Modeling language constructs with compatibility intervals

Pavlo Kapustin
University of Bergen
pavlo.kapustin@uib.no

Michael Kapustin
Moscow Institute of Physics and Technology
michael.kapustin@gmail.com

Abstract

We describe a representation for modeling meaning of natural language constructs that is closely related to fuzzy sets. Same as fuzzy sets, it allows to express quantitative relationships between different concepts, and is designed to support vagueness and imprecision common to natural language. We compare the representations using several examples, and argue that in some cases the proposed representation may be a good alternative to the fuzzy set based representation, and that it may also be easier to learn from data.

1 Introduction

Consider the sentence "A young lady came into the room". The word "lady" informs us about the gender of the person, and both words in the construct "young lady" tell us something about the age of the person. To be able to make such inferences one needs quantitative information about the relation of the construct "young lady" to the property "age", i.e. what ages are compatible with this description.

One representation that allows to explicitly capture this type of quantitative information is based on fuzzy sets and suggested by Lotfi Zadeh in his earlier papers (Zadeh, 1971, 1972). However, there is currently not much research in applying this representation to problems of computational linguistics and natural language processing (Carvalho et al., 2012; Novák, 2017).

We believe that one of the reasons for this may be that membership functions are relatively complex objects, and this may possibly increase the complexity of operations like different transformations and learning from data when using the fuzzy set based representation. In this paper we briefly describe a representation closely related to fuzzy sets that we call compatibility intervals, and argue that in some cases it may be a good alternative to the fuzzy set based representation, and may also be easier to learn from data.

2 Related work

In his early works, Lotfi Zadeh suggests modeling meaning of certain types of adjectives (e.g. "small", "medium", "large") as fuzzy sets, and some lingustic hedges (e.g. "very", "slightly" — as operators, acting on these fuzzy sets (Zadeh, 1971, 1972). Hersh and Caramazza (1976) introduce logical and linguistic interpretations of membership functions, showing the difference between them in an experimental setting.

Novák (2017) describes Fuzzy Natural Logic, a mathematical theory attempting to model the semantics of natural language, that includes Theory of Evaluative Linguistic Expressions (Novák, 2008).

In M. Kapustin and P. Kapustin (2015) we describe a framework for computational interpreting of natural language fragments, and suggest modeling meaning of words as operators. P. Kapustin (2015) describes an application that implements and tests some features of this framework in a simplified setting.

Runkler (2016) describes an approach for generation of linguistically meaningful membership functions from word vectors.

In P. Kapustin and M. Kapustin (2019b) we describe a couple of approaches that can be used for modeling meaning of natural language constructs using fuzzy sets and discuss some examples. We

discuss how people relate some language constructs to compatibility intervals in an experimental study (P. Kapustin and M. Kapustin, 2019a).

Schwarzchild and Wilkinson (2002) describe interval-based semantics for comparatives. Abrusán and Spector (2008) describe semantics of degree questions based on intervals.

3 Different interpretations

Lotfi Zadeh's further work on linguistic variables (Zadeh, 1975a, 1975b, 1975c) and possibility theory (Zadeh, 1999) introduce the term "compatibility", clarifying interpretation of the values of the membership functions: they can be seen as degrees of compatibility between the value of the function argument and the construct that is described by the membership function.

Similarly to Hersh and Caramazza (1976), we distinguish between logical and linguistic interpretations of membership functions. Consider fig. 1: "young1" corresponds to the logical interpretation, reflecting the fact that infants and newborns are, indeed, as young as one can be. On the other hand, "young2" corresponds to the linguistic interpretation, reflecting the fact that when we use the word "young", we usually refer to ages other than newborns and infants. Similarly, we normally do not use "often" when something occurs always or almost always, and we normally do not use "seldom" when something occurs never or almost never. On the other hand, the usage of word "old" does not seem to differ from what is "logically" correct: we may say "old" about someone who is 80 or 100 years old.

We think that some differences between logical and linguistic interpretations may be related to scalar implicatures and similar phenomena, and believe that this needs to be investigated further.

Difference between logical and linguistic interpretations has some interesting implications. Consider fig. 2. Here we apply negation, implemented as Zadeh's complementation (Zadeh, 1972), to constructs "young1", "young2" and "old". While such negation seems to work well with the logical interpretation, it gives somewhat unexpected results with the linguistic interpretation: according to $not(\mu_{\text{young2}})$, it appears that infants are less "not young" than newborns, which is not correct.

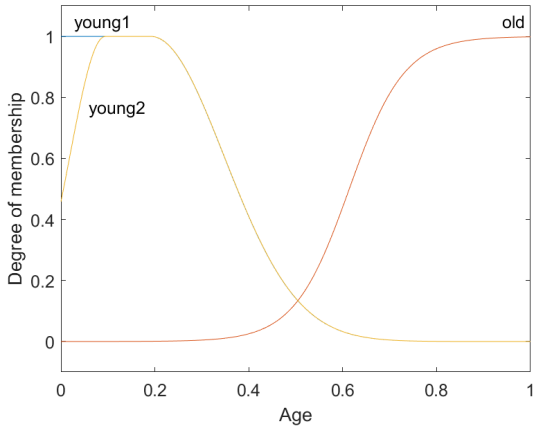

Figure 1: Different meanings of "young".

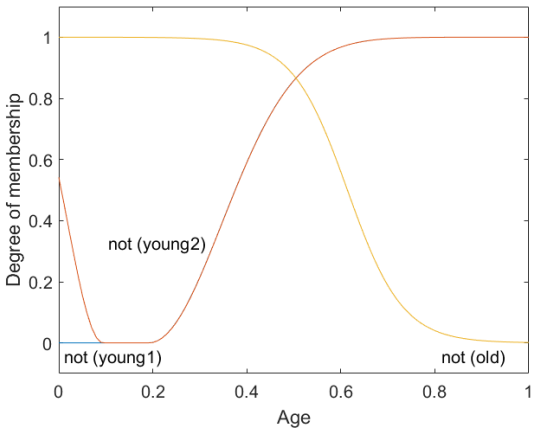

Figure 2: Logical "not", applied to "young1" and "young2".

Consider fig. 3. Here we apply intensifier "very", implemented as Zadeh's concentration (Zadeh, 1972), to constructs "young1", "young2" and "old". Again, while such implementation of "very" seems to work well with the logical interpretation, it gives somewhat unexpected results with the linguistic interpretation: according to $very(\mu_{\text{young2}})$, it appears that infants are "very young" to a lower degree than they are "young". In addition, the width of the interval in which the membership degree is equal to one is unaffected by the application of "very", which seems to be incorrect.

We believe that logical and linguistic interpretations complement each other, each of them modeling different aspects of the meaning of natural language constructs, and for some words may need to be modeled as separate membership functions.

We think that finding mathematical functions and their transformations that correctly capture important details about language constructs may not always be that easy. This seems to become even more challenging if we are working with the linguistic interpretation, as in this case the shape of the functions often becomes more complex.

4 Representation based on compatibility intervals

Here we briefly describe the representation based on compatibility intervals and discuss some examples.

Compatibility interval is an interval of property values on some scale that are compatible with a given language construct (here term "property" is used in a relatively general sense). The discussion of scales is a large topic of separate research, and is out of scope of this paper.

Compatibility intervals consist of the main subinterval with high compatibility, and optional left ("increasing") and right ("decreasing") subintervals adjacent to the main subinterval. The following invariants are maintained: all the values in the main subinterval have equal (high) compatibility, and the closer the values are to the main subinterval the higher their compatibility is.

Here we provide examples of some intervals for different age groups (we use double hyphens between the start and the end of the main subinterval, and single hyphens between the start and the end of the left and the right subintervals).

```
child: [0 -- 15-20]
young1: [0 -- 30-50]
young2: [0-18 -- 30-50]
adult: [15-20 -- 100]
middle-aged: [40-45 -- 65-70]
old: [70-80 -- 100]
```

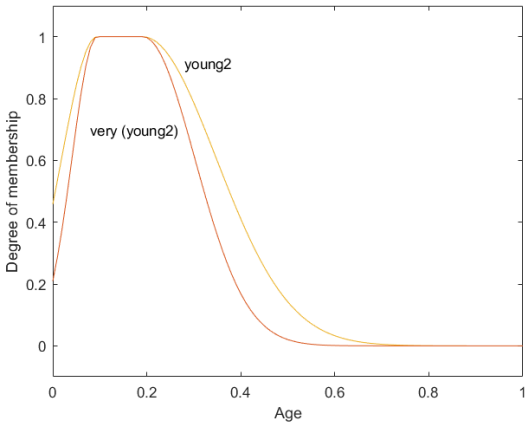

Figure 3: Intensifier "very", applied to "young2".

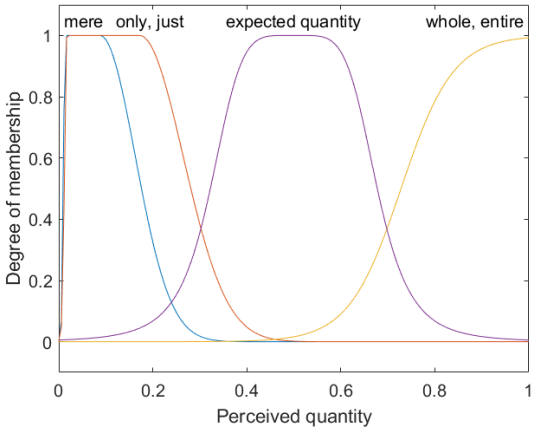

Figure 4: "Mere", "only", "just", "whole" and "entire" related to perceived quantity.

Compatibility intervals may have including or excluding boundaries.

In P. Kapustin and M. Kapustin (2019b) we suggest how words "mere", "only" and "just" can be modeled using fuzzy sets (fig. 4). These words may be used with quantities that are perceived as very small, but not with zero quantities (neither "a mere zero", "only zero" or "only no one" seem to make sense). This fact is not very convenient to model using membership functions. Using the representation based on compatibility intervals, we could model this with an excluding boundary:

```
mere: (0 -- 0.1-0.3]
only, just: (0 -- 0.2-0.4]
expected quantity: [0.2-0.4 -- 0.6-0.8]
whole, entire: [0.6-0.8 -- 1]
```

In case of compatibility intervals, instead of using one membership function to model the relation of a language construct to a certain property, one or several compatibility intervals can be used (several intervals would correspond to a membership function with more than one maximum, something that could be used for modeling ambiguity). For example, suppose that "teenager" is defined by the interval:

```
teenager: [10-13 -- 17-20]
```

Then something corresponding to "not teenager" with meaning similar to Zadeh's complementation (Zadeh, 1972) could be modeled by inverting the interval, for example:

```
[0 -- 10-13)
(17-20 -- 100]
```

Consider the construct "not teenager" in the context of the phrase "you are not a teenager" (that may mean "younger child" or "adult"). Compared to membership functions, as long as intervals don't have membership values, if we would like to represent which of the two intervals corresponding to "not teenager" is more "possible", we would need to store this information separately or extend the representation to optionally contain the membership degree of the main subinterval.

Same as with membership functions, if we are interested in both logical and linguistic interpretations, we need to store two different intervals (or two different sets of intervals, if a construct is represented by several intervals) for some constructs.

Also, as with membership functions, if we attempt to invert interval "young2", trying to model negation similar to Zadeh's complementation (Zadeh, 1972), we would get unexpected results (similar to what is described on fig. 2).

On the other hand, it seems to be easier to model intensifier "very" for the linguistic interpretation using compatibility intervals. If we would like to do something similar to Zadeh's concentration (Zadeh, 1972), we could, for example, use a very simple model: for constructs whose compatibility interval starts at the left end of the scale, we could leave the left boundary of the left subinterval unchanged, multiplying all the other boundaries with a certain factor. For constructs whose compatibility interval ends at the right end of the scale, we could do the opposite: leave the right boundary of the right subinterval unchanged, multiplying all the other boundaries with the inverse of the same factor.

```
young2: [0-18 -- 30-50]
very (young2): [0-16.20 -- 27-45]
old: [60-80 -- 100.0]
very (old): [66.67-88.89 -- 100]
```

Note that we no longer have the unexpected result compared to what was described on fig. 3.

This example is provided to illustrate that it may be easier to control what effects are achieved by doing different transformations when using compatibility intervals, rather than membership functions, because one can move boundaries of the subintervals independently.

5 Discussion

Presented representation, while being closely related to fuzzy sets, is different in some important respects. Compatibility intervals move attention away from specific mathematical functions, curve shapes and membership degrees, rather focusing on the intervals of values that are compatible with the language construct.

Compared to membership functions, compatibility intervals are relatively simple objects that are defined by several numbers. That's why we hope that certain operations, including different transformations and learning from data, may be easier to implement with compatibility intervals than with fuzzy sets.

We believe that moving the focus away from membership degrees has mostly positive effect, as one only needs to think about interval boundaries when implementing operations on intervals. However, we can imagine situations when explicit membership degrees may be desirable. We described one such case related to ambiguity, along with possible solutions.

In general, we think it is positive that there is no clear boundary between the representation based on fuzzy sets and the representation based on compatibility intervals, which means that in some cases it may be possible to use benefits of both of the representations.

We discussed why it is important to distinguish between logical and linguistic interpretations when modeling meaning of natural language constructs using fuzzy sets or compatibility intervals.

Representations based on fuzzy sets and compatibility intervals, being closely related, are both able to quantitatively represent what language constructs tell us about certain properties, allowing to capture important aspect of meaning of these constructs. That's why we hope that more researchers in the field of computational linguistics and natural language processing become interested in this area.

Acknowledgements

We thank Vadim Kimmelman and Csaba Veres for helpful discussions and comments. We thank anonymous reviewers for helpful feedback.

References

Abrusán, M., and B. Spector (2008). An interval-based semantics for degree questions: negative islands and their obviation. In *Proceedings of the 27th West Coast Conference on Formal Linguistics, Somerville, MA,* 17–26. Citeseer.

Carvalho, J. P., F. Batista, and L. Coheur (2012). A critical survey on the use of fuzzy sets in speech and natural language processing. In *Fuzzy Systems (FUZZ-IEEE), 2012 IEEE International Conference on,* 1–8. IEEE.

Hersh, H. M., and A. Caramazza (1976). A fuzzy set approach to modifiers and vagueness in natural language. *Journal of Experimental Psychology: General* 105 (3): 254.

Kapustin, M., and P. Kapustin (2015). Modeling meaning: computational interpreting and understanding of natural language fragments. *arXiv preprint arXiv:1505.08149.*

Kapustin, P. (2015). Computational comprehension of spatial directions expressed in natural language. Master's thesis, The University of Bergen.

Kapustin, P., and M. Kapustin (2019a). Language constructs as compatibility intervals: an experimental study. In preparation.

Kapustin, P., and M. Kapustin (2019b). Modeling language constructs with fuzzy sets: some approaches, examples and interpretations. In submission.

Novák, V. (2008). A comprehensive theory of trichotomous evaluative linguistic expressions. *Fuzzy Sets and Systems* 159 (22): 2939–2969.

Novák, V. (2017). Fuzzy logic in natural language processing. In *Fuzzy Systems (FUZZ-IEEE), 2017 IEEE International Conference on,* 1–6. IEEE.

Runkler, T. A. (2016). Generation of linguistic membership functions from word vectors. In *Fuzzy Systems (FUZZ-IEEE), 2016 IEEE International Conference on,* 993–999. IEEE.

Schwarzchild, R., and K. Wilkinson (2002). Quantifiers in comparatives: A semantics of degree based on intervals. *Natural language semantics* 10 (1): 1–41.

Zadeh, L. A. (1971). Quantitative fuzzy semantics. *Information sciences* 3 (2): 159–176.

Zadeh, L. A. (1972). A fuzzy-set-theoretic interpretation of linguistic hedges.

Zadeh, L. A. (1975a). The concept of a linguistic variable and its application to approximate reasoning—I. *Information sciences* 8 (3): 199–249.

Zadeh, L. A. (1975b). The concept of a linguistic variable and its application to approximate reasoning—II. *Information sciences* 8 (4): 301–357.

Zadeh, L. A. (1975c). The concept of a linguistic variable and its application to approximate reasoning-III. *Information sciences* 9 (1): 43–80.

Zadeh, L. A. (1999). Fuzzy sets as a basis for a theory of possibility. *Fuzzy sets and systems* 100 (1): 9–34.