Linguistic summarization of data with probabilistic fuzzy quantifiers

F. Díaz-Hermida, A. Bugarín.
Department of Electronics and Computer Science
University of Santiago de Compostela
15782 Santiago de Compostela. Spain.
{felix.diaz,alberto.bugarin.diz}@usc.es

Abstract

In this work we study how a probabilistic based model of fuzzy quantification can be used to build quantified fuzzy summaries. Given a particular set of data, we describe how a set of expressions endowed with a semantics that is convenient and comprehensive for human consumption can be generated for linguistically summarizing data.

Keywords: fuzzy quantification, linguistic summaries, determiner fuzzification schemes.

1 Introduction

Although fuzzy quantification has been presented in the literature as a powerful mechanism for summarizing data [7, 8, 9], building of summaries is not always based on plausible models of fuzzy quantification [1, 2, 3, 6].

In [6] three main issues in fuzzy quantification are identified: interpretation, reasoning and summarization. Interpretation is the main basic goal, as the other issues depend on a correct modeling of fuzzy quantifiers. Both reasoning and summarization depend on the first issue, since a plausible behavior cannot be expected for these tasks when non-plausible models are used for quantified fuzzy reasoning. Therefore, building of linguistic summaries based on fuzzy quantifiers demands using plausible fuzzy quantification models in order to be successful.

In this paper, we show that probabilistic fuzzy quantification models [2, 3, 4] fulfill a number of properties that make them appropriate for the summarization problem. We focus our research in the use of the F^A model [2, 4]. The F^A model is a (non-standard) *Determiner Fuzzification Scheme*[1] (DFS) [6] that has a very solid theoretical behavior and is much more plausible than most of the models defined in the literature. Moreover, the model fulfills other adequacy criteria that makes it very useful for practical applications. As a key point, the fuzzy operators induced by the model are the product tnorm and the probabilistic sum tconorm. This fact makes the F^A model essentially different from the standard models defined in [6].

We can identify two main issues in summarizing data with fuzzy quantifiers. The first one is to discover the best set of quantified expressions that is capable to adequately describe a given set of data, and to build summaries based on them. In the second problem, fuzzy quantifiers are used to guide the retrieval or ranking of fuzzy quantified patterns.

In this paper we deal with the first summarization problem. Given a set of data, we analyze the key concepts to build a set of summaries involving quantifiers to describe the data using the F^A quantification model. An algorithm that accomplishes this task is proposed following these criteria.

2 The F^A quantifier fuzzification mechanisms

To overcome Zadeh's framework to fuzzy quantification Glöckner [6] reconsiders the problem of fuzzy quantification as the problem of looking for adequate procedures to transform semi-fuzzy quantifiers (specification means) into fuzzy quantifiers (operational means). Fuzzy quantifiers are just a fuzzy generalization of crisp or classic quantifiers:

Definition 1: Fuzzy Quantifier. [6, pag. 66] An n-ary fuzzy quantifier $\widetilde{Q}$ on a base set $E \neq \varnothing$ is a map-

[1] Determiner fuzzification schemes are fuzzy quantification models fulfilling a very strict axiomatic framework [6], that assures an excellent theoretical behavior. Standard DFSs induce standard fuzzy operators (*min* tnorm and *max* tconorm).

ping $\widetilde{Q} : \widetilde{P}(E)^n \longrightarrow \mathbf{I} = [\mathbf{0}, \mathbf{1}]$. Here $\widetilde{P}(E)$ denotes the fuzzy powerset of E.

A fuzzy quantifier assigns a gradual result to each choice of $X_1, \ldots, X_n \in \widetilde{P}(E)$. An example of a fuzzy quantifier could be $\widetilde{\mathbf{all}} : \widetilde{P}(E)^2 \longrightarrow \mathbf{I}$ and a reasonable definition for it: $\widetilde{\mathbf{all}}(X_1, X_2) = \inf\{\max(1 - \mu_{X_1}(e), \mu_{X_2}(e)) : e \in E\}$.

Although a certain consensus may be achieved to accept this previous expression as a suitable definition for $\widetilde{\mathbf{all}}$ this is not the unique one. The problem of establishing consistent fuzzy definitions for quantifiers (e.g., *"at least eighty percent"*) is faced in [6] by introducing the concept of semi-fuzzy quantifiers. A semi-fuzzy quantifier represents a medium point between classic quantifiers and fuzzy quantifiers, and it is close but is far more general than the idea of Zadeh's linguistic quantifiers [10].

Definition 2: Semi-fuzzy Quantifier. [6, pag. 71] An n-ary semi-fuzzy quantifier Q on a base set $E \neq \varnothing$ is a mapping $Q : P(E)^n \longrightarrow \mathbf{I}$.

Q assigns a gradual result to each pair of crisp sets $(Y_1, \ldots, Y_n)$.

An example of a semi-fuzzy quantifiers with the meaning of *"at least about 80%"* is:

$$ab{\geq}\mathbf{80\%}(\mathbf{Y_1}, \mathbf{Y_2}) = \begin{cases} S_{0.5,0.8}\left(\frac{|Y_1 \cap Y_2|}{|Y_1|}\right) & Y_1 \neq \varnothing \\ 1 & Y_1 = \varnothing \end{cases}$$

where $S_{0.5,0.8}(x)$ is a Zadeh's S function.

In order to evaluate fuzzy quantified sentences mechanisms that are capable to transform semi-fuzzy quantifiers into fuzzy quantifiers are needed, i.e., mappings with domain in the universe of semi-fuzzy quantifiers and range in the universe of fuzzy quantifiers. Glöckner names these mechanisms *quantifier fuzzification mechanisms.*

Definition 3: [6, pag. 74]A quantifier fuzzification mechanism (QFM) F assigns to each semi-fuzzy quantifier $Q : P(E)^n \to \mathbf{I}$ a corresponding fuzzy quantifier $F(Q) : \widetilde{P}(E)^n \to \mathbf{I}$ of the same arity $n \in \mathbb{N}$ and on the same base set.

In [2, 4] the finite QFM F^A is proposed. This model is based on a probabilistic interpretation of fuzzy sets, although it can also be interpreted in a fuzzy way [2, chapter three].

Under the interpretation of fuzzy sets used in the F^A model, membership grades are interpreted as probabilities, and independence between probabilities of different elements is assumed.

Definition 4: F^A. Let $Q : P(E)^n \to \mathbf{I}$ be a semi-fuzzy quantifier, E finite. The QFM F^A is defined as

$$\begin{aligned} &F^A(Q)(X_1, \ldots, X_n) \qquad (1)\\ &= \sum_{Y_1 \in P(E)} \cdots \sum_{Y_n \in P(E)} m_{X_1}(Y_1) \ldots m_{X_n}(Y_n)\, Q(Y_1, \ldots, Y_n) \end{aligned}$$

for all $X_1, \ldots, X_n \in \widetilde{P}(E)$, where $m_X(Y) = \prod_{e \in Y} \mu_X(e) \prod_{e \in E \setminus Y} (1 - \mu_X(e))$.

This model is a finite DFS [2], and to our knowledge the unique non-standard known DFS. Moreover, the theoretical behaviour of the model is excellent [2, chapter three], fulfilling some properties that makes it very adequate for applications. In [4, 2] we have tested the behaviour of the F^A and other quantification models in the information retrieval basic task , providing that the F^A model can work in realistic scenarios.

3 Ruspini quantified partitions and probabilistic QFMs

Let us suppose we use a set of semi-fuzzy quantifiers (*"nearly none"*, *"a few"*, *"several"*, *"many"* and *"nearly all"*) to split the quantification universe. If semi-fuzzy quantifiers can be interpreted in a probabilistic way; that is, for a certain proportion x, $\sum_{l_i \in L} l_i(x) = 1$ where the $l_i s$ are the fuzzy numbers used in the definition of the quantifiers, then the F^A model allows us to interpret fuzzy quantifiers in a probabilistic way.

Definition 5: A set of semi-fuzzy quantifiers $Q_1, \ldots, Q_r : P^n(E) \to \mathbf{I}$ forms a probabilistic Ruspini partition of the quantification universe if for all $Y_1, \ldots, Y_n \in P(E)$ it holds that $Q_1(Y_1, \ldots, Y_n) + \ldots + Q_r(Y_1, \ldots, Y_n) = 1$.

In figure 1 we show a Ruspini quantified partition of the quantification universe.

The application of the F^A model on a Ruspini partition of quantifiers can be interpreted as a probability, due to the following property, which is fulfilled by this and other probabilistic models [2, chapter three].

Definition 6: [2] A QFM F^A fulfills the property of probabilistic interpretation of quantifiers if for all probabilistic Ruspini partitions of the quantification universe $Q_1, \ldots, Q_r : P(E)^n \to \mathbf{I}$ it holds that[2]

$$F(Q_1)(X_1, \ldots, X_n) + \ldots + F(Q_r)(X_1, \ldots, X_n) = 1$$

[2] This property allows us to use the entropy to measure the dispersion of the data for the fuzzy quantifiers. If the entropy is low then most of the probability is concentrated in a small set of quantifiers that describes very well the data (e.g., *"many temperatures are low in January"*). If the entropy is high then the probability is distributed between several quantifiers.

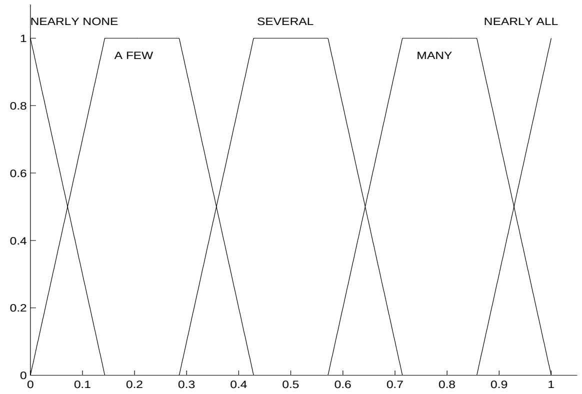

Figure 1: Fuzzy quantified proportional Ruspini partition.

Moreover, for an easy definition of summarization algorithms, we will use Ruspini quantified partitions fulfilling the next symmetry relation: $l_i(x) = l_{w-i-1}(1-x)$. In this way, the Ruspini quantified partition is closed for the antonym property; e.g., *"many"* is the antonym of *"a few"* and *"nearly all"* is the antonym of *"nearly none"*. Thinking in crisp data, if we know *"many data are A"* then we also know *"few data are not A"*, that is also a quantifier of the Ruspini partition. In the same way, if we know *"many data are related to label l"*, we also know the rest of data are associated to quantifiers *"nearly none"* or *"a few"*.

Probabilistic interpretation of quantifiers also allows us to merge quantifiers in a very intuitive way. In figure 2 an hypothetical example of the process to merge quantifiers is presented. We are using a very simple result: the evaluation of the combination of two quantifiers is the sum of the evaluation of the individual quantifiers.

4 Using quantifiers in data summarization

Different kinds of fuzzy quantified summaries can be found in the literature [7, 8, 9], generally linking the building of summaries to specific applications. We can classify them into two different summarization problems:

Explaining the data with fuzzy quantifiers. The problem we deal with in this paper. For a given set of data, we look for procedures that allow us to obtain a group of quantifiers that describes the set of data conveniently. Summaries are associated to these quantifiers.

Quantified pattern detection. In this case, we have some knowledge about the quantification patterns that are of interest and can be found in the data. Data are analyzed in search of these patterns and results are ranked using evaluation results of quantified patterns and other criteria as the amount of data covered by the patterns.

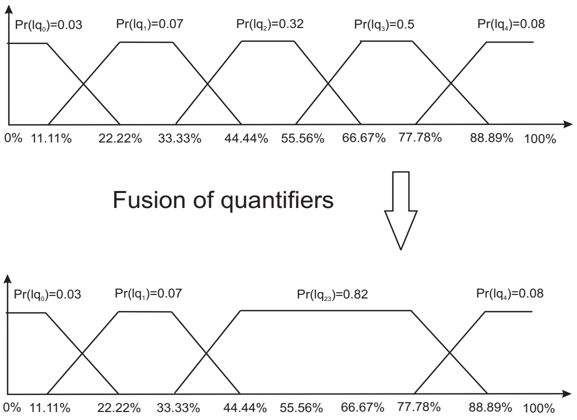

Figure 2: Example of the process of merging of quantifiers. In the bottom of the figure the semi-fuzzy quantifier related to the label lq_{23} has associated the 82% of the probability.

Summaries in the first problem are related with the issue of learning the best set of quantifiers to describe the data. By means of quantifiers, we adapt the data to a granularity level that is reasonable for human consumption. That is, we explain the distribution of the data by means of fuzzy quantifiers. In the second problem, a rough idea on the kind of summaries we are searching for in the data is the starting point. Fuzzy quantification is needed for the evaluation of the quantification patterns. Other criteria could be needed to select or rank the interesting summaries, as fuzzy quantification models are only used to analyze the validity of the patterns.

Generating a summary as *"many temperatures are hot but a few are warm"* for a given daily temperature data set is an example of the first problem, as quantified expressions are used to explain the amount of data associated to each linguistic label (i.e., *cold, warm* and *hot*).

The second problem is generally application dependent. For example, it could be interesting in a given application to evaluate the relationship between the number of people injured in traffic accidents and the part of the day they occur, thus generating summaries as *"most of the summer days, number of people injured in the morning was higher than in the afternoon"*. Here, *most* is a fuzzy quantifier and *higher* a fuzzy comparator. And possibly to compare previous pattern with another ones *"several summer days,* *number of people injured in the morning was less than in the afternoon"*

5 Some principles for building quantified summaries

Let us consider a base set E, and let $L = \{l_0, \ldots, l_{v-1}\}$ be a linguistic variable defined on a referential universe associated to some particular property of the elements of E (i.e., E could be formed by a set of days and L a linguistic variable defined on the universe of temperatures). Let $F_Q = \left\{f_{Q_0}, \ldots, f_{Q_{W-1}}\right\}$ be a fuzzy quantified partition of the proportional universe $[0, \ldots, 100\%]$ (i.e., *"nearly none"*, *"a few"*, *"several"*, *"many"*, *"nearly all"*). Then the objective is to find a set of quantified expression of the form $Q_i\ Es$ are l_j that conveniently summarizes the data.

Generated summaries should be adequate for human consumption. Following data mining principles and conversational principles[3] [5, p. 204] we establish a set of criteria to guide the summarization process:

Quality of the summaries. Degree of fulfillment of the summaries should be high. Moreover, well behaved fuzzy quantification models are needed in order to guarantee the quality of the results.

Compatibility of the summaries. A summary composed of the following expressions: *"nearly all the days are hot"* and *"nearly all the days are warm"* is not adequate to human consumption. A good summary should be composed of "compatible quantifiers" covering approximately 100 percent of the data set.

Unambiguity of the summaries. Ambiguous summaries as *"between 25% and 75% of the days are hot"* and *"between 50% and 90% of the days are warm"* are not adequate. It should be noted these two summaries are compatible, but ambiguous.

Minimization of the number of summaries. Summaries consisting of a small number of expressions should be preferred. "Fusion of quantifiers" and other techniques could be employed to increase the quality of the summaries. Moreover, following data mining techniques, summaries explaining a large set of data are more adequate.

6 Building summaries with the F^A model

In this section we propose a greedy algorithm for building summaries using fuzzy quantifiers, following the principles in the previous section.

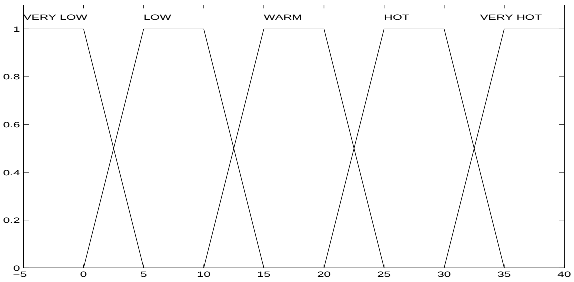

Figure 3: "Fuzzy Temperatures" linguistic variable, in ^{o}C.

Months	April				
Labels\Quantifiers	nn	f	s	m	na
very low	1	0	0	0	0
low	0	0.72	0.28	0	0
warm	0	0	0.28	0.72	0
hot	1	0	0	0	0
very hot	1	0	0	0	0

Table 1: Evaluation of the quantifiers for the different temperature labels in April.

Let us consider a linguistic variable associated to a certain variable of the data (e.g., *"fuzzy temperature"* in figure 3) and a Ruspini unary proportional quantified partition of the quantification universe (e.g., the one presented in figure 1). Then, a way of summarizing the data is to evaluate the combinations between linguistic labels and linguistic quantifiers. Pairs of label/quantifiers with a high degree of fulfillment are adequate candidates to generate summaries. We will present this idea with an example.

Let us consider a temperature data set (in our experimentation, monthly temperatures in Madrid in January, April and June 2006) (figure 4).

When applying the F^A model to all the possible combinations of temperature labels and quantifiers the results presented in table 1 are obtained. We have evaluated $F^A(Q_i)(lab_j(temperatures))$ for each possible (i, j), *"nn=nearly none, f=a few, s=several, m=many* and *na=nearly all"*.

The results shown in table 1 are intuitive. E.g., we find in April that *"many temperatures are warm"* and *"few temperatures are low"* are respectively the most adequate summaries using quantifiers few and $many$. Previous matrix is a good summary of the fuzzy set, that adapts the data to the granularities of the fuzzy linguistic variable *temperature* and *fuzzy quantified partition.*

[3] We have taken into account *Conversational Maxims of Grice's theory.*

Now we sketch the main ideas of an algorithm to transform previous matrix in a fuzzy quantified summary. For simplicity, a greedy strategy is described.

We firstly introduce an index to measure the adequacy of a given series of quantifiers $Qs = \{Q_i, \ldots, Q_h\}$ to summarize a set of data. Definition is based on a sequence or series of quantifiers because consecutive quantifiers of high probability will be merged. We need two previous definitions:

Definition: Specificity of a quantifier. Let $Q : P(E) \rightarrow \mathbf{I}$ be a unary proportional semi-fuzzy quantifier defined by means of a proportional fuzzy number $fn : [0,1] \rightarrow \mathbf{I}$. We define the specificity of Q as: $Spec(Q) = \int_0^1 fn(x)\,dx$. Specificity of a quantifier allows us to compare different quantifiers of a Ruspini quantified partition. In general, given two quantifiers of similar probabilities, the most specific one is preferred to summarize a set of data[4].

Definition: Threshold to merge quantifiers: When we found a consecutive series of quantifiers $Qs = \{Q_i, \ldots, Q_h\}$ with a high probability it could be convenient to merge the quantifiers in order to generate a single quantified expression instead of several ones. We will use a threshold to make the decision of merging a series of quantifiers depending on its probability: $u\_fusion = Spec(q_i)^h$. In this manner, less specific (or wider) quantifiers requires a greater probability to be merged with other quantifiers. The $h \leq 1$ parameter allows us to adjust the probability required to merge the quantifiers.

Definition: Adequacy K index of a $Qs = \{Q_i, \ldots, Q_h\}$ series of quantifiers. Let $Qs = \{Q_i, \ldots, Q_h\}$ be a given series of quantifiers, then we will use the next index K to measure the adequacy of the series of quantifiers to be an adequate summary of a data set:

$$K(Qs) = \sum_{i \leq j \leq h} \Pr(Q_j) \cdot (j-1) \cdot \frac{1}{(Spec(Q_j))}$$

In previous expression:

- $\Pr(Q_j)$ is the probability of the quantifier. In this way, we look for sequences with a high probability of fulfillment.
- $j - 1$ is used to promote quantifiers covering a significant amount of data, following data mining principles. Quantifier *"nearly all"*, that covers a significant amount of data, is preferred to the quantifier *"a few"* that does not cover too much data.
- $\frac{1}{(Spec(Q_j))}$ is used to promote more specific quantifiers.

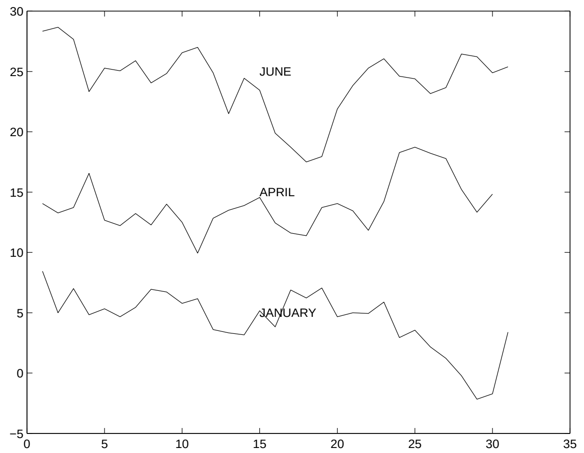

Figure 4: Average daily temperatures in January, April and June of 2006 in Madrid.

In this way, the K index would be high for Qs covering a large set of data, with a high probability and a low specificity.

The idea underlying the algorithm is quite simplistic. First we compute a matrix indicating which quantifiers could possibly be merged. Then we look for the best quantifier (or series of quantifiers) to summarize the data (the one that maximizes the K index) and generate the corresponding expression. Then, we repeat the process. In order to generate more understandable expressions, previously used linguistic labels will not be repeated (for human consumption, quantified expressions *"nearly all are hot"* and *"a few are hot"* are incompatible). We also discard generating expressions covering more than 100% of the data. If we interpret the quantifiers as fuzzy numbers, we do not want the sum of the quantifiers associated to the summaries to be greater than 100% (i.e, a summary of data consisting of the expressions *"nearly all are hot"* and *"nearly all are warm"* are clearly not adequate for human consumption). To prevent this, we stop the generation of summaries when the generation of a new summary implies covering more than 100% of the data.

Based on these considerations, we propose the algorithm shown in table 2 to build a quantified summary of a set of data.

Some of the results we have obtained with the previous algorithm for the mentioned temperature data set are:

[4] Here, specificity of a quantifier is not defined in the same way that specificity of a fuzzy set. The idea of the specificity is to measure the amount of possible proportions covered by a proportional quantifier defined by means of a fuzzy number in $[0, 1]$.

```
Computing a quantified linguistic summary
$FL = \{l_0, \ldots, l_{v-1}\}$
      a linguistic variable
$QS = \{Q_0, \ldots, Q_{w-1}\}$
      a Ruspini quantified partition
$R\,[i,j] = F^A\left(l_i, QS_j\right)$,
      $i = 0, \ldots, v-1, j = 0, \ldots, w-1$
$Fusion\,[0 : v-1, 0 : w-1] =$
      Possible fusions of quantifiers
$while$ we do not cover enough data
      look for the best series of fusionable
      quantifiers $Qs = \{Q_i, \ldots, Q_h\}$
            without repeating linguistic labels
            without the summaries surpass 100%.
      Generate a quantified expression for $Qs$
end while
```

Table 2: Algorithm for generating a quantified summary of a set of data

January: *"Many temperatures are low", "a few temperatures are very low"*

April: *"Many temperatures are warm", "a few temperatures are low"*

June: *"Several or many temperatures are hot"*, *"a few temperatures are warm"*.

For example, in the case of January previous algorithm generates a summary indicating that *"many temperatures were low"*. Daily temperatures for which the *"low"* description is not adequate were only *"a few"* and are associated to the label *"very low"*. We want to emphasize the simplicity and intuitiveness of generated summaries, very adequate for human consumption.

6.1 Conclusions and future work

In this work we have presented an initial proposal to build linguistic quantified summaries with probabilistic quantification models. Given a particular data set, we have studied how to build a set of fuzzy quantified expressions for summarizing the data, convenient for human consumption.

Several relevant open questions still remain in relation to the proposal. Other criteria and algorithms to guide the building of summaries should be studied. Moreover, our approach deals with unary proportional quantified expressions. Using other quantifiers have not been studied, although a similar approach could be used to generate summaries with binary proportional quantifiers. In *"Q of the $\underline{l}$ man are $\underline{r}$ paid"*, by fixing the label associated to the first variable "man" (e.g. $l = young$) a similar procedure could be used to compute a quantified set of expressions explaing how many man are "*$\underline{r}$ paid*" given they are young.

Another important question is to make a deep study of the second problem of summarizing with fuzzy quantifiers (detect the most interesting set of quantified patterns to summarize particular events).

References

[1] M. Delgado, D. Sánchez, and M. A. Vila. Fuzzy cardinality based evaluation of quantified sentences. *International Journal of Approximate Reasoning*, 23(1):23–66, 2000.

[2] F. Díaz-Hermida. *Modelos de cuantificacón borrosa basados en una interpretación probabilística y su aplicación en recuperación de información.* PhD thesis, Universidad de Santiago de Compostela, 2006.

[3] F. Díaz-Hermida, A. Bugarín, P. Cariñena, and S. Barro. Voting model based evaluation of fuzzy quantified sentences: a general framework. *Fuzzy Sets and Systems*, 146:97–120, 2004.

[4] F. Díaz-Hermida, David. E. Losada, A. Bugarín, and S. Barro. A probabilistic quantifier fuzzification mechanism: The model and its evaluation for information retrieval. *IEEE Transactions on Fuzzy Systems*, 13(1):688–700, 2005.

[5] L.F.T. Gamut. *Logic, Language and Meaning*, volume II of *Logic, Language, and Meaning.* The University of Chicago Press, 1984.

[6] I. Glöckner. *Fuzzy Quantifiers: A Computational Theory.* Springer, 2006.

[7] I. Glöckner. Optimal selection of proportional bounding quantifiers in linguistic data summarization. In *Proceedings of SMPS2006 - Special Session on Applications and Modelling of Imprecise Operators, Published in Soft Methods for Integrated Uncertainty Modelling, LNCS, Springer*, pages 173–181, 2006.

[8] A. Wilbik J. Kacprzyk and S. Zadrozny. Linguistic summarization of time series using a fuzzy quantifier driven aggregation. *Fuzzy Sets and Systems*, 159:1485–1499, 2008.

[9] R.R. Yager. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Transactions on Systems, Man and Cybernetics*, 18(1):183–191, 1988.

[10] L.A. Zadeh. A computational approach to fuzzy quantifiers in natural languages. *Comp. and Machs. with Appls.*, 8:149–184, 1983.