aggregations are commonly applied to such data, linguistic summaries capture essential features and relationships in the data and better support human users to understand complex data sets. The basic quality measure of linguistic summaries is the truth value, describing the validity of the sentence. Several methods for calculating the truth value have been proposed. In this paper we analyze several popular methods and show a strange, contradictory behavior in case of extended protoforms, which can result in misleading or non-intuitive results to the user. These results highlight the need for further research into linguistic summarization and the computation of truth values for real data sets.

Index Terms—linguistic summarization, extended protoform, truth value, dual summary, paradox

I. INTRODUCTION

Data drives a vast majority of modern business, industrial, and governance processes. Over the past decades the data has enabled us to gain deep and significant insights into social, political, and economic structures around the world. However, estimates of the continual growth of data volumes generated annually is as high as 61% CAGR [1] and given that humanity is generating approximately 2.5 exabytes of data daily, we produce far more data than can be readily accessed and processed to support important decision processes. While automation of such processes is one important research direction, human decision makers are still required in many critical domains. Decision support systems aide decision makers by summarizing and aggregating data, and transforming it into formats and feature spaces that humans can more readily interpret and understand. While many statistical and machine-driven data analysis methods utilize quantitative data, humans are more naturally disposed to interpret visual and natural language representations of information. While graphs are a common way to visually represent data relationships and insights, they still require training and experience to be able to interpret quickly and accurately [2]. Natural language remains the most salient means of communicating with humans. Computational models for linguistic summarization of data aim to provide human-centered methods for communicating insights and critical information from quantitative data sources.

For instance, Dubois and Prade [3], [4] proposed representation and reasoning for gradual inference rules for linguistic summarization in the form “the more X is F, the more/the less Y is G” that could summarize various relationships. Such rules expressed a progressive change of the degree to which the entity Y satisfies the gradual property G when the degree to which the entity X satisfies the gradual property F is modified. This concept was later investigated by Oudini et. al [5], [6] and Wilbik et. al [7]. Rasmussen and Yager [8] discussed the benefit of using fuzzy sets in data summaries based on generalized association rules. Gradual functional dependencies were investigated, and a query language called SummarySQL was proposed. Bosc et al. [9] discussed the use of fuzzy cardinalities for linguistic summarization. The SAINTETIQ model [10] provides the user synthetic views of groups of tuples over the database. In [11], the authors proposed a summarization procedure to describe long-term trends of change in human behavior, e.g., “the quality of the ‘wake up’ behavior has been decreasing in the last month” or “the quality of the ‘morning routine’ is constant but has been highly unstable in the last month.”

We will follow the approach of Yager [12], in the form “Q objects in Y are P.” This approach was considerably advanced and then implemented by Kacprzyk [13], Kacprzyk and Yager [14], and Kacprzyk et al. [15]–[17]. This approach has also been applied to different types of data: numerical [18]–[20], time series [21]–[23], sensor data [11], [24], texts [25], videos [26]–[28] and processes [29], [30].

In this paper we focus on the linguistic summaries of numerical data and the extended protoform [12], i.e., “Q R objects in Y are P.” In our work we have encountered a strange behaviour of linguistic summaries that we describe and investigate herein.

This paper is structured as follows. The next section introduces the background of linguistic summaries and Section III demonstrates the encountered paradox. We then demonstrate how different methods for calculating the truth value behave in this paradoxical situation. Last, are some concluding remarks and recommendations for further study.

This work was supported in part by the Michigan Tech Visiting Professor program, which is funded by a grant to the Michigan Tech Provost’s Office from the State of Michigan’s King-Chavez-Parks Initiative, and the Michigan Tech Institute of Computing and Cybersystems.
II. BACKGROUND

We consider linguistic data summaries as quantified propositions with two possible protoforms (or templates) [12]:

- simple protoform:
  \[ Qy's \ are \ P; \]  
  e.g. Most boxes are large

- extended protoform:
  \[ Q \ Ry's \ are \ P; \]  
  e.g. Most large boxes are heavy

where \( Q \) is the quantifier, \( P \) is the summarizer, and \( R \) is an optional qualifier, which are all modeled as fuzzy sets over appropriate domains.

The truth value, describing the validity of the summary, is the basic measure of the quality of the summary; thus, it is an essential part of the summary. Many methods for calculating the truth value have been proposed; for some examples, see [31]. The truth value is not the only quality measure of a linguistic summary. Kacprzyk et al. [17], [32] proposed four additional measures, namely the degree of specificity, the degree of appropriateness, the degree of covering, and the length of the summary. Bugarin et al. [33] differentiated between evaluating a single summary sentence and a set of summaries. They proposed several measures that capture aspects, such as coverage, length, and specificity. An overview of different quality criteria can be found in [34]. For completeness purposes, we now briefly describe the four methods that we use in this work.

A. Zadeh’s calculus of quantified propositions

In this approach [35], the truth value, denoted as \( T_Z \), of a simple protoform summary is calculated as

\[
T_Z(Qy's \ are \ P) = \mu_Q \left( \frac{1}{n} \sum_{i=1}^{n} \mu_P(y_i) \right),
\]

the extended protoform summary is calculated as

\[
T_Z(Q \ Ry's \ are \ P) = \mu_Q \left( \frac{\sum_{i=1}^{n} \mu_P(y_i) \wedge \mu_R(y_i)}{\sum_{i=1}^{n} \mu_R(y_i)} \right),
\]

where \( n \) is the number of objects in the data, and \( \mu_P, \mu_Q, \) and \( \mu_R \) are the membership functions of the summarizer, the qualifier, and the quantifier, respectively. The notation \( \wedge \) denotes the minimum operator; in the more general case, it can be any \( t \)-norm [36].

B. Method based on Sugeno integral

Using the approach in [20], the truth value, denoted as \( T_S \), is computed by

\[
T_S(Qy's \ are \ P) = T_S(Q_1 \ Ry's \ are \ P) \wedge T_S(Q_2 \ Ry's \ are \ P)
\]

where quantifier \( Q \) is split into \( Q_1 \) and \( Q_2 \), so that \( Q_1 \) is a non-increasing quantifier and \( Q_2 \) is non-decreasing. For the non-decreasing quantifier \( Q_2 \) the truth value is calculated as

\[
T_S(Q_2 \ Ry's \ are \ P) = \max_{\beta \in [0,\max(R(y_i))]} \left( \max_{\alpha \in [0,1]} \{ \alpha \wedge Q_2(P_{\alpha \beta}) \} \right),
\]

where \( R_\beta \{ y_i \in Y : \mu_R(y_i) \geq \beta \} \) and

\[
P_{\alpha \beta} = \frac{\{ y_i \in R_\beta : \mu_P(y_i) \geq \alpha \}}{|R_\beta|}, \quad \forall |R_\beta| > 0.
\]

For the non-increasing quantifier \( Q_1 \), the truth value is calculated for the dual summary to “\( Q_2 R \ y's \ are \ P. \)” Namely, “\( \bar{Q}_2 R \ y's \ are \ \bar{P} \)” where \( \bar{\cdot} \) denotes antonym and \( \bar{\cdot} \) is complement.

C. GD method

The truth value according to the GD method [37] is calculated as

\[
GD(Qy's \ are \ P) = \sum_{c \in CR(P/R)} ER(P/R,c) \times \mu_Q(c),
\]

where

\[
ER(P/R,c) = \sum_{c = C(P/R,\alpha_i)} (\alpha_i - \alpha_{i+1}), \quad \forall c \in CR(P/R),
\]

\[
CR(P/R) = \left\{ \frac{|(P \cap R)|_\alpha}{|R|} : \alpha \in M(P/R) \right\},
\]

\[
M(P/R) = M(P \cap R) \cup M(R),
\]

\[
M(P) = \{ \alpha \in [0,1] : \exists y_i, \mu_P(y_i) = \alpha \},
\]

\[
C(P/R,\alpha_i) = \frac{|(P \cap R)|_\alpha}{|R|},
\]

If \( R \) is not a normal fuzzy set, it is first normalized. The fuzzy set \( P \cap R \) is scaled using the same factor used in the normalization of \( R \).

D. ZS method

The truth value according to the ZS method [37] is calculated as

\[
ZS(Qy's \ are \ P) = \max_{c \in CR(P/R)} \min\{ ES(P/R,c), \mu_Q(c) \}
\]

where

\[
ES(P/R,c) = \max \left\{ \alpha \in M(P/R) : c = \frac{|(P \cap R)|_\alpha}{|R|}, \quad \forall c \in CR(P/R), \right\}
\]

and \( M(P/R) \) and \( CR(P/R) \) are defined as in GD method at (7d) and (7c), respectively. Similar to GD, if fuzzy set \( R \) is normalized, then the fuzzy set \( P \cap R \) is scaled using the same factor used in the normalization of \( R \).
E. Properties for evaluating linguistic summaries

Delgado, Sanchez and Vila [37] proposed a set of properties for evaluating linguistic summaries: six for simple protoforms and eight for extended protoforms. Here, we reintroduce only the properties for extended protoforms.

- (Crisp case). If A and D are crisp, then the (known) result of the evaluation must be \( Q \left( \frac{|P \cap R|}{|R|} \right) \), where Q is relative quantifier.
- In the case \( R = X \) and for relative quantifiers, the resulting evaluation method is a valid method for the evaluation of simple protoform sentences.
- Evaluation must be time-efficient.
- If \( R \subseteq P \) and \( R \) is a normal set then the evaluation method must return the value \( Q(1) \).
- If \( D \cap A = \emptyset \) then the evaluation method must return the value \( Q(0) \).
- Evaluation must be coherent with fuzzy logic in the case of the quantifiers “exist” and “all”.
- Evaluation must allow us to use any quantifier, i.e. any possibility distribution over \([0,1]\).
- Evaluation must not be too “strict”, i.e. given a quantifier \( Q \) in the rational interval \([0,1]\) with \( Q \neq \emptyset \) and \( Q \neq H = \{p/q \text{ with } p \in \{0,\ldots,n\}, q \in \{1,\ldots,n\} \} \) we must be able to find fuzzy sets \( P \) and \( \bar{P} \) so that the evaluation of the sentence is not in \([0,1]\).

In [37] it was shown, that method ZS fulfills all the above mentioned properties.

III. Paradox - Examples

Consider a set of objects \( \{o_1, o_2, \ldots, o_n\} \) that have properties, \( f_1, f_2, \ldots, f_m \). Those properties can be described with linguistic values, e.g., for feature \( f_1 \), we have \( lv_{1,1}, lv_{1,2}, \ldots \). Hence, for each object we can calculate the membership degree to which object \( o_i \) is \( lv_{k,1} \) for feature \( f_k \).

Without loss of generality summarizer \( P \) is \( lv_{1,1} \) and qualifier \( R \) is \( lv_{2,1} \). We consider the form of linguistic summary, “Most R objects are P.”

We also look at the dual linguistic summary [38], i.e., “most \( \bar{R} \) objects are \( \bar{P} \)” where \( \bar{R} \) denotes antonym of \( R \) and \( \bar{P} \) denotes complement. The summary and the dual summary should have equal truth value. Without this symmetry we may arrive at a set of inconsistent summaries.

Consider a family of sets with the following characteristics: \( O = \{o_i, i = 1, \ldots, n; \mu_P(o_i) \geq \mu_R(o_i)\} \). We observe a strange behavior, that is, a paradox when \( \max \mu_R(o_i) \) is getting smaller and tends to a value of 0.

As an illustrative example, consider a set of puppies \( \{p_1, p_2, \ldots, p_5\} \). Those puppies can be characterized by the properties such as size, fur quality, and appearance. In this example size can be described as small, medium, big. Quality of the fur will be described with linguistic values soft, coarse, or rough. Appearance will be described as either cute or ugly (not cute). We consider four sets of five puppies, for which we have given their degree of softness and cuteness. Those sets are shown in Tables I-V. \( \varepsilon \) from set 4, shown in Table IV, is an extremely small value, yet greater than 0.

The reader may notice that the degree of cuteness is the same for the five sets and only the degree of softness varies. Three of the puppies have zero membership for soft fur, while two puppies have a degree of membership that gets smaller for each subsequent set. We are interested in the truth value of the linguistic summary “most soft puppies are cute” and its dual summary “most soft puppies are not cute.”, as this degree of membership falls toward 0.

We used five methods for calculating the truth value: two methods based on Zadeh’s calculus of quantified propositions (with min and Łukasiewicz as t-norms), the Sugeno integral aggregation [20], and the ZS and GD methods proposed by Delgado, Sanchez, and Vila in [37]. Note that in early work [14], [16] calculations of truth value using Zadeh’s calculus assumed only monotonic non-decreasing quantifiers. Problems with this assumption were highlighted in [24], [38], and following these authors we ignore this assumption.

Quantifier most is defined as a trapezoidal membership function \( \text{Trap}[0.6, 0.8, 1, 1] \) and can be calculated as \( \max\{\min\{5x - 3, 1\}, 0\} \). Quantifier most is antonym of most, i.e., \( \text{Trap}[0,0,0.2,0,4] \) and can be calculated as \( \max\{\min\{-5x + 2, 1\}, 0\} \).

Truth value results are shown in Table VI. We now use these results to illustrate a few issues with the quality measures. First, only one method (the Sugeno integral) managed to obtain equal truth values for the summary and its dual summary across all sets of puppies. All other methods failed to obtain equal truth values for the summary and its dual for both Sets 3 and 4. Those sets contain some puppies that have very small non-zero membership in softness. The GD method also failed to obtain equal truth values for the summary and its dual on
Set 2.

Moreover, compare the truth values obtained for the summaries calculated for Sets 4 and 5. Firstly, Zadeh’s calculus fails for Set 5, and the choice of t-norm determines the result on Set 4. For the ZS and GD methods the results are more concerning. For small, arbitrary \( \varepsilon > 0 \) the truth values of "Most soft puppies are cute" is 1, but changes to 0 for \( \varepsilon = 0 \). This is a significant change in truth value for a very small change in the data. This change is also contradictory with human intuition, which would suggest that the truth value should approach 0 smoothly as \( \varepsilon \downarrow 0 \). Furthermore, this step change in truth value at this limit violates one of the required properties for an evaluation method for linguistic summaries, namely that the evaluation must not be too strict.

Set 4 seems to support two possible interpretations. Either "most soft puppies are cute":
1) is true to a very small degree because there is a small degree of support for this in the data (which follows the way of reasoning applied in a rule bases); or,
2) is entirely true as those puppies with a small degree of softness are to a high degree cute.

The ZS and GD methods give a result corresponding to (2), however this seems to go against intuition regarding this data, which says that we don’t know much about the cuteness of puppies that are to a high degree also soft, and hence cannot be confident in the statement "most soft puppies are cute". In other words, the data suggests that "most (all) cute puppies are not soft", which seems at odds with "Most soft puppies are cute".

Unfortunately the Sugeno integral is not the complete solution to avoiding paradoxical truth values for linguistic summaries. Consider the set of puppies shown in Table VII. For this set the summary “most soft puppies are cute” has a truth of 1, which is also contradictory to human intuition. Clearly, four of the soft puppies – those with 0.99 membership in being soft – are not cute at all, having a membership of merely 0.01 in cute.

### IV. CONCLUDING REMARKS

We have shown herein a paradox within extended protoform linguistic summaries in the case of objects with low qualifier membership degree for many of the evaluation methods presented. The paradox is clearly visible when we compare the truth values of a summary and its antonym. Only the method based on the Sugeno integral [20] is able to fulfill the requirement of equality of these truth values.

One might argue that in real applications such summaries are never interesting to the user because of low membership values of the qualifier, hence low values of degree of focus [39]. However, given the desirability of producing linguistic summaries from real data sets, it is important to understand how the selected methods will handle these corner cases, and avoid results that violate human intuition (without good cause).

In our future work we will work towards improving the evaluation method based on Sugeno integral, so that inconsistent results will not occur.

### ACKNOWLEDGMENT

No puppies were harmed in this research, although some humans were unhappy with puppies being described as ugly. We would like to emphasize that while not all puppies may be cute, or soft, or even puppies anymore, they all deserve a loving home; so, please think of your local animal shelter or pet rescue the next time you’re looking for a new animal companion.

### REFERENCES
