Natural Language Generation of Explanations of Fuzzy Inference Decisions

Ismaïl Baaj, Jean-Philippe Poli

Natural Language Generation of Explanations of Fuzzy Inference Decisions

Ismaïl Baaj and Jean-Philippe Poli
CEA, LIST
91191 Gif-sur-Yvette cedex, France.

Abstract

As Artificial Intelligence and fuzzy systems are at the center of the emergence of advanced technologies such as autonomous vehicles or medical decision support systems, a problem of trust from a human point of view is strongly appearing. In this article, we tackle the problem of explanation of a fuzzy inference system decision in its entirety: from the conception of an algorithm that produces a textual explanation to its evaluation.

We define a function which is able to associate to any activated fuzzy rule, the structure responsible of its activation degree. To assess our method, we defined a protocol to evaluate AI-generated explanation, and made an experiment: explanations obtained from the classification of pastas. Despite limitations, the results show a good transparency of the reasoning, consistency and good global effectiveness in generated explanations.

Keywords: Explainable Artificial Intelligence, Fuzzy Inference System

1 Introduction

In our daily life, Artificial Intelligence plays a significant role in our choices and in our actions. Intelligent systems are able to perceive, represent and decide almost instantly with a very high accuracy. The models are often so complex that humans may have difficulties understanding their internal behaviors: it highlights a problem of trust in the system from a human point of view, especially when their control can only be exercised *a posteriori*. Indeed, when Artificial Intelligence is used in risky environments in which respect for human life can be engaged, it must be able to explain and inform its decisions, assumptions and reasoning. Ideally, a form of symmetrical communication may arise between intelligent systems and humans, whether they are experts or not. The interaction between the system and the end users can take advantage of the progress made in natural language processing: intelligent systems need to offer a dialog in which they can describe their reasoning without a lack of precision, and be able to explain each concept they use in a human understandable way.

Fuzzy Inference System (FIS) seems to be a good candidate for the explainable Artificial Intelligence field [1], as fuzzy logic is able to deal with the vagueness of natural language and to reason under uncertainty. Moreover, Zadeh introduced with fuzzy logic the methodology for "Computing with words" (CWW) [2].

However, major difficulties exist when it comes to produce simple, general, coherent and accurate explanations of the reasoning operated by fuzzy systems. On the one hand, from the social sciences, explanation is a very complex set of statements used by humans to communicate [3]. On the other hand, three limits pointed by Zadeh remain for CWW: words are less precise than numbers, precision carries a cost, and if numbers are respected, words are not [4]. Nonetheless, the need for an explanatory capacity for fuzzy rule-based system is growing. With the adoption of a right to explanation in Europe and the XAI program launched by DARPA, the research community actively focus on them. Fuzzy inference engines contain a lot of information, e.g. activated rules and their associated data, which can be very helpful to build the explanation of a decision made by the system. To provide such a functionality, many questions arise: How to express a result ? What would be the form and the content of the explanation ? How to make accurate associations of precise linguistic terms with situations presenting an uncertainty ? How to evaluate the quality of an explanation ?

In this paper, we will strive to answer these questions, notably by presenting the challenges related to fuzzy inference explanation in section 2. We will then demonstrate via a proof of concept that it is possible to produce textual explanation in natural language from it (section 3). We also faced the open problem of the evaluation of AI-generated explanations (section 4) which leads to the development of an evaluation protocol along with an evaluation use case. The results of this evaluation is presented at the end of section 4 and we conclude with some research perspectives.

2 Background

Firstly discussed in the 1970's with the MYCIN experiments [5], the production of explanations for expert systems has emerged in the 80's. The main goal was the user acceptance of the conclusions delivered by intelligent systems. Three specific types of explanations has been distinguished: inference process, elements from the knowledge base and the operated strategy [6]. For fuzzy expert systems, an explanatory capacity is currently at a very experimental stage. Researchers have tried to interpret a Mamdani distribution composed of trapezoid fuzzy sets [7], have created a link between Linguistic Description of Data (LDD) and Natural Language Generation (NLG) [8] and have emphasized that users prefer decisions accompanied by an explanation in the context of a fuzzy expert system [9]. Starting from these researches, we first focus on how to enunciate a fuzzy expert system decision and how to clearly identify its responsible causes.

2.1 Result enunciation

The utterance of such result may consist in enumerating possibilities in natural language, notably if they are uncertain, ambiguous or under any numerical representation. Even if it is known that humans have difficulties when they have to deal with numbers [10], it appears that the use of statements in natural language to designate them is not an easy task, as demonstrated for instance in risk management [11] or medicine [12]. Researchers have established many uncertainty scales that match a precise vocabulary and a probability, like the one for intelligence agencies [13] or the one applied to climatology (Table 1).

Phrase	**Likelihood of occurrence/outcome**
Virtually certain	>99%
Very likely	>90%
Likely	>66%
About as likely as not	33% to 66%
Unlikely	<33%
Very unlikely	<10%
Exceptionally unlikely	<1%

Table 1: IPCC uncertainty scale [14]

The use of such scales can be useful for fuzzy logic: indeed, it is possible to get linguistic characterization of fuzzy sets (e.g. with linguistic modifiers [15]) or of numeric data inputs (e.g. with protoforms [16]). A vocabulary accurate enough is necessary to extract relevant information: researchers suggest terms such as “very”, “more or less”, “completely”, “quite”, “fairly”, “extremely” or “somewhat” when qualifiers for fuzzy sets were introduced [17], but some of them can lead to ambiguities. For instance, it is not easy to distinguish which is the most preponderant between “more or less” or “somewhat”.

2.2 Justification extraction

The justification of a result inferred by a fuzzy system is strongly related to the activation of the rules, which are not sufficient to explain the result. Indeed, the difficulty also resides in the interdependence of the different conclusions that concerns the same output, in particular when the rules are complex.

The simplest solution is to aggregate the different activated rules into an explanation. However, regarding a given output, rules may contain irrelevant antecedents or superfluous content that users would not expect in an explanation. Moreover the explanation thus formed could contain contradictions of the kind “it is A and not A”, that are not always senseless contradiction as pointed by Sauerland [18]. Furthermore, as fuzzy systems are not interpretable by nature [19], the system modelling is determinant.

However, as dense and rich are fuzzy logic and fuzzy systems, all of these issues are still open to make fuzzy expert systems explainable. In this paper, we present a first attempt to some of them.

3 Textual explanation generation

In this work, we focus on the explanation of classification results by a fuzzy expert system. In particular, we used the Mamdani inference with the majority aggregation. It takes as input the rule base R and the execution trace that contains all the activated fuzzy rules and their related data. Three successive steps are needed (Figure 1) to produce an explanation in natural language:

- *Justification extraction*: creates links between the conclusions and the fuzzy rule base;
- *Explanation formatting*: selects the relevant links within those created before;
- *Text Generation*: adds qualifiers to any information and render a textual explanation.

Figure 1: Structure of the explanation generation process

Before presenting these steps in detail, we first define some notations which are helpful to understand this section.

3.1 Notations

A fuzzy IF-THEN rule $r = (p, c)$ is composed of a premise p and a conclusion c. Its activation degree related to the computation of p is noted α_r. For the sake of simplicity, we denote α_e the fuzzy value of the expression e. Thus, in the Mamdani system we target, $\alpha_r = \alpha_p$.

Given a finite set of pairs formed by a linguistic variable and one of its terms (fuzzy sets):

$$\mathcal{V} = \{(v, A), \text{ with } v \text{ a linguistic variable and } A \text{ in } T_v\}$$

where T_v is the set of terms of v.

For the premises of the rules, let:

- $e = (v, A) \in \mathcal{V}$ denote a fuzzy proposition;
- $(e, \neg)$ be the negation of a fuzzy expression e;
- $(e_1, e_2, \circ)$ be a binary expression, with e_1, e_2 fuzzy expressions and $\circ$ a logical operator AND (t-norm) or OR (t-conorm) also noted $\wedge$ and $\vee$ respectively.

We remind that a conclusion c of a fuzzy IF-THEN rule $r = (p, c)$ in a Mamdani system is of the form $c = (o, A)$ composed of an output linguistic variable o and one of its terms A (such as $A \in T_o$). We define by C_o the set of conclusions used in R which use the output linguistic variable o. For each conclusion $c \in C_o$, let R_c be defined as $R_c = \{r = (p, c)\}$, i.e. the set of rules whose conclusion is c. Let α_c^* be the highest activation degree among all the activation degrees α_r for each rule $r = (p, c) \in R_c$.

3.2 Justification extraction

We define in this section a function $\mathscr{R}$ that reduces a premise p of a fuzzy rule in the way that it only contains the elements responsible of its activation degree denoted here α_p, asserting $\alpha_{\mathscr{R}(p)} = \alpha_p = \alpha_r$.
$\mathscr{R}$ keeps the premise as it is for the simplest form of a premise, in other words, a fuzzy proposition and a negation of a fuzzy proposition. Let $e = (v, A) \in \mathcal{V}$ a fuzzy proposition:

$$\mathscr{R}(e) = e$$
$$\mathscr{R}((e, \neg)) = (e, \neg)$$

For a conjunction, for instance $e = (e_1, e_2, \wedge)$, where e_1 and e_2 are fuzzy expressions, $\mathscr{R}$ reduces e when one of the two operands has an activation degree equal to zero:

$$\mathscr{R}(e) = \begin{cases} \mathscr{R}(e_1) \text{ if } \alpha_{e_1} = 0 \text{ and } \alpha_{e_2} \neq 0 \\ \mathscr{R}(e_2) \text{ if } \alpha_{e_1} \neq 0 \text{ and } \alpha_{e_2} = 0 \\ (\mathscr{R}(e_1), \mathscr{R}(e_2), \wedge) \text{ otherwise.} \end{cases}$$

For a disjunction $e = (e_1, e_2, \vee)$, a threshold of reduction $Th \in [0, 1]$ that can be arbitrarily $Th = 0.75$ is needed and determines if one expression that composes e is insignificant regarding the value $\Delta =\mid \alpha_{e_1} - \alpha_{e_2} \mid$. $\mathscr{R}$ also takes into account when $\alpha_{e_1} = 0$, $\alpha_{e_2} = 0$ or both.

$$\mathscr{R}(e) = \begin{cases} (\mathscr{R}(e_1), \mathscr{R}(e_2), \wedge) \text{ if } \alpha_{e_1} = 0 \text{ and } \alpha_{e_2} = 0 \\ \mathscr{R}(e_1) \text{ if } \alpha_{e_1} \neq 0 \text{ and } \alpha_{e_2} = 0 \\ \mathscr{R}(e_2) \text{ if } \alpha_{e_1} = 0 \text{ and } \alpha_{e_2} \neq 0 \\ \mathscr{R}(e_1) \text{ if } \Delta \geq T_h \text{ and } \alpha_{e_1} > \alpha_{e_2} \\ \mathscr{R}(e_2) \text{ if } \Delta \geq T_h \text{ and } \alpha_{e_2} > \alpha_{e_1} \\ (\mathscr{R}(e_1), \mathscr{R}(e_2), \vee) \text{ otherwise.} \end{cases}$$

The reader may notice that we changed the operator when $\alpha_{e_1} = 0$ and $\alpha_{e_2} = 0$ without taking the negation of its operands: it will be taken into consideration in the text generation phase.

The reduction of a negation of premise $(e, \neg)$ requires to satisfy $\alpha_{\mathscr{R}(((e,\neg),\neg))} = \alpha_{\mathscr{R}(e)}$. As we have previously dealt when e is a fuzzy proposition, we suppose that $e = (e_1, e_2, \circ)$ is either a conjunction or a disjunction here. By creating $e'_1 = (e_1, \neg)$ and $e'_2 = (e_2, \neg)$, and the value $\Delta =\mid \alpha_{e_1} - \alpha_{e_2} \mid$, we can reduce e when $\circ$ is the AND operator:

$$\mathscr{R}((e,\neg))=\begin{cases}(\mathscr{R}(e'_1),\mathscr{R}(e'_2),\wedge) \text{ if } \alpha_{e_1}=1 \text{ and } \alpha_{e_2}=1\\ \mathscr{R}(e'_1) \text{ if } \alpha_{e_1}\neq 1 \text{ and } \alpha_{e_2}=1\\ \mathscr{R}(e'_2) \text{ if } \alpha_{e_1}=1 \text{ and } \alpha_{e_2}\neq 1\\ \mathscr{R}(e'_1) \text{ if } \Delta\geq T_h \text{ and } \alpha_{e_1}<\alpha_{e_2}\\ \mathscr{R}(e'_2) \text{ if } \Delta\geq T_h \text{ and } \alpha_{e_2}<\alpha_{e_1}\\ (\mathscr{R}(e'_1),\mathscr{R}(e'_2),\vee) \text{ otherwise.}\end{cases}$$

And also if $\circ$ is OR:

$$\mathscr{R}((e,\neg))=\begin{cases}\mathscr{R}(e'_1) \text{ if } \alpha_{e_1}=1 \text{ and } \alpha_{e_2}\neq 1\\ \mathscr{R}(e'_2) \text{ if } \alpha_{e_1}\neq 1 \text{ and } \alpha_{e_2}=1\\ (\mathscr{R}(e'_1),\mathscr{R}(e'_2),\wedge) \text{ otherwise.}\end{cases}$$

Now $\mathscr{R}$ has been defined, we can define the *Justification* of a conclusion c as the set of rules with an activation degree equal to α_c^* and their related reduced premise :

$$\text{Justification}(c)=\{(\mathscr{R}(p),c) \mid r{=}(p,c)\in R_c \text{ such as } \alpha_r{=}\alpha_c^*\}$$

3.3 Explanation formatting

Explanation formatting consists of two steps:

- step 1 minimizes and simplifies the Justification(c) of each conclusion c;
- step 2 sorts the conclusions by decreasing relevance.

For the first step, we obtain for any *Justification* of a conclusion c an expression E_c by joining the premises of the minimized set of fuzzy rules of Justification(c) into a single sentence with the co-ordinating conjunction "and". The minimization method is described in [20] and is applied to a fuzzy rules set. In this algorithm, each fuzzy membership predicate (i.e. a linguistic term) accompanied with proper constraints is used to obtain a minimized set of fuzzy rules by factorizing. It is important to notice that the justifications do not contain any negation of binary expression, with the help of $\mathscr{R}$. They are only composed of fuzzy propositions, negations of fuzzy proposition, $\wedge$ and $\vee$. To avoid the use of negation as much as we can, we replace the negation of a fuzzy proposition $((v,A),\neg)$ with an activation degree equal to 0 with its related fuzzy proposition (v,A).

For the second step, we suggest that an explanation needs to enunciate conclusions from the most possible to the least to preserve consistency. Thus, the sort is performed regarding the activation degrees α_c^* for all the conclusions c, in descending order.

3.4 Text generation

The goal is to enrich the explanation with useful characteristics of the situation, by finding for each conclusion c:

- an accurate linguistic terminology regarding α_c^*;
- for each fuzzy proposition (v, A) in E_c a qualifier (e.g. the one suggested in table 1) regarding the value of the membership function μ_A of A applied on the input;
- for each negation of fuzzy proposition $((u, B), \neg)$ a qualifier regarding the negation of the value of the membership function μ_B of B applied on the input.

In our opinion, the main drawback of the IPCC uncertainty scale (see table 1) is the lack of qualifier when there is no doubt. We thus adapted the scale to our work by adding three qualifiers:

- *definitely* when the value is equal to 1 (100%), that can be applied to any piece of justification;
- *not* when the value is equal to 0, that can be applied to fuzzy propositions only;
- *impossible* when the value is equal to 0, that can be applied on conclusions only.

Finally, the textual explanation of an output linguistic variable o is based on each conclusion $c \in C_o$ sorted by decreasing α_c^* and decorated with their accurate linguistic terminology and their associated expression E_c. In order to produce text in natural language in a consistent form in terms of morphology, grammar and conjugation, we use a realization engine called SimpleNLG [21].

4 Evaluation of explanations

The evaluation of an explanation requires the definition of a test protocol and of criteria that characterize a good explanatory capacity, and therefore the quality of the explanations [22].

Presently, to our knowledge, there is no consensus around a good methodology for evaluating explanations in the literature despite it is an active research topic. However, concerning the evaluation, researchers have identified some properties [23], some techniques [24] and criteria in particular in recommender systems [25]. Some experiments have also been conducted on expert systems, either in terms of performances [26] or of user acceptance [27, 28]. Recently, [9] emphasizes the benefits of this functionality for fuzzy expert systems, by broadcasting a web survey to evaluate the accompaniment of a decision by an explanation.

From the literature, we can distinguish three main topics to characterize an explanation:

- Natural language: its evaluation resides nowadays in the evaluation of a NLG layer like suggested by Alonso [9]. Reiter and Belz described ways to assess the quality of text produced by NLG systems [29], but in the

case of explanations, many choices like the tense need also to be studied.

- Human-Computer Interaction: after reading these explanations, the interaction between the human and the intelligent system and their relation may evolve. For instance, the user can change his own opinion if he is convinced by the system.
- Content and form need strong features from the way human use explanations to communicate. From a social sciences point of view, Tim Miller argues that the most important criteria are probability, simplicity, generality and coherence with prior beliefs [3]. Except for the coherence that has been defined by Thagard [30], criteria do not have a well developed formalism.

Based on this literature, we designed a protocol to evaluate the explanations produced by our algorithm.

4.1 Evaluation protocol

In our opinion, the protocol of evaluation of explanations must be a survey questionnaire destined to expert and non-expert people with precise issues related to the quality of explanation. The data used as example should be explanations from decisions took after reasoning on a problem that most of the people can understand. Indeed, this guarantees that the surveyed people can be a very large public, and it avoids some biases in the results, such as the response bias [31].

Based on the previous properties, we built a survey questionnaire with 17 facts (Table 2) assessed with a Likert scale. We added a comment part for each, like suggested by Moore who declared that they are often the most interesting parts to understand the frustrations of the user with the system, and help to improve explanations [24].

Natural language	Human-Computer Interaction	Content and form
1. Overall, explanations are written in a correct English	4. Explanations are simple to use and easy to read	10. Length of explanations is adequate
2. Conjugation choices are appropriate and adequate	5. Explanations help to make decisions faster than without	11. Explanations are repetitive
3. Grammatical form of sentences is satisfying	6. Explanations let you change your opinion about your expectations	12. It is difficult to read explanations until the end
	7. Explanations help to take good decisions and are convincing	13. Content layout and order of elements in explanations are satisfying
	8. Data and explanations are enough to trust the system	14. All causes are identified in explanations
	9. Explanations express indirectly the way of the system is reasoning	15. Explanations are sufficient in the sense that they do not contain superfluous information and do not miss one
		16. Overall, explanations seem consistent
		17. Explanations are true

Table 2: Facts to be evaluated

In our opinion a relevant test scenario must involve vagueness, must be easy to understand for a large audience, and must be solved by a fuzzy expert system. We choose to develop a system able to recognize types of pasta regarding different features such as length (L), width/diameter (W/D), longitudinal profile (LP), cross section (CS) and surface aspect (SA). We authored rules with the help of a book [32], restricted to 8 pasta types (Table 3) : Bucatini, Capellini, Fusilli, Linguine, Maccheroni, Penne (either short or very short as Pennette exists), Spaghetti and Ziti (either long with Ziti rigati or short). We then collected and measured 37 samples of these pasta to create a proper data set. A part of it has been used to set membership functions and the other has been kept for evaluation. Linguistic variables as length and width/diameter are continuous variables while the other are discrete. We chose not to consider strong fuzzy

partitions. Nevertheless, the intersection between two consecutive fuzzy sets is not empty.

Type	L	W/D	LP	CS	SA
Bucatini	Long	Thin	Straight	Hollow	Smooth
Capellini	Long	Hair thin	Straight	Solid	Smooth
Fusilli	Very short	Large	Twisted	Solid	Smooth
Linguine	Long	Thin	Straight	Solid	Smooth
Maccheroni	Short	Medium	Straight	Hollow	Striated
Penne	Short or very short	Medium	Sheared	Hollow	Striated
Spaghetti	Long	Very thin	Straight	Solid	Smooth
Ziti	Long or short	Medium	Straight	Hollow	Smooth

Table 3: Pasta rule base

4.2 Experiments and results

Two anonymous web survey questionnaires were created: the first targets experts in artificial intelligence and the other targets anonymous people. Each of them display the pictures of each of 8 different pasta before any assessment. Both of them contains three explanations with a content reflecting three different situations:

- a very sure result: "*This pasta is definitely a Fusilli because longitudinal profile is twisted and cross section is solid and surface is smooth and length is definitely very short and width/diameter is definitely large.*".
- an ambiguous situation with two conclusions being possible: "*This pasta is likely a Capellini because longitudinal profile is straight and cross section is solid and surface is smooth and length is likely long and width/diameter is likely hair thin. There is another choice: the pasta can be unlikely a Spaghetti because longitudinal profile is straight and cross section is solid and surface is smooth and length is likely long and width/diameter is unlikely very thin.*".
- one very unlikely result and all the reasons why each other conclusions are impossible: "*This pasta is exceptionally unlikely a Spaghetti because longitudinal profile is straight and cross section is solid and surface is smooth and length is exceptionally unlikely long and width/diameter is definitely very thin. Some conclusions are not possible:*
  - *It's impossible to have a Linguine because width/diameter is not thin.*
  - *It's impossible to have a Capellini because width/diameter is not hair thin.*
  - *It's impossible to have a Bucatini because cross section is not hollow and width/diameter is not thin.*

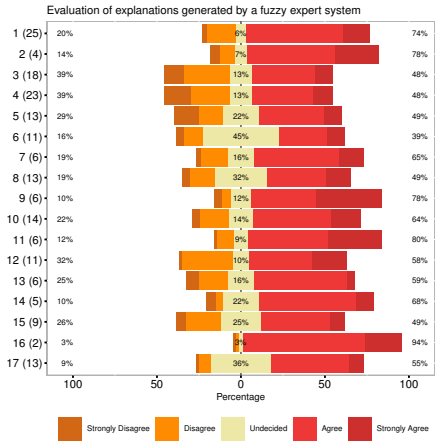

Figure 2: Results of the survey on explanations of pasta classification

- *It's impossible to have a Maccheroni because cross section is not hollow and surface is not striated and length is not short and width/diameter is not medium.*
- *It's impossible to have a Fusilli because longitudinal profile is not twisted and length is not very short and width/diameter is not large.*
- *It's impossible to have a Penne because longitudinal profile is not sheared and cross section is not hollow and surface is not striated and length is not short and length is not very short and width/diameter is not medium.*
- *It's impossible to have a Ziti because cross section is not hollow and width/diameter is not medium.*"

The last explanation allows us to test a particular case: if the system is not confident in its decision, it dismisses all other decisions that are not possible and explains why they are not. The facts presented in table 2 were assessed with a Likert scale of 5 steps (strongly agree, agree, undecided, disagree, strongly disagree).

4.3 Results

In total we got 69 responses with 185 comments. We decided to merge the two studies into one as the first survey had not enough participants (only 9). The results is displayed in Figure 2. Each fact is presented with its index and its related number of comments between parenthesis.

In terms of natural language, users were satisfied with the English (74% approved) and the choice to use the present as conjugation (78% agreed). However, they were more undecided about the grammar as 39% disapproved and 48% approved that it is correct enough. They notably said that the explanations are unusual (too much *and*) but remain understandable. A lack of punctuation has been pointed out, and some morphology mistakes have been noticed (e.g.

food is not used with article *a* before, *can be unlikely* is not correct...). Also, few people find unnatural the qualifiers for uncertainty and the variable named "longitudinal profile". The fact is the algorithm adds systematically a qualifier and thus weigh down the sentences. This point must be improved in a future version.

We also noticed that some users encounter difficulties while reading the explanations (as 39% people disagreed and 48% other agreed with fact 4 and only 49% agreed with fact 5) but all agreed that they strongly express the way our system reasons (78% people agreed with fact 9). Users are sufficiently convinced by explanations (only 18% surveyed disagreed with fact 7) but felt there were not enough cases to fully trust the system (only 49% people agreed with fact 8), i.e. the test seems too short to them. Following the comments, fact 8 and especially fact 6 were sometimes misunderstood and subject to personal interpretation of the question. This is a difficulty when using a survey questionnaire: we did not want to take too much time to the panel. There is also a balance to find here between the time a user can spend and the willingness to ask a lot of questions. For instance, at the beginning, we wanted to question about the 17 facts for each explanation.

Even if the length of the explanations seems satisfying to the users (64% agreed with fact 10), they argue that they are too repetitive (80% of them for fact 11). This is particularly shown by fact 12 (58% agreed) about the difficulty to read the explanation until the end. Nonetheless, the provided explanations are perceived as extremely consistent (94% agreed with fact 16), and all causes are correctly identified to most of the people (68% agreed with fact 14, only 10% disagreed).

About the choice of the use case, some comments claim the detection of pasta does not need explanations, and others claim it is a difficult example even with the images (fact 17). This emphasizes the difficulty to find a good use case, with a good balance between complexity and accessibility.

5 Conclusion

In this article, we treated the problem of generation of textual explanation in fuzzy inference systems from one end to the other: generation of explanations, introduction of an evaluation use case and finally evaluation by a survey questionnaire. This complete chain allowed us to highlight the locks and the difficulties of explaining decisions in rule-based systems.

We designed a new algorithm to produce explanations from the trace of a fuzzy rule-based system. The algorithm consists in three successive steps that build a textual explanation sufficiently understandable and consistent: justification extraction, explanation formatting and text generation. The key of our work is the function $\mathscr{R}$ that reduces a set of rules to the structure responsible of their activation degrees, and demonstrate a good transparency of the reasoning.

We also designed a protocol to evaluate explanations composed of a survey questionnaire with 17 issues split into 3 main categories (natural language,

human-computer interaction, content and form) and of a test scenario that has been designed for the largest audience. 69 surveyed commented that the explanations were useful, consistent and transparent. However, it has been pointed out that the explanations remain too uncommon regarding the way human communicate.

This paper also shows that the NLG part is paramount and not yet ready to support the construction of explanations. A better understanding of the way humans built explanations is needed. Moreover, the algorithm performs well on small rule bases but must be improved to be applied on real world rule bases.

References

[1] J. M. Alonso, C. Castiello, and C. Mencar, "A bibliometric analysis of the explainable artificial intelligence research field," in *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations*, 2018, pp. 3–15.

[2] L. A. Zadeh, "Fuzzy logic = computing with words," *IEEE transactions on fuzzy systems*, vol. 4, no. 2, pp. 103–111, 1996.

[3] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1 – 38, 2019.

[4] L. A. Zadeh, *Computing with words: Principal concepts and ideas*. Springer, 2012, vol. 277.

[5] E. H. Shortliffe and B. G. Buchanan, "A model of inexact reasoning in medicine," *Mathematical biosciences*, vol. 23, no. 3-4, pp. 351–379, 1975.

[6] B. Chandrasekaran, M. C. Tanner, and J. R. Josephson, "Explaining control strategies in problem solving," *IEEE Intelligent Systems*, no. 1, pp. 9–15, 1989.

[7] C. Moraga, "An essay on the interpretability of Mamdani systems," in *Combining Experimentation and Theory*. Springer, 2012, pp. 61–72.

[8] A. Ramos-Soto, A. Bugarín, and S. Barro, "Fuzzy sets across the natural language generation pipeline," *Progress in artificial intelligence*, vol. 5, no. 4, pp. 261–276, 2016.

[9] J. M. Alonso, A. Ramos-Soto, E. Reiter, and K. van Deemter, "An exploratory study on the benefits of using natural language for explaining fuzzy rule-based systems," in *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2017, pp. 1–6.

[10] M. Galesic and R. Garcia-Retamero, "Statistical numeracy for health: A cross-cultural comparison with probabilistic national samples," *Archives of Internal Medicine*, vol. 170, no. 5, pp. 462–468, 2010.

[11] H. David, "Describing probability the limitations of natural language," in *EMEA Proceedings*, 2005.

[12] B. J. O'Brien, "Words or numbers? the evaluation of probability expressions in general practice." *JR Coll Gen Pract*, vol. 39, no. 320, pp. 98–100, 1989.

[13] S. Kent. (1964) Words of estimative probability. [Online]. Available: https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/sherman-kent-and-the-board-of-national-estimates-collected-essays/6words.html

[14] D. V. Budescu, H.-H. Por, and S. B. Broomell, "Effective communication of uncertainty in the IPCC reports," *Climatic change*, vol. 113, no. 2, pp. 181–200, 2012.

[15] E. E. Kerre and M. De Cock, "Linguistic modifiers: an overview," in *Fuzzy logic and soft computing*. Springer, 1999, pp. 69–85.

[16] L. A. Zadeh, "A prototype-centered approach to adding deduction capability to search engines-the concept of protoform," in *Intelligent Systems, 2002. Proceedings. 2002 First International IEEE Symposium*, vol. 1. IEEE, 2002, pp. 2–3.

[17] L. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning," *Information Sciences*, vol. 8, no. 3, pp. 199 – 249, 1975. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0020025575900365

[18] U. Sauerland, "Vagueness in language: The case against fuzzy logic revisited," *Reasoning under Vagueness-Logical, Philosophical, and Linguistic Perspectives, Studies in Logic series of College Publications*, 2011.

[19] J. M. Alonso, C. Castiello, and C. Mencar, "Interpretability of fuzzy systems: Current research trends and prospects," in *Springer Handbook of Computational Intelligence*. Springer, 2015, pp. 219–237.

[20] R. Rovatti, R. Guerrieri, and G. Baccarani, "An enhanced two-level boolean synthesis methodology for fuzzy rules minimization," *IEEE Transactions on Fuzzy Systems*, vol. 3, no. 3, pp. 288–299, 1995.

[21] A. Gatt and E. Reiter, "Simplenlg: A realisation engine for practical applications," in *Proceedings of the 12th European Workshop on Natural Language Generation*. Association for Computational Linguistics, 2009, pp. 90–93.

[22] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[23] M. Narayanan, E. Chen, J. He, B. Kim, S. Gershman, and F. Doshi-Velez, "How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation," *arXiv preprint arXiv:1802.00682*, 2018.

[24] J. Moore, *Assessment of Explanation Systems.* Technology Assessment in Software Applications, 1994.

[25] N. Tintarev and J. Masthoff, "Designing and evaluating explanations for recommender systems," in *Recommender systems handbook.* Springer, 2011, pp. 479–510.

[26] B. Buchanan, "Rule based expert systems," *The MYCIN Experiments of the Stanford Heuristic Programming Project*, 1984.

[27] G. Carenini, V. O Mittal, and J. Moore, "Generating patient-specific interactive natural language explanations," *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pp. 5–9, 02 1994.

[28] R. L Ye and P. Johnson, "The impact of explanation facilities on user acceptance of expert systems advice," *MIS Q.*, vol. 19, pp. 157–172, 06 1995.

[29] E. Reiter and A. Belz, "An investigation into the validity of some metrics for automatically evaluating natural language generation systems," *Computational Linguistics*, vol. 35, no. 4, pp. 529–558, 2009.

[30] P. Thagard, "Explanatory coherence," *Behavioral and brain sciences*, vol. 12, no. 3, pp. 435–467, 1989.

[31] A. Furnham, "Response bias, social desirability and dissimulation," *Personality and Individual Differences*, vol. 7, no. 3, pp. 385 – 400, 1986.

[32] G. L. Legendre, "Pasta by design," *Architectural Design*, vol. 81, no. 4, pp. 100–101, 2011.