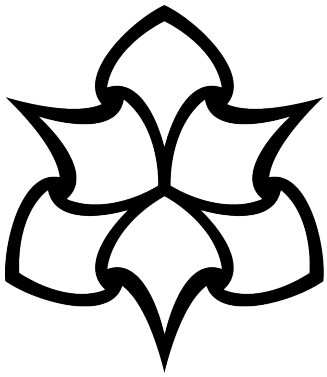

Interpreting Human Responses in Dialogue Systems using Fuzzy Semantic Similarity Measures

Naeemeh Adel, Keeley Crockett
School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University, Chester Street, Manchester, M1 5GD, UK
N.Adel@mmu.ac.uk

David Chandran
Institute of Psychiatry, Psychology & Neuroscience, Kings College London, 16 De Crespigny Park, London, SE5 8AF, UK

Joao P. Carvalho
INESC-ID / Instituto Superior Tecnico, Universidade de Lisboa, Portugal

***Abstract*— Dialogue systems are automated systems that interact with humans using natural language. Much work has been done on dialogue management and learning using a range of computational intelligence based approaches, however the complexity of human dialogue in different contexts still presents many challenges. The key impact of work presented in this paper is to use fuzzy semantic similarity measures embedded within a dialogue system to allow a machine to semantically comprehend human utterances in a given context and thus communicate more effectively with a human in a specific domain using natural language. To achieve this, perception based words should be understood by a machine in context of the dialogue. In this work, a simple question and answer dialogue system is implemented for a café customer satisfaction feedback survey. Both fuzzy and crisp semantic similarity measures are used within the dialogue engine to assess the accuracy and robustness of rule firing. Results from a 32 participant study, show that the fuzzy measure improves rule matching within the dialogue system by 21.88% compared with the crisp measure known as STASIS, thus providing a more natural and fluid dialogue exchange.**

Keywords*— *dialogue systems, conversational agents, fuzzy semantic similarity measures, fuzzy natural language

I. INTRODUCTION

Dialogue Systems (DS) are applications, which effectively replace human experts by interacting with users through natural language dialogue to provide a type of service or advice [1]. In order for a DS to engage with humans, they must be able to handle extended natural language dialogue relating to complex tasks and potentially engage in decision-making. In this sense, agents are helpful tools for human-machine interaction, allowing the input of data via natural language, processing sentences, and returning answers appropriately through text. DS, sometimes known as conversational agents, have been used in a wide range of applications such as customer service [1], help desk support [2], Educational [3,4,5,6], Cognitive Behavioural Therapy for young adults [7], insurance [8] and healthcare [9]. Dialogue understanding has become more valuable to companies with the easier ability to gain insights from unstructured text through Google's AutoML and natural language API [10], to Amazon's use of supervised machine learning to allow correct interpretation of natural language vocabulary reducing, for example, the detection of false positive responses [11]. For spoken DS, task based systems which utilise deep reinforcement learning techniques in their dialogue management systems are also becoming more available to industry [12]. What makes a successful DS is the ability for the machine to understand and interpret the human's natural language response in the context of the conversation.

Traditionally, DS used a pattern matching method to determine the most suitable response through computation of rule strengths for all matched occurrences of scripted patterns in the context of the system. The pattern matching approach has shown effectiveness and flexibility to develop extended dialogue applications [1, 13, 14] especially when coupled with ruled based matching algorithms to produce controlled responses and offer flexibility to sustain dialogues with users. However, scripting patterns is known as a laborious and time-consuming task with many flaws. More recently, some DS have opted to use short text semantic similarity measures (STSM) in place of pattern matching [6, 14, 15]. Utilising STSM within a DS is more effective than other techniques because it replaces the scripted patterns by a few natural language sentences in each rule. Evaluation of STSM based systems has been shown to improve the robustness of the system in terms of increasing the number of correctly fired rules, thus maintaining the conversational flow and increasing usability [15, 16]. However, when traditional STSM are used, they do not sufficiently match the fuzziness of natural language i.e. the human perception-based words, leading to a fundamental meaning of the human utterance in the dialogue context being misunderstood, causing incorrect firing of a rule, leading to incorrect flow of conversation and even wrong tasks being suggested.

Fuzzy Sentence Similarity Measures (FSSM) are algorithms that can compare two or more short texts or phrases which contain human perception-based words, and will return a numeric measure of similarity (composed of both semantic and syntactic elements) of meaning between them. This paper utilises one such measure known as FUSE (FUzzy Similarity mEasure) [17] which uses both WordNet [18] and a series of fuzzy ontologies which have been modelled from human representations using Interval Type-2 fuzzy sets [17]. FUSE has

been shown to model *intra-personal* and *inter-personal* uncertainties of fuzzy words representative of natural language.

This paper describes the creation and evaluation of a simple DS which utilises the FUSE measure to match human utterances to a set of fuzzy phrases with a rule-based system. The aim is to improve the robustness of rule matching within the DS compared with the use of a crisp similarity measure in a market research scenario where the capture of rich descriptive dialogue is important in gaining customer insight. A fuzzy DS can be used to automate the analysis of unstructured answers given to open ended questions, allowing for richer insight when collecting survey data. For example, an understanding of the dialogue, can lead to further probing to obtain more descriptive answers that provide greater insight into why a particular answer was given. This paper aims to address the following research question:

Can a Fuzzy Sentence Similarity Measure (FSSM) be incorporated into a dialogue system to improve rule matching ability from user utterance compared with a traditional STSM?

This paper is organised as follows; Section II provides a brief overview of dialogue systems and illustrates the differences between the use of traditional pattern matching and semantic similarity measures with the management of the human-machine conversation. Section III describes the design of a simple dialogue system that comprises of an FSSM, for collating human responses for evaluating customer feedback in a café and section IV describes the experimental methodology and results. Finally, section V presents the conclusions and future work.

II. DIALOGUE SYSTEMS

In this section, we briefly examine the dialogue engine within the DS, which is used to maintain conversational flow. We review and highlight typical problems associated with pattern matching and outline why the use of STSS overcomes some of the problems.

```
rule <tle-help-desk>
a:0.01
c:%att_name%
p:50 * something wrong * pc*
p:50 * something wrong * pc
p:50 * something wrong * computer*
p:50 * computer* * faulty*
p:50 * pc* faulty*
p:50 * computer* broken*
p:50 * pc* broken*
p:50 * computer *nt work*
p:50 * pc* *nt work*
p:50 * curing * fault * computer*
p:50 * curing * fault * pc*
p:50 * fault* * pc*
p:50 * fault* computer*
p:50 * pc * fault*
p:50 * computer * fault*
p:50 * problem * pc*
p:50 * problem * computer*
r: Please can you explain what the problem is? *<set att_service_type PC_fault>
```

Fig. 1 Pattern matching rule

A) *Strengths and Weaknesses of Pattern Matching*

A dialogue system, sometimes referred to as a conversational agent (CA) is a computer program which interacts with a user through natural language dialogue and provides some form of service [1, 2, 19, 20, 21], however, they typically suffer from high maintenance in updating dialogue patterns for new scenarios due to the huge number of language patterns within the scripts. Typically DS work off scripts, which are organized into contexts, consisting of hierarchically organized rules with combining patterns and associated responses (see Figure. 1 for an example of a pattern matching rule). Scripts need to capture a wide variety of inputs and hence many rules are required, each of which deals with an input pattern and the possible variations and an associated response [5, 14, 16]. InfoChat is one such pattern matching system which utilises the sophisticated PatternScript scripting language [22] and has been adapted over the years for use in intelligent conversational tutorial systems [6]. Figure. 1 shows an example of a pattern matching rule, *<tle-help-desk>* which has been encoded using the scripting language provided with the agent InfoChat. The rule uses default values for (*a*)ctivation and (*p*)attern matching strength, has a (*c*)ondition (that the variable *att_name* has a value) and a response consisting both of a text and the setting of a variable *<set att_service_type PC_fault>*. Figure. 1 illustrates that scripting patterns is inefficient, results in domain instability and high maintenance costs. Whilst pattern matching scripting engines are a mature technology and robust, to some degree to expected user input, scripting is an art form and requires good knowledge of the language and the ability to perform in-depth knowledge engineering of the domain [1, 4, 16].

B) *Semantic Similarity Measures*

In a Semantic Dialogue System, each rule is matched in accordance with a pre-determined semantic similarity threshold, which is set initially through empirical evaluation and depends upon the sensitivity of rules within a context. A simple rule (Figure. 2) comprises of a set of prototypical sentences, (*s*), where the similarity with the user utterance is calculated using a STSM. Each rule has a series of responses, (*r*), which are provided to the user and can be randomly selected. Each rule also has an associated default rule, which would fire if the user utterance failed to match any prototypical sentences within the rule. O'Shea et al [15] devised a semantic scripting language which incorporated an STSS through adapting the pattern matching language of InfoChat [16] which encompasses the

```
rule <tle-help-desk>
c:%att_name%
s: There is a problem with my computer
r: Please can you explain what the problem is? *<set att_service_type PC_fault>
```

Fig 2. Semantic rule

ability to extract patterns to set variables, set rule conditions and freeze, promote and demote rules.

In a semantic system, prototypical sentence rules are compared with user utterances using a pre-selected STSS algorithm and the rule with the highest similarity match would fire. The most obvious benefit of using semantic rules is that no patterns are required and more importantly the semantic meaning of the utterance can be captured and acted upon within the dialogue context. Aljameel [4] used a hybrid similarity approach, combining an STSM with limited patterns, to construct an Arabic conversational intelligent tutoring system for the education of autistic children. The conversational agent processed Arabic utterances using a novel crisp STSM which utilised the cosine similarity measure to solve the word order issue associated with the Arabic language. Consequently, this reduced the number of scripts and rules required. Through empirical evaluation of two versions of the system, the use of an STSM reduced the number of unrecognised human utterances to 5.4% compared to 38% in the pattern scripted version and, hence, the systems incorrect responses were reduced to 3.6% compared to 10.2% in the pattern scripted version [4]. Similar improvements on the benefits of utilising a STSM within DS are also reported in [23]. In this paper, we will replace the traditional semantic similarity measure with a Fuzzy semantic similarity measure to evaluate the effectiveness of a DS through a reduction in the incorrect responses and unrecognised human utterances compared with using an STSM.

III. A Simple Dialogue System for Collating User Responses

A) Overview

In this section, we describe a simple question and answer dialogue system that utilises the FUSE semantic similarity measure [17], to match user utterances to different categories of responses to each question. The dialog structure is therefore a linear sequence of questions, where each question response has three possible branches. The aim is to distinguish between human perceptions of fuzzy words in nine categories to assess if the correct rule fires in response to natural language used within the human utterance. FUSE [17] is an ontology based similarity measure that uses Interval Type-2 fuzzy sets to model relationships between categories of human perception based words. The FUSE algorithm identifies fuzzy words in a human utterance and determines their similarity in context of both the semantic and syntactic construction of the sentence. Currently FUSE consists of nine fuzzy categories each containing a series of fuzzy words. These categories are Size/Distance, Age, Temperature, Worth, Level of Membership, Frequency, Brightness, Strength and Speed. Initial selection and methodology for word population can be found in [17]. An experiment originally described in [17] was used to capture human ratings to create the fuzzy ontology for these categories where words were modelled based on Mendel's Hao-Mendel Approach (HMA) using Interval Type-2 fuzzy sets [24]. A full description of the FUSE algorithm and the general approach on how the fuzzy word models and measures in each category were derived is given in [17].

TABLE I: Café Feedback Dialogue Questions Mapped to Fuzzy Categories

Question	Category	Question Asked
Q1	Size/Distance	Using descriptive words, how would you describe the size of the queue?
Q2	Temperature	How would you describe the temperature of the cafe?
Q3	Brightness	How would you describe the brightness of the cafe?
Q4	Age	Using descriptive words, how would you describe the age of the barista that served you?
Q5	Speed	Once you placed your order, how quickly was your drink made and served to you?
Q6	Strength	Looking up from your screen to the first person you see, how would you describe their physical strength?
Q7	Frequency	How frequently do you visit this cafe?
Q8	Level of Membership	How did todays visit meet your expectation?
Q9	Worth	How would you describe your experience overall today?

B) Design of a Dialogue System for Café Feedback

In order to establish if a FSSM could be used in a dialogue system, a simple question and answer system was designed to obtain feedback from participants who visited a local café. This was done using a knowledge engineering approach and involved gathering information about typical questions asked in a customer satisfaction online questionnaire concerning customer satisfaction levels in high street cafes. Existing survey questions were either a mixture of dichotomous questions, multiple choice, Likert scale questions or free text. Within the proposed Café feedback DS, each question selected had to be transformed into one which would allow the user to provide descriptive textual answers in order to gather as much data as possible to evaluate the impact of the fuzzy semantic measure. To ensure all the categories in FUSE were covered, nine questions where created (Table I), each one covering responses that would contain words or synonyms of words from each fuzzy category. Each question formulates a question-rule within the DS where each rule can have three responses which represent full coverage of the categories as defuzzified word values obtained through human experts and Type-II modelling using HMA approach [17].

The rule responses were divided into three thresholds of high, medium and low, and words (and word synonyms) within each category would fall under each threshold. The threshold for each category varies as the number of words and measurements in each category varies (dependent on human perceptions [17]). The thresholds in each of the nine categories were selected based

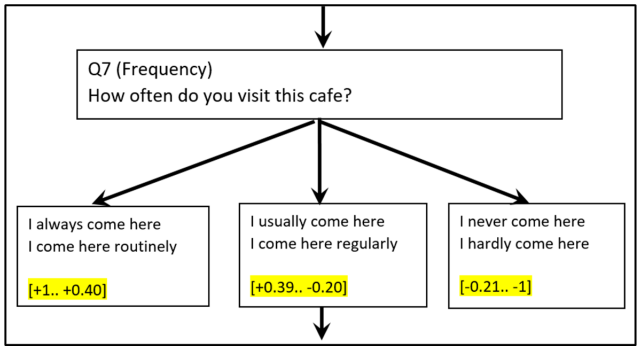

Fig 3. Frequency threshold

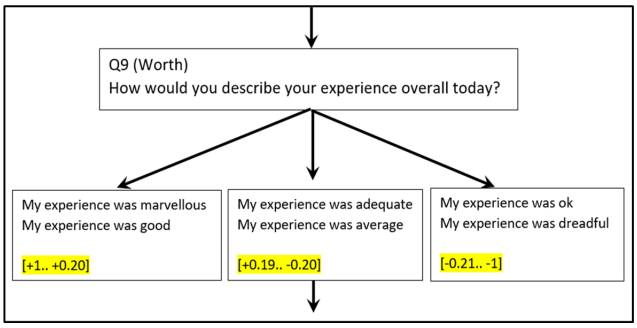

Fig 4. Worth threshold

on the words in that specific category. An example is shown in Figures 3 and 4 for the two categories of *Frequency* and *Worth*.

Considering Figure. 3, for the category *Frequency*, the low threshold begins at [-1] and ends at [+0.40], with the last word to fall in this threshold being *Everytime*, and the next word after this which begins the mid threshold is *Occasionally* at [+0.39], and this threshold continues up to [-0.20], and even though this is now a negative value, it still falls in the mid threshold for this category, and the low threshold starts at [-0.21] and ends at [-1]. Examining Figure. 4 for category *Worth*, the high threshold starts at [+1] and ends at [+0.20], the mid threshold begins at [+0.19] and ends at [-0.20], and the low threshold begins at [-0.21] and ends at [-1]; thus there was not a single fixed threshold

```
<default-rule1><size/distance>
s: It was long
s: It was huge
r: Using descriptive words, how would you describe the size of the queue?
*<set att_size-distance-high>
c: temperature_context

<Default-rule2><size/distance>
s: It was average
s: It was regular
r: Using descriptive words, how would you describe the size of the queue?
*<set att_size-distance-medium>
c: temperature_context

<Default-rule3><size/distance>
s: It was tiny
s: It was small
r: Using descriptive words, how would you describe the size of the queue?
*<set att_size-distance-low>
c: temperature_context
```

Fig 5. Sample Rules for Size/ distance category

for all nine categories, as the words and there values varied in each category. In order to determine the specific high, medium and low thresholds for each fuzzy category, two English language experts independently grouped the words for each category. In the case of disagreement, a third expert was asked to cast the deciding vote.

C) Scripting

Each question (Table I) was scripted into a context which represented a category. Three English prototypical sentences were used in each rule to enable coverage of either the high, medium or the low thresholds. In addition, there were initialisation and conclusion contexts. Figure. 5 shows three rules (*r*) from the *Size/Distance* category. Each dialogue exchange between human and machine generated a human utterance that was compared to the prototypical sentences in each rule. In each context, the rule where the (*s*)sentence gave the highest similarity score compared with the human utterance, was analysed and fired through FUSE. An attribute is set i.e. *att_size-distance-high* becomes true if *default-rule1* fires and a

```
**********************************************************************************************************************
Hello, my name is Fusion.
I am going to ask you a set of questions relating to today's experience in the cafe.
When writing your answers it is very important to use complete sentences rather than short word answers
and please make sure all words are spelled correctly, and no numbers or symbols are used."
Now let's begin...
**********************************************************************************************************************

Q1)     Using descriptive words, how would you describe the size of the queue? It was excessively long
Q2)     How would you describe the temperature of the cafe? The temperature of the cafe is warm
Q3)     How would you describe the brightness of the cafe? the cafe had very bright lights
Q4)     Using descriptive words, how would you describe the age of the barista that served you? fairly young, possibly a student
Q5)     Once you placed your order, how quickly was your drink made and served to you? the service was fast
Q6)     Looking up from your screen to the first person you see, how would you describe their physical strength? They appear reasonably strong
Q7)     How frequently do you visit this cafe? i come here often as its spacious
Q8)     How did todays visit meet your expectation? generally a very good experience as usual
Q9)     How would you describe your experience overall today? it was superb, really liked it

**********************************************************************************************************************
Thank you! You have reached the end of the questions. Please inform the researcher you have finished
**********************************************************************************************************************
```

Fig 6. Simple Interface Design

TABLE II: RESULTS OF FUSION DS WITH FUSE VS STASIS SSM

Category	FUSE TP	FUSE TP%	FUSE FP	FUSE FP%	STASIS TP	STASIS TP%	STASIS FP	STASIS FP%
Q1 Size/Distance	26	81.25	6	18.75	20	62.50	12	37.50
Q2 Temperature	31	96.88	1	3.13	21	65.63	11	34.38
Q3 Brightness	27	84.38	5	15.63	27	84.38	5	15.63
Q4 Age	24	75.00	8	25.00	17	53.13	15	46.88
Q5 Speed	31	96.88	1	3.13	26	81.25	6	18.75
Q6 Strength	24	75.00	8	25.00	16	50.00	16	50.00
Q7 Frequency	27	84.38	5	15.63	14	43.75	18	56.25
Q8 Level of Membership	31	96.88	1	3.13	23	71.88	9	28.13
Q9 Worth	32	100.00	0	0.00	26	81.25	6	18.75
Average %TP Rate	**FUSE: 87.85%**				**STASIS: 65.97%**			

change in context will occur, denoted by the '*c:*' identifier. As this is a simple linear DS the change in context is always set to the context of the next question until all questions have been asked. Figure. 6 shows an example of a participants answers.

On initiation of the system, the DS begins with the simple message:

"Hello, My name is Fusion. I am going to ask you a set of questions relating to today's experience in the cafe. When writing your answers it is very important to use complete sentences rather than short word answers and please make sure all words are spelled correctly, and no numbers or symbols are used. Now let's begin...".

After all questions were asked the final message was "*Thank you! You have reached the end of the questions. Please inform the researcher you have finished.*"

A log file recorded all dialogue, including the semantic similarity score for each rule during the completion of the survey. In this version of the system, all human utterances were recorded, with incorrect utterances failing to match any rules in each context also being recorded.

IV. EXPERIMENTAL DESIGN

A) Experimental Methodology

Following Manchester Metropolitan Universities ethical approval process (Ethos number: 11759), 32 participants were recruited through an advertising campaign through the University. After agreeing to take part, and agreeing a suitable time, participants were given a voucher to purchase a drink at one of two cafes within the University. On purchasing a beverage, the participant was asked to sit down and observe their environment for 10-15 minutes. Once finished, they notified the researcher (who was sat independently) and began to complete the café feedback survey using the DS about their experience and visit to the café. During this interaction, the typed user utterances for each answer is run through the DS and compared with the thresholds for the corresponding category. For analysis purposes, each user utterance was taken and compared with the two sentences for each of the high, medium and low threshold sentences. The similarity is calculated for each sentence pair using FUSE and the results are recorded and the highest similarity rating is noted for each interaction. All dialogue exchanges are recorded in a log for analysis. Once completed, the participants completed a short usability questionnaire, with questions comparable to those used to typically assess usability of DS [25, 26].

To analyse the results, a dataset consisting of 288 rows was compiled of all user responses to all questions, along with the semantic similarity measurement for each rule calculated using FUSE. For comparison purposes, the same rules and responses were also fired through a well-established similarity measurement known as STASIS [27]. STASIS is not able to capture the meaning of fuzzy words. STASIS only caters for crisp values and uses WordNet and Browns Corpus to find similarity rating for sentence pairs [27].

B) Results

Table II shows the results from all 32 participants for the TP and FP values run for both FUSE and STASIS and shows the percentage of correct TP for FUSE compared with that of STASIS. The fuzzy words assigned to each of the thresholds are examined and if the DS has picked up the correct sentence match then this is counted as a True Positive (TP) and given a score of 1. If the highest similarity rating has not fallen under the correct threshold of words, then it is classed as a False Positive (FP) and given a score of 0.

As can be seen from the results in Table II, FUSE has an average TP rating of 87.85% and STASIS has an average TP rating of only 65.97%. The average TP rating represents the total number of correctly fired rules that are also correctly matched with the user utterances and are therefore a true positive. These results show that the fuzzy dictionary of words modelled within the FUSE categories increases the similarity

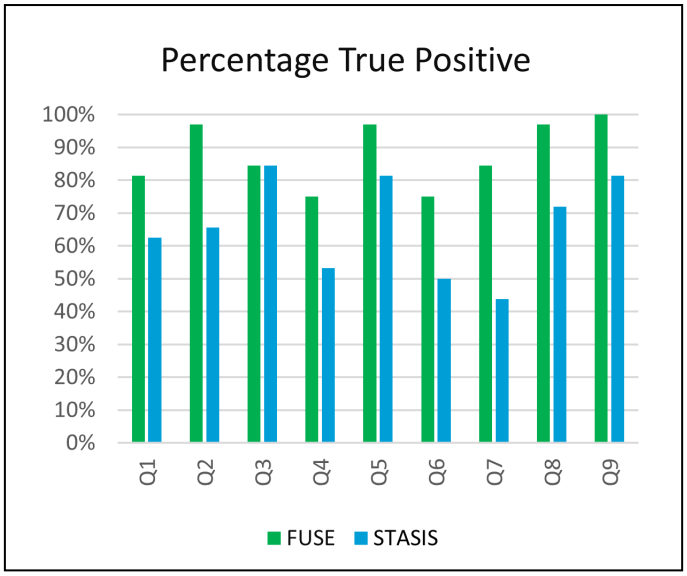

Fig 7. Percentage of TP values for FUSE vs STASIS

rating when compared with that of human utterances as opposed to just crisp values.

Figure. 7 shows the percentage of correctly matched user utterances using FUSE and STASIS. Each question is designed to represent a separate category for comparison purposes, therefore even though STASIS does not have a fuzzy dictionary and only uses WordNet it can still be used in this scenario to compare the effects of fuzzy words vs crisp values. It can be seen in Figure. 7 that for all the nine categories, with the exception of *Brightness* (Q3), FUSE always has a higher TP rating then STASIS, meaning it has a higher number of true positive matches that fired under the correct threshold. For Q3 (*Brightness*), both FUSE and STASIS scored the same, meaning they both fired the same correct thresholds.

C) Discussion

Overall, the results have shown that a DS that utilises the FUSE measure to determine which rule fires, provides a higher average TP rating using fuzzy words as opposed to STASIS that only uses crisp values. There was an improvement of 21.88% in the average TP rating as can be seen in Table II when compared with STASIS, where fuzzy words are not taken into consideration. There were however, some rules that did not fire correctly and this section provides some in-depth analysis of those rules to feed into future work on the system.

In total, 8 (out of 288) of the user utterances contained some numerical responses as well as just words; an example is shown below of an instance where the DS asked the question relating to the category *Age*:

Q4) Using descriptive words, how would you describe the age of the barista that served you?

User Utterance: The physical appearance of the barista tells that she was in her 30's

Both FUSE and STASIS picked this up as belonging to the low category, consisting of words such as *baby*, *young*, *child*, etc; when according to the two English language experts, it should be in the mid threshold containing words such as *adult*, *middle-aged*, *grownup* etc. On the other hand, when the DS asked the question relating to the category *Size/Distance*:

Q1) Using descriptive words, how would you describe the size of the queue?

User Utterance: The size of the queue was 2-3 people long with a wait time of no longer than 1 minute.

Both FUSE and STASIS picked this up as being in the mid threshold, containing words such as *average*, *standard*, *middle*, and the two English language experts agreed that this can be classed as a TP and is in the correct threshold.

Neither FUSE nor STASIS was able to deal with the effect of the inclusion of negation words within utterances. For example, when the DS asked the question relating to the category *Brightness*:

Q3) How would you describe the brightness of the cafe?

User Utterance: The light level of the cafe is not bright

Both FUSE and STASIS picked this up as the high threshold because of the word *bright*, when in effect due to the use of the word *not*, it actually means it was dark. Therefore is this case, the correct rule category did not fire (i.e. *bright* was identified as being in the high threshold by the English language experts, but the presence of the word *not* would contradict this and it should be the in the low threshold).

An additional example of negations leading to an incorrect rule firing was when the DS asked the question relating to the category *Strength*:

Q6) Looking up from your screen to the first person you see, how would you describe their physical strength?

User Utterance: I would describe them as lean and not very strong.

Both FUSE and STASIS picked this up as belonging to the high threshold due to the word *strong (*and had an increased intensity in FUSE to the hedge word very*)*, when in fact because of the use of the word *not* it actually should belong to the low or mid thresholds and this was also confirmed by the two English language experts.

There were some instances where FUSE correctly matched a rule and STASIS did not. One example of this is when the DS asked the question relating to the category *Size/Distance*:

Q1) Using descriptive words, how would you describe the size of the queue?

User Utterance: The size of the queue was huge.

FUSE picked this up as belonging to the high threshold with a similarity value of ((D1) It was long: 0.57554), and STASIS picked this up as belonging to the low threshold, with a similarity value of ((D3) It was small: 0.53459). The high threshold is correct, since it holds words such as *big*, *massive* and *huge*. Although the difference in the two similarity ratings are small, it is down to the fact that the high threshold actually

holds the word *huge* therefore this is the threshold it must fall under for it to be a TP [17].

An instance when STASIS correctly matched a rule and FUSE did not is when the DS asked the question relating to the category *Brightness*:

Q3) How would you describe the brightness of the cafe?

User Utterance: It was fairly bright

STASIS picked this up as belonging to the high threshold with a similarity value of ((D1) The cafe was bright: 0.36442), and FUSE picked this up as belonging to the mid threshold with a similarity value of ((D2) The cafe was luminous: 0.67367). The high threshold is correct as it holds words such as *sunny*, *radiant* and *bright*.

D) Effect on Usability

All participants completed a short usability survey comprising of 13 Likert scale questions with allowable free text, following completion of the task. A full in-depth usability analysis is beyond the scope of this paper, but it is important to highlight that the inclusion of a FSSM into the DS did not appear to negatively affect the usability of the system. In summary, 94% agreed or strongly agreed that a DS could be used as a mechanism to answer survey questions in the future. 90% of participants reported no inconsistences when using the system and 91% found the system easy to interact with and intuitive to use.

V. Conclusion And Further Work

This paper has described the development of a simple linear DS that incorporated the FUSE semantic similarity algorithm. The semantic similarity of user utterances and rules was compared using both FUSE and STASIS in order to determine which of the three rules in each category would fire. The results show that the average TP of FUSE is 87.85% which is an improvement of 21.88% when compared with STASIS rule firing rating (65.97%). Given the original research question, we conclude that a Fuzzy Sentence Similarity Measure (FSSM) can be incorporated into a dialogue system to improve rule matching ability from a user utterance compared with a traditional STSM. A weakness of utilising FUSE was its inability to deal with the word "Not" within the dialogue, which caused misfiring of rules. Future work will address this issue by looking at ways to apply the fuzzy NOT operator to the associated word.

Despite the simplicity of the DS, a number of issues have been recognised. Firstly, neither measure (STASIS or FUSE) were able to produce correct rule firings when a negation word was used to form part of the utterance. All though hedges had been considered as an addition to the FUSE fuzzy dictionary [17], negation words were not included in the similarity calculation within FUSE. Secondly, FUSE is very much dependent on the fuzzy dictionary created in previous work, which were generated from many empirical experiments [17] where humans rated words within categories and then within the context of general sentences. In this paper, it is clear that the context of perception-based words does matter when used by a FSSM in a DS. Further work will include the evaluation of a second, more substantial prototype DS, which will incorporate other fuzzy similarity measures [28] and revisit the impact of hedge words.

Acknowledgment

This work was supported by national funds through FCT, Fundação para a Ciência e a Tecnologia, under project UIDB/50021/2020.

References

[1] J. D. O'Shea, Z. Bandar, K. Crockett, "Systems Engineering and Conversational Agents", In *Intelligence-Based Systems Engineering*, Intelligent Systems Reference Library, Springer, Berlin, Heidelberg, 2011, vol. 10, pp. 201-232.

[2] L. Ozaeta, M. Graña, 2018. A View of the State of the Art of Dialogue Systems. In: de Cos Juez F. et al. (eds) *Hybrid Artificial Intelligent Systems*. *HAIS*. 2018. Lecture Notes in Computer Science, vol. 10870. Springer, Cham https://doi.org/10.1007/978-3-319-92639-1_59

[3] L. Lin, P. Ginns, T. Wang, P. Zhang, 2020. Using a pedagogical agent to deliver conversational style instruction: What benefits can you obtain?. *Computers & Education*, 143, p.103658.

[4] S.S. Aljameel, "Development of an Arabic conversational intelligent tutoring system for education of children with autism spectrum disorder", PhD dissertation, School of Computing, Maths and Digital Technology, Manchester Metropolitan University (MMU), 2018.

[5] S.S. Aljameel, J.D. O'Shea, K. Crockett, A. Latham, and M Kaleem, 2019. LANA-I: an Arabic conversational intelligent tutoring system for children with ASD. In *Intelligent Computing-Proceedings of the Computing Conference*, pp. 498-516. Springer, Cham.

[6] A. Latham, K. Crockett, D. McLean, 2014. An adaptation algorithm for an intelligent natural language tutoring system. *Computers & Education*, 71, pp. 97-110.

[7] K.K. Fitzpatrick, A. Darcy, M. Vierhile, 2017. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health*, vol. 4, no. 2, p.e19.

[8] F. Koetter, M. Blohm, J. Drawehn, M. Kochanowski, J. Goetzer, D. Graziotin and S. Wagner, 2019, February. Conversational Agents for Insurance Companies: From Theory to Practice. In *International Conference on Agents and Artificial Intelligence*, pp. 338-362. Springer, Cham.

[9] J.L.Z. Montenegro, C.A. da Costa, R. da Rosa Righi, 2019. Survey of Conversational Agents in Health. *Expert Systems with Applications*, vol. 129, pp. 56-67, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2019.03.054.

[10] Google Natural Language, Jan. 01, 2020. [Online]. Available: https://cloud.google.com/natural-language/#overview. [Accessed Jan. 3, 2020].

[11] Alexa and Alexa Device FAQs, Feb. 23, 2016. [Online] Available: https://www.amazon.com/gp/help/customer/display.html?tag=skim1x169757-20&nodeId=201602230. [Accessed Jan. 3, 2020].

[12] J. He, B. Wang, M. Mingming Fu, T. Yang and X. Zhao, Hierarchical attention and knowledge matching networks with information enhancement for end-to-end task-oriented dialog systems, *IEEE Access*, vol. 7, pp. 18871–18883, 2019.

[13] R.R.A. Pazos, B.J.J. González, L.M.A. Aguirre, F.J.A. Martínez and H.H.J Fraire, 2013. Natural Language Interfaces to Databases: An Analysis of the State of the Art. In: *Recent Advances on Hybrid Intelligent Systems*, O. Castillo, P. Melin and J. Kacprzyk, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 463-480.

[14] J.D. O'Shea, "A framework for applying short text semantic similarity in goal-oriented conversational agents", PhD dissertation, School of Computing, Maths and Digital Technology, Manchester Metropolitan University (MMU), 2010.

[15] K. O'Shea, K. Crockett, Z. Bandar, J.D. O'Shea, Erratum: An approach to conversational agent design using semantic sentence

similarity (Appl Intelligence) *Applied Intelligence*, vol. 40, no. 1, pp. 199-199, 2014.

[16] C. Curry, "A framework for developing a conversational agent to improve normal age-associated memory loss and increase subjective wellbeing", PhD dissertation, School of Computing, Maths and Digital Technology, Manchester Metropolitan University (MMU), 2019.

[17] N. Adel, K. Crockett, A. Crispin, D. Chandran and J.P. Carvalho, Jul. 2018, FUSE (Fuzzy Similarity Measure)-A measure for determining fuzzy short text similarity using Interval Type-2 fuzzy sets. In *2018 IEEE International Conference on Fuzzy Systems* (*FUZZ-IEEE*) pp. 1-8, IEEE.

[18] Princeton University, "About Wordnet". [Online]. Available: http://wordnet.princeton.edu/ [Accessed Jun. 13, 2014].

[19] J.G. Harms, P. Kucherbaev, A. Bozzon and G.J. Houben. 2018. Approaches for Dialog Management in Conversational Agents. *IEEE Internet Computing*, vol. 23, no. 2, pp.13-22.

[20] J.B. Aujogue, A. Aussem, 2019. Hierarchical Recurrent Attention Networks for Context-Aware Education Chatbots. In *2019 International Joint Conference on Neural Networks* (*IJCNN*) pp. 1-8, IEEE.

[21] J. Lester, K. Branting, B. Mott, "Conversational Agents". CRC Press LLC. [Online]. Available: https://www.ida.liu.se/~729A15/mtrl/Lester_et_al.pdf [Accessed Jun. 16, 2015].

[22] D. Michie, C. Sammut, Infochat Scripter's Manual, Convagent Ltd, Manchester, UK, 2001.

[23] M. Kaleem, J.D. O'Shea, K. Crockett, 2014. Word order variation and string similarity algorithm to reduce pattern scripting in pattern matching conversational agents. In *2014 14th UK Workshop on Computational Intelligence* (*UKCI*) pp. 1-8: IEEE, ISBN: 978-1-4799-5538-1, DOI: 10.1109/UKCI.2014.6930180.

[24] M. Hao, J.M., Mendel, 2015. Encoding words into normal interval type-2 fuzzy sets: HM approach, *IEEE Transactions on Fuzzy Systems*, vol. 24, no. 4, pp. 865-879.

[25] J. Deriu, A. Rodrigo, A. Otegi, G. Echegoyen, S. Rosset, E. Agirre and M. Cieliebak, 2019. Survey on Evaluation Methods for Dialogue Systems. *arXiv preprint arXiv:1905.04071.*

[26] X. Chen, J. Mi, M. Jia, Y. Han, M. Zhou, T. Wu and D. Guan, October 2019. Chat with Smart Conversational Agents: How to Evaluate Chat Experience in Smart Home. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, pp. 1-6, ACM.

[27] Y. Li, D. McLean, Z. Bandar, J.D. O'Shea and K. Crockett, 2006. Sentence similarity based on semantic nets and corpus statistics, *IEEE transactions on knowledge and data engineering*, vol. 18, no. 8, pp. 1138-1150.

[28] V. Cross, V. Morenko, K. Crockett and N Adel, 2019, June. Ontological and fuzzy set similarity between perception-based words. In *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* pp. 1-6, IEEE.