# Finding fuzziness in Neural Network Models of Language Processing

Kanishka Misra and Julia Taylor Rayz

**Abstract** Humans often communicate by using imprecise language, suggesting that fuzzy concepts with unclear boundaries are prevalent in language use. In this paper, we test the extent to which models trained to capture the distributional statistics of language show correspondence to fuzzy-membership patterns. Using the task of natural language inference, we test a recent state of the art model on the classical case of temperature, by examining its mapping of temperature data to fuzzy-perceptions such as *cool*, *hot*, etc. We find the model to show patterns that are similar to classical fuzzy-set theoretic formulations of linguistic hedges, albeit with a substantial amount of noise, suggesting that models trained solely on language show promise in encoding fuzziness.

## 1 Introduction

Fuzziness of variables such as HOT, TALL, etc. is often reflected in language use. Humans tend to use vague constructions such as *"Joe is quite tall"* or *"it's very cold outside"* in their everyday language and are still able to successfully communicate their intent. Fuzzy sets [22] provide us with a calculus for handling such imprecise expressions, and have been used in numerous engineering applications, but have yet to be fully adopted by the Natural Language Processing (NLP) community [2]. Computing with Words (CWW) [24] initiatives followed a similar trend: they are described by engineers and computer scientists and used in some applications, but are not trending in publications of natural language processing (NLP) fields.

A prevalent hypothesis in the computational linguistics and natural language processing communities is that the distributional statistics of language use can guide the process of acquiring word meaning — that *"You shall know the meaning of a*

Kanishka Misra and Julia Taylor Rayz
Purdue University, West Lafayette, IN 47906, USA
e-mail: kmisra@purdue.edu, jtaylor1@purdue.edu

*word by the company it keeps"* [4]. This distributional hypothesis has been the basis of a number of prominent language representation architectures, ranging from early neural network-based vector space models such as `word2vec` [11] to more recent models that fall under the category of *"pre-trained language models,"* such as BERT [3]. A key factor in models that follow the distributional hypothesis is that their knowledge is acquired solely from linguistic input, and reflects the extent to which the statistics present in the language can inform the model about the world. Indeed, recent studies have shown models such as BERT to show great feasibility in capturing factual [12] and categorical knowledge [19] by investigating their outputs on word prediction prompts such as *"A robin is a ___ ."* While these analyses shed light on how well prediction in such models reflects what is generally known to be true in the world (in a bivalent setting of strict truth and false), much is unknown about the manifestation of vagueness within the semantic knowledge that they capture.

We argue in this paper that models that are built to interpret and act upon language should also be able to account for vagueness in categories and predicates, which have borderline cases where the exact truth value is unclear. This is especially important in the case of systems involving rational human interaction, where, for example, a computer can communicate to the user how *icy* a road is on a particular wintry night by approximating it from precise data [15], and for this to be made possible, our models of language processing must encode aspects vagueness or fuzziness. To test this, we investigate a state of the art NLP model (RoBERTa$_{large}$ [8]) in its ability to encode fuzziness of language, specifically in this case, language about a classical fuzzy concept: TEMPERATURE. We formulate the probing experiment as a Natural Language Inference (NLI) task where the model is supplied with precise descriptions about temperature data and the model is evaluated based on the extent to which that data maps to opportunistically selected fuzzy categories such as *freezing, warm, etc.*

## 2 The concept of vagueness and the vagueness of concepts

We define vagueness to be the phenomenon underlying the uncertainty in demarcating concepts, i.e., the lack of clear-cut boundaries separating membership to one concept with the other.

### 2.1 Psychological significance

Experimentally, vagueness has been of interest in the psychological study of concepts. The seminal work of Rosch [14] showed consistently that people perceived category membership as a graded phenomenon, as opposed to being a matter of crisp boundaries. Labov [6] experimented on the boundaries of commonplace con-

cepts such as CUP, MUG, and BOWL, by presenting human subjects with exemplar drawing of the objects. He found subjects to be consistent in attributing names to object drawings with typical visual characteristics, while showing large inconsistencies between each other while naming drawings where the object was visually intermediate between categories (for instance being too tall to be a member of BOWL but too wide to be a CUP), indicating the close relationship of lack of consensus with vagueness. Additionally, the work of McCloskey and Glucksberg [10] involved experiments where human subjects had to explicitly rate whether a particular category (e.g. FRUIT) applied to multiple items in consideration (e.g. APPLE, BANANA). The authors found that when the same group of people were asked to perform the experiment again, after a few weeks, the subjects tended to be highly inconsistent with their original judgements for borderline cases (e.g. TOMATO), suggesting that vagueness also involves individual uncertainty.

## 2.2 Handling vagueness using fuzzy interpretations of linguistic hedges

Linguistic hedges [7], such as *sort of, technically, etc.,* are central to vagueness in language. Zadeh's seminal work [22] proposes a calculus to handle such fuzziness in scalar adjectives such as *tall*, *old*, etc. Subsequent work [23] proposed to explicitly treat hedges, such as *very, more-or-less, slightly* as operations on pre-defined fuzzy sets. Usually, the definition of the membership functions and fuzzy sets is left to the experts, as it is one of the most difficult tasks in a successful fuzzy expert systems [9]. For instance, Zadeh proposed that for a given fuzzy set $F = \{(x, \mu_F(x))\}$, hedges such as *very, extremely, etc.* can be modeled using "concentration" functions that pull the curve of memberships for $F$ inward, such as *very* $F = \mu_F^2(x)$. Meanwhile, hedges such as *slightly, almost* were modeled using "dilation" functions that spread out the values of the membership function, such as $\mu_F^{1/2}(x)$. Zadeh also proposed more complex interpretation of hedges, by employing the use of "intensifiers." An example of applying such recommendations to the concept of tall is shown in fig. 1. Whether or not Zadeh's proposal for fuzzy-interpretations of linguistic hedges is correct, they provide a useful method for handling vague predicates in semantics and serve as inspiration for some of our analyses.

While fuzzy sets in general, and linguistic hedges—modeled as operators—are defined by experts to model data, we wonder whether very large amount of data can model such linguistics hedges. Given a membership function, and some point $x$, one can determine whether this point $x$ is a member of a fuzzy set, and to what degree. What we are interested in is an answer to this question: given a point $x$, and approximate membership value of some fuzzy set, is it possible to reconstruct a fuzzy membership function for $x \in X$ where $|X|$ is very large.

In other words, we can assume that if somebody says that it is 90°F outside, we find a membership degree of *hot, very hot, slightly hot, etc.* using an expert-determined fuzzy membership function. What is less clear is whether an oracle (let us assume that we have such a thing) would be able to map 90°F to some relative po-

**Fig. 1** Possible membership transformations for linguistic hedges of *tall* (with membership function $\mu(h)$, generated using a general bell curve): (1) *slightly tall* – $\mu^{1/2}(h)$; (2) *very tall* – $\mu^2(h)$; and (3) *extremely tall* – $\mu^4(h)$, where $h$ is height. Note that this is solely for illustrative purposes and is not an exhaustive list of hedges of *tall*.
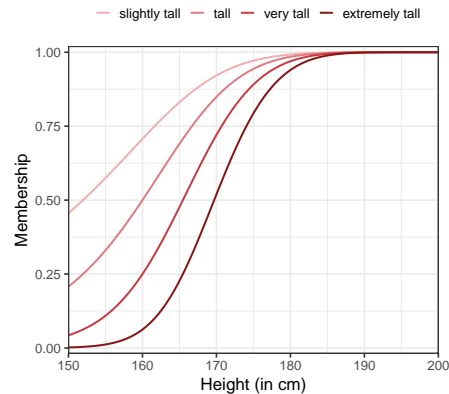
sition of *hot*, *very-hot*, etc. without *knowing* the actual equations for the membership functions.

More specifically, we wonder if the notion of fuzzy sets and hedges—modeled as proposed by Zadeh—can be reconstructed from very large data. If such reconstruction can be done, we should see graphs that are qualitatively similar to that of fig. 1. However, it should be noted that the linguistic hedges are indeed linguistic in nature, and thus, they can allow synonym substitutions. In particular, somebody very very tall may be called a giant, and our data may show correlation with that word choice, but not with "very very tall." While this raises the complexity of analysis, we note that it is essential in order to facilitate conversational interactions between computers and humans, but leave the fine-grained details for future work.

## 3 Pre-trained Language Models

Recent advances in the field of NLP have lead to the rise of a class of highly parameterized neural network-based models called pre-trained language models (PLMs). These models are trained on vast amounts of text using the language modeling objective — estimating probabilities of missing words in context.[1] A common architecture that governs the computation in these models is the transformer architecture [16]. The attention module in the transformer architecture allows PLMs to 're-member' all words in a given sentence context, while predicting the missing word, thereby making its internal representations more finely attuned to the entire context. Though originally proposed for ranking sentences (by assigning them probabilities), the language modeling objective guides PLMs into learning general-purpose language representations which can be fine-tuned to any supervised learning task end-to-end, i.e., all representations get updated in order to optimize for a given task(s).

---

[1] we assume a general form of the language modeling objective which also includes its bidirectional sense

Their expressiveness and adaptability has enabled PLMs to achieve state of the art results on several high-level NLP tasks, such as Question Answering, Natural Language Inference, Reading Comprehension, etc. PLMs come in two main variants: (1) Incremental PLMs, which are trained to predict words when conditioned on a left context, e.g., *"I went to the library to ___"*; and (2) Masked PLMs, which have bidirectional context while predicting tokens, e.g., *"I went to the ___ to read."* Examples of Incremental PLMs include GPT2 [13] while those of Masked PLMs include BERT [3] and RoBERTa [8].

## 4 Vagueness in Natural Language Inference Models

In this section, we present our behavioral experiment targeting the extent to which NLP models, trained on very large data, are sensitive to fuzziness/vagueness of concepts. We specifically consider the classical fuzzy concept of TEMPERATURE, where, for instance, the transition from *cool* to *warm* is gradual, and there is no clear-cut separation between when someone perceives their surrounding to fall under either of the two categories. To probe for how models encode the fuzziness of TEMPERATURE, we rely on the task of natural language inference (NLI) [1].

NLI, also referred to as "recognizing textual entailment" (RTE), is the task of predicting the logical entailment relation between pairs of sentences. Specifically, a model trained to do this task is supplied with two sentences, a premise and a hypothesis, and it has to predict whether the premise entails the hypothesis. For instance, the sentence *"a dog is running towards the man"* logically entails *"an animal is moving towards a human"*, while *"the boy is mowing the lawn"* contradicts *"the boy is sleeping"*. Note that it is assumed that the premise and the hypothesis must refer to the same event in order for this task to be carried out. NLI features prominently in many language processing benchmarks [18, 17], highlighting its importance in progress towards achieving true natural language understanding, and by extension, an understanding of the world.

We argue that the task of NLI is suitable in our endeavor of exploring fuzziness in models due to the amount of flexibility it offers. Models trained to perform NLI are capable of making—in theory—an infinite number of inferences for a given premise, thereby enabling the testing of a number of pre-defined phenomena of interest. For instance, Wang et al. [18] formulate many natural language understanding phenomena such as lexical entailment (*dog* entails *animal*), monotonic reasoning (*I do not have a blue book* does not entail *I have no books*), symmetry (*John married Josh* entails *Josh married John* but *Larry likes Sally* does not entail *Sally likes Larry*), etc. as NLI tasks in order to test a range of models. A caveat in investigating semantic phenomena by studying model predictions on such a task is that the logical relation between the premise and the hypothesis is pre-defined, making it easy for the researcher to relate the model's predictions with its knowledge about the specific phenomena of interest. However, when it comes to the study of vagueness or fuzziness—which is arguably an integral component of understanding the

world, (see section 2)—knowing exactly that a temperature of 50° F entails *cool*
or *warm* is un-achievable due to the epistemic uncertainty surrounding the concept
of vagueness itself [21]. However, the exploration of vagueness within the models'
behavioral responses in an NLI setting can still be possible if one supplies multiple
plausible hypotheses for a given premise, by relying on commonsense knowledge
and intuition. Comparing the behavior of the model in its predictions on the different
hypotheses, then, facilitates roughly exploring the model's encoding of the fuzziness
of the given concept. Revisiting the earlier example, one can compare the extent to
which NLI models prefer attributing an entailment relation to a temperature of 50°
F with a number of fuzzy categories that may encompass the concept of TEMPER-
ATURE, such as *warm, cool, moderate,* etc. In our demonstrative experiments, de-
scribed below, we test an existing NLI model to classify a natural language premise
describing the temperature into five arbitrarily chosen fuzzy categories: *freezing,
cold, cool, warm, hot*.

## 4.1 Methodology

### 4.1.1 Model investigated

We perform our experiments on RoBERTa$_{large}$ [8], fine-tuned on the MNLI dataset
[20]. The RoBERTa model's core architecture is similar to that of BERT [3], i.e.,
it uses a transformer trained to bidirectionally predict missing words in context.
Unlike BERT, it does not have the next sentence prediction task, and is trained on
10-times the amount of text data (160 GB) as compared to BERT (16 GB). To fine-
tune on MNLI, the model accepts two sentences, a premise and a hypothesis, in the
following format:

```
<cls> Premise </s></s> Hypothesis </s>,
```

where $<cls>$ is a classification token, and $</s>$ acts as a sentence separator.
The model estimates scores (in the form of log-probabilities) for the three classes:
entailment, contradiction, and neutral, by applying a linear layer that accepts the
representation for the $<cls>$ token and produces a three-dimensional vector with
scores for each label. We follow this format of representing the input to the model
in our experiments.

**Table 1** Configurations used to generate the analytical dataset. **Note:** The columns "Location
Phrase" and "Fuzzy Category" encompass all rows.

| Unit | Temp. Range | Location Phrase | Fuzzy Category | Total Instances |
|------|-------------|-----------------|----------------|-----------------|
| No unit | $[-50, 122]$ | *No-Location, in the bedroom* | *freezing,* | 5190 |
| Fahrenheit (°F) | $[-50, 122]$ | *in the living room, in the basement* | *cold, cool,* | 5190 |
| Celsius (°C) | $[-50, 50]$ | *outside,inside* | *warm, hot* | 3030 |

### 4.1.2 Stimuli

Like any standard NLI input, our stimuli are composed of pairs of premise and hypothesis sentences. The premise sentences are natural language descriptions of temperature measurements, with the option of a specified location as well as the unit of measurement. For example, *"It is $n$ degrees outside."* For every premise, we construct a hypothesis about each of the plausible fuzzy category to be inferred from the temperature value specified in the premise — *freezing, cold, cool, warm, hot*. For example, *"It is warm in the basement."* Our premise and hypothesis templates are shown below, with text in '{}' indicating an optional element:

> **Premise Template:** It is `[temperature]` degrees {`units`} `[location-phrase]`.
> **Hypothesis Template:** It is `[fuzzy-category]` `[location-phrase]`.

In summary, we vary between three different unit options, six different location phrases, and five different fuzzy sets in generating our premise-hypothesis pairs, amounting to 54 different settings for every degree of measurement. The configurations we apply to generate our analytical dataset is presented in table 1. The location phrases are selected opportunistically. In total we test on 13,410 different sentence pairs.

### 4.1.3 Analysis and Results

To examine the model's behavior in making fuzzy inferences from a given natural language description of a temperature setting, we compute the probability the model assigns to the "entailment" label for every premise-hypothesis pair. Formally, for a given pair $\langle P_i, H_{ij} \rangle$, where $i$ is the temperature in the premise, and $j$ is the perceived fuzzy category specified in the hypothesis, we compute the entailment score $\mathcal{E}$ as:

$$\mathcal{E} = p(\text{"entailment"} \mid \langle P_i, H_{ij} \rangle). \tag{1}$$

Since this computation is probabilistic, we cannot translate it directly into a membership function. We instead focus on the patterns the model shows in its entailment scores for all the fuzzy categories we consider, in order to characterize its behavior in terms of how well it can mimic qualitative properties of membership functions, both typical and atypical. Table 2 shows two randomly selected premises, their corresponding hypotheses, and $\mathcal{E}$-values computed using the model that we study.

The results of our experiments are collectively shown in fig. 2. Each plot in fig. 2 represents the model's $\mathcal{E}$-values for each of the five hypotheses types in a given unit of measurement and location setting. To better understand general patterns in each setting, we smooth out the noise in our measurements by fitting a Generalized Additive Model (GAM) [5]. The first thing to note from our results is that, overall, model shows noisy but qualitatively sound numeric knowledge patterns based on how the raw temperature values relate to the model's synthetic perception — large values typically show greater $\mathcal{E}$-values for the *warm* and *hot* hypotheses than for hypotheses representing *freezing, cold,* and *cool*. This is especially noticeable upon

**Table 2** Examples of the extent to which RoBERTa$_{large}$—fine-tuned on the MNLI dataset [20]—is able to infer one of five arbitrarily chosen temperature categories when given the temperature of a given room (no unit specified). **Note:** "Entailment Scores" are not the same as fuzzy membership values, and are instead used to perform the behavioral analysis of the model.

| Premise | Hypotheses | Entailment Score ($\mathcal{E}$) |
| --- | --- | --- |
| | It is **freezing** in the bedroom. | 0.956 |
| | It is **cold** in the bedroom. | 0.961 |
| It is 0 degrees in the bedroom. | It is **cool** in the bedroom. | 0.962 |
| | It is **warm** in the bedroom. | 0.009 |
| | It is **hot** in the bedroom. | 0.004 |
| | It is **freezing** in the living room. | <0.001 |
| | It is **cold** in the living room. | 0.002 |
| It is 70 degrees in the living room. | It is **cool** in the living room. | 0.928 |
| | It is **warm** in the living room. | 0.902 |
| | It is **hot** in the living room. | 0.713 |

comparing no-unit (fig. 2a) and Fahrenheit (fig. 2b) conditions with that of Celsius (fig. 2c), we see that the model has approximately grasped what is considered to be "cold" in the Celsius condition — the model shows almost zero-membership values for *warm* and *hot* when the temperature specified in the premise is below 0°C and gradually goes up to positive $\mathcal{E}$-values as the temperature increases. In the Fahrenheit condition, however, the model is unable to show exactly the same pattern (0°C = 32°F), suggesting a slight lack of internal consistency. Revisiting the Celsius condition, the model is rather noisy, as compared to its no-unit and Fahrenheit counterparts — it shows similar $\mathcal{E}$-values for *cool, cold, warm,* and *hot* categories in this condition, deviating slightly from intuitive perceptions about temperature. For instance, at 50°C, the perception of the model leans more towards *cold* than *hot*. The model's patterns when no unit of measurement is specified in the premise show strong correspondence with that in the Fahrenheit condition (RMSE[2] between No-unit and Fahrenheit is 0.09, while that between No-unit and Celsius is 0.16). This suggests that when no information is available, the model defaults to inferring Fahrenheit as the unit of measuring temperature. We hypothesize that this is likely because the model is primarily trained on American-influenced text (e.g. Reddit).

In the context of computational treatment of fuzziness, we find the model to show strikingly noticeable similarities to fuzzy hedges [7, 23]. More prominently, the $\mathcal{E}$-values for *hot* bear qualitative resemblance to a 'concentration' transformation of the $\mathcal{E}$-values for *warm*. That is, if we take the membership of *warm* and raise it to the power of some value greater than 1 (say, $\lambda$), we should get a hedged version of *warm* ($H$ = *very, extremely*, etc.) that could potentially be mapped to the *hot* category, pulling the curve inward, as proposed by Zadeh [23]:
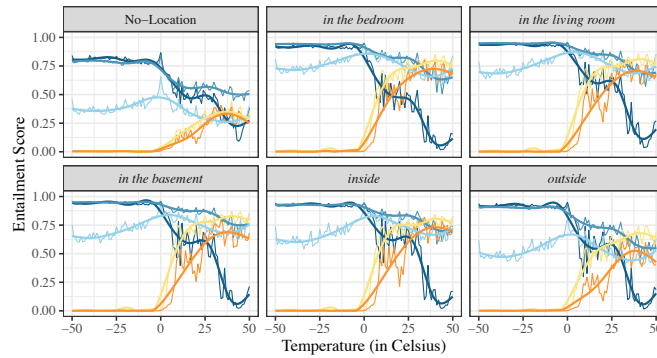
---

[2] root mean-squared error
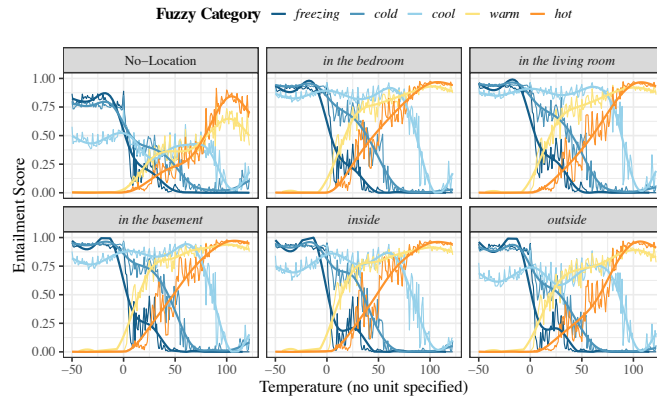
(a)
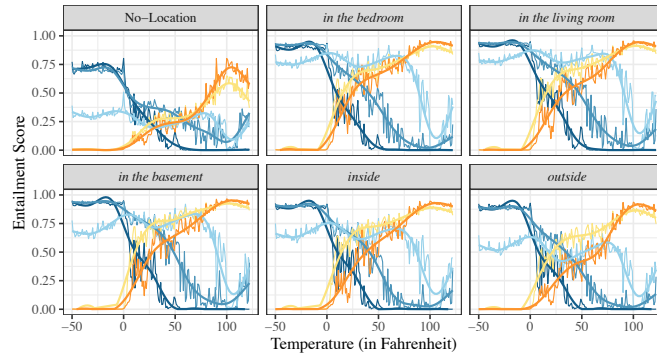


(b)



(c)

**Fig. 2** Entailment scores (y-axis) from RoBERTa$_{large}$ for various fuzzy categories, given premise with specific temperature (x-axis): **(a)** No unit; **(b)** Fahrenheit; and **(c)** Celsius. Scores are smoothed using a generalized additive model (with default parameters, in R).

$$hot \approx H \; warm = \int_U \mu^\lambda(x) \qquad\qquad (\lambda > 1)$$

Empirically, we perform a naïve search between $[1, 8]$ (100 values) and minimize the RMSE between $\mathcal{E}_{hot}$ and $\mathcal{E}_{warm}^\lambda$ to find the best value for $\lambda$ as per the above equation, and show results in fig. 3. We find the best $\lambda$ value to be close to 2 in all three cases (with Fahrenheit showing the greatest deviation), suggesting that *hot* is approximately equal to *very warm* (if Zadeh's recommendation in [23] is to be followed for *very*). We also find the $\mathcal{E}$-value for *warm* to be greater than that of *hot* approximately $77.64\%, 75.33\%$, and $98.51\%$ of the time for the no-unit, Fahrenheit, and Celsius conditions, respectively. This empirically supports the above notion that highlights the possibility of hedging *warm* to produce *hot*, and also largely follows the 'fuzzy' interpretation of semantic entailment ($P$ (*hot* or hedged-*warm*) entails $Q$ (*warm*) iff $\mu_P(x) \leq \mu_Q(x)$), as noted by [7], with its strongest effect being included in the Celsius condition. Finally, we also observe a tendency of the *hot* and *warm* curves to show a decrease in extremely hot temperatures, suggesting the existence of another fuzzy set that is applicable to more extreme temperatures. This is likely to be an antonym of *freezing* or something that is synonymous to *very hot*. However, since there isn't a clear word with comparable frequency,[3] we leave this analysis for future work.
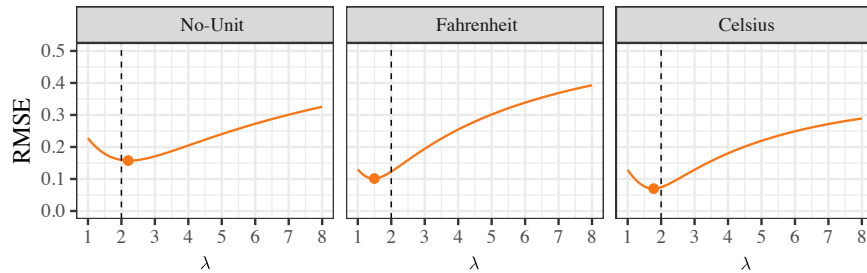


**Fig. 3** Points representing the best $\lambda$ values for minimizing $\mathrm{RMSE}(\mathcal{E}_{hot}, \mathcal{E}_{warm}^\lambda)$ using a naïve search among 100 values in $[1, 8]$. Dashed line represents $\lambda$ corresponding to *very* as per Zadeh [23].

## 5 General Discussion and Conclusion

More often than not, natural language expressions have a graded—rather than absolute—measurement of truth and falsehood [7]. This graded property of language owes itself to the vague boundaries in concepts such as TALL, FRUIT, etc.,

---

[3] we experimented with *scorching*, *blistering*, and *balmy*, but all three have considerably greater differences in frequency with *freezing*, as compared to that between *cold-hot* and *cool-warm*. This can be verified by using a public resource such as Google's n-gram viewer.

that do not have a clear-cut criteria for membership. The development of fuzzy set theory [22] has provided us with a useful toolkit to handle imprecision in language, e.g. "*the coffee is very warm*," by formulating membership functions to model data. We argue in this paper that computational models—such as PLMs—that learn language representations must account for the fuzziness contained in language use. To this end, we formulated a demonstrative experiment by investigating a current state of the art model (RoBERTa$_{large}$) for the classical fuzzy case of TEMPERATURE—to what extent can it map natural language premises such as *"it is 40°F inside"* to hypotheses carrying fuzzy categories such as *"it is cold inside"* Our tests follow the task of natural language inference [1], which allows broad-coverage testing of a variety of semantic phenomena encoded in models. Furthermore, the model we test was fine-tuned to perform the NLI task, making our investigation a natural environment for the model. We primarily focus on our model's behavioral responses to the supplied premise-hypotheses pairs by examining its entailment-scores (the degree to which the premise entails the hypothesis, according to the model) for various fuzzy categories mentioned in the hypotheses, and analyse their patterns. Overall, we found the model to show a noisy-but-sound grasp of temperature perception, with substantial differences between cold and warm temperature categories towards the extreme ends of the scale. While the model showed sound patterns of temperature perception across our unit-conditions, we found it to show slightly lower correspondence with intuition when the temperature mentioned in the premise was in Celsius. This, together with the fact that its patterns on the Fahrenheit condition are found to be strongly similar to that in the no-unit condition suggest that the model has a bias towards American English. In our main analyses, the model showed similar patterns in its entailment scores across the fuzzy categories as standard membership functions formulated for showing hedging effects. For instance, the model's entailment scores for the category of *hot* mimicked the entailment-scores of a concentrated version of *warm*, pulling its curve inward. However this was not the case in the colder temperatures, suggesting a lack of consistency in this behavior. This indicates, tentatively, that state of the art models show noticeable signs of approximately showing patterns akin to that of fuzzy membership functions in graded cases such as TEMPERATURE. We therefore use this paper as a call towards a systematic study of the manifestation of vagueness in state of the art language models, as well as how multi-valued logic such as fuzzy logic can be incorporated within them in order to better handle imprecise language.

# References

1. Bowman, S., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 632–642 (2015)
2. Carvalho, J.P., Batista, F., Coheur, L.: A critical survey on the use of fuzzy sets in speech and natural language processing. In: 2012 IEEE international conference on fuzzy systems, pp. 1–8. IEEE (2012)

3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019)
4. Firth, J.R.: A synopsis of linguistic theory, 1930-1955. Studies in linguistic analysis (1957)
5. Hastie, T.J., Tibshirani, R.J.: Generalized additive models, vol. 43. CRC press (1990)
6. Labov, W., Bailey, C., Shuy, R.: The boundaries of words and their meanings. 1973 pp. 340–73 (1973)
7. Lakoff, G.: Hedges: A study in meaning criteria and the logic of fuzzy concepts. In: Contemporary research in philosophical logic and linguistic semantics, pp. 221–271. Springer (1975)
8. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692 (2019)
9. Martine, D.: Linguistic hedges: a quantifier based approach. Soft Computing Systems: Design, Management and Applications **87**, 142 (2002)
10. McCloskey, M.E., Glucksberg, S.: Natural categories: Well defined or fuzzy sets? Memory & Cognition **6**(4), 462–472 (1978)
11. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, vol. 26, pp. 3111–3119 (2013)
12. Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., Miller, A.: Language models as knowledge bases? In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 2463–2473 (2019)
13. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
14. Rosch, E.: Cognitive representations of semantic categories. Journal of experimental psychology: General **104**(3), 192 (1975)
15. Van Deemter, K.: Not exactly: In praise of vagueness. Oxford University Press (2012)
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30**, 5998–6008 (2017)
17. Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: Superglue: A stickier benchmark for general-purpose language understanding systems. In: Advances in Neural Information Processing Systems, pp. 3266–3280 (2019)
18. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: Glue: A multi-task benchmark and analysis platform for natural language understanding. In: International Conference on Learning Representations (2018)
19. Weir, N., Poliak, A., Van Durme, B.: Probing neural language models for human tacit assumptions. In: S. Denison, M. Mack, Y. Xu, B.C. Armstrong (eds.) Proceedings of the 42nd Annual Conference of the Cognitive Science Society, pp. 377–383. Cognitive Science Society (2020)
20. Williams, A., Nangia, N., Bowman, S.: A broad-coverage challenge corpus for sentence understanding through inference. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 1112–1122. Association for Computational Linguistics (2018)
21. Williamson, T.: Vagueness. Routledge (1994)
22. Zadeh, L.A.: Fuzzy sets. Information and Control **8**, 338–353 (1965)
23. Zadeh, L.A.: A fuzzy-set-theoretic interpretation of linguistic hedges (1972)
24. Zadeh, L.A.: Fuzzy logic = computing with words. In: Computing with Words in Information/Intelligent Systems 1, pp. 3–23. Springer (1999)