Towards Semantic Classification of Dialog using Contextual Prediction Networks

A Doboli (alex.doboli@stonybrook.edu)
Department of Electrical and Computer Engineering, Stony Brook University
Stony Brook, NY 11794-2350 USA

G Villuri (gnaneswar.villuri@stonybrook.edu)
Department of Electrical and Computer Engineering, Stony Brook University
Stony Brook, NY 11794-2350 USA

Abstract

Semantic classification distinguishes inputs based on their meaning (e.g., interpretation) not their static features, as in traditional classification. Existing transformer models seem to have limited capabilities for semantic classification. This paper presents our ongoing work on the semantic classification of the dialog sentences produced by human subjects during problem solving, including the used data set, and the gained insight from using transformer models for classification. A new theoretical model, called Contextual Prediction Networks, is suggested for semantic classification of dialog sentences.

Keywords: semantic classification; dialog; explainability; transformer models

Introduction

Theories in the philosophy of art argue that art objects possess both exhibited (EXP) and nonexhibited properties (NEXP) (Dickie, 1969). EXPs are physical, visible properties, like shapes, color, texture, etc., while NEXPs represent meaning-producing interpretations of EXPs, created using the norms of a given historical, social, political, or artistic context. For example, religious paintings of the Italian Renaissance incorporate a certain symbolistic (NEXPs) of the exhibited scenes (EXPs) that is defined by the beliefs of that period (Baxandall, 1985). Hence, understanding art, including tasks like placing (classifying) artwork into different art genres, requires considering both EXPs and NEXPs. It can be argued that the importance of EXPs and NEXPs in understanding semantics is important not only in art but for any human activity in general. This paper focuses on analyzing the importance of EXPs and NEXPs in understanding some semantic aspects of human dialog during programming problem solving.

We define *semantic classification* as the activity of classifying inputs, e.g., the speech sentences during dialog, based on their interpretation, hence NEXPs. In contrast, traditional classification methods utilize only EXPs, which are static, expressed features of the input data. For example, the sentence ``So do we add three times three to the array?'' produced during a programming problem solving exercise by human subjects was classified by DistilBERT (Sanh, Debut, & et al., 2020) to indicate an analysis step, such as the subjects analyzing their code. But equally well, the sentence could be an elaboration step, in which new details are added to the solution, with the subject asking for the opinion of the other participants. The actual meaning of the sentence results not only from its EXPs, e.g., the labels representing the words in the sentence, but also its interpretation, such as the role the sentence has in the flow of the dialog during problem solving.

This paper summarizes our ongoing work on the semantic classification of the dialog sentences produced by human subjects during programming problem solving to understand how the meaning of the sentences influenced the success or failure of the solving process. The paper describes the semantic classification problem and the data set utilized to study the problem. The insight obtained by using state-of-the-art transformer models, like BERT (Devlin, Chang, & et al., 2019), DistilBERT (Sanh et al., 2020) and Roberta (Liu, Ott, & et al., 2019), for semantic classification is also discussed. The paper ends by suggesting a new theoretical model for semantic classification based on the observed limitations of the transformer models in generating explainable semantic classification. The new model is called Contextual Prediction Networks (CPNs).

Semantic Classification

The studied semantic classification problem was to classify the dialog sentences into five categories depending on their role in problem solving. To simplify the presentation, each category was also labeled using a separate color. The five categories used for classification were as follows: (i) Analysis of the problem requirements (color Yellow), (ii) Formulation of the overall solution approach (color Grey), (iii) Elaboration (detailing) of the solution (color Blue), (iv) Analysis of the solution (color Green), and (v) Modifying the solution (color Red). The remaining part of the document refers to the color labels instead of the category types.

This problem requires semantic classification as the decision to which category a statement is added depends on the effect of performing (interpreting) the statement, and not only on the words (seen as static labels) that form the sentence. A high similarity of the effects of two sentence determines that they pertain to the same category, not necessarily the similarity of their words, like in traditional classification. For example, the sentences ``compute the sum after initializing their values'' and ``find the product after reading the input file'' belong to category Elaboration of the solution (color Blue), even though their words are dissimilar. However, their actions are similar. The degree to which the outcome of a sentence (NEXPs) can be calculated only from its composing words (EXPs) is unknown.

Data Set

The collected data represent verbal discussions during programming problem solving. Thirty teams of undergraduate students (i.e. three students in a team) were required to create programming code to solve a problem. Each team individually worked for twenty minutes. The verbal discussions between the team members were recorded and utilized for automated speaker tracking and then converted into text (Duke & Doboli, 2022a, 2022b). The dataset sizes for the thirty teams were between 56 and 289 sentences, with a total of 3714 sentences that were used for classifier training and testing.

Transformer Performance

Three transformer models, BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2020) and Roberta (Liu et al., 2019), were trained on the data set comprising the sentences recorded during the problem solving experiment. Table 1 summarizes the classification performance of the transformer models.

Table 1: Performance of the transformer models

Model	Green	Blue	Grey	Yellow	Red
BERT	79.70%	77.70%	54.70%	34%	0%
DistilBERT	85%	73%	47.70%	14%	0%
Roberta	27.30%	76.70%	0%	0%	0%

Note that DistilBERT offered the best performance followed by BERT. Roberta had the lowest performance. Sentences in category Green were recognized with the least error by both DistilBERT and BERT followed by classifying sentences in category Blue. The classification error is higher for the other three categories, Grey, Yellow, and Red. The results are explained by the fact that most of the sentences in the data set are in categories Green and Blue, with significantly fewer sentences belonging to the other three categories.

Insight into the Transformer Classification

The classification results of DistilBERT were analyzed for explainability using the tool Transformers Interpret (*https://github.com/cdpierse/transformers-interpret*, n.d.). The explainability outputs offer attribution scores, which describe the degree to which the words in a sentence positively or negatively assign the sentence to a category.

The explainability analysis shows that some words and punctuation marks, like “yeah”, “?”, “it”, “that”, “this”, “we”, “to”, “okay”, and so on, are positively linked to category Green, while words, e.g., “comparing”, “file”, “inside”, “sorted”, “less” etc., are negatively linked. Some of these elements are similar to what manual analysis indicated, i.e. questions are a good predictor of category Green, however, other elements are less indicative, as attributed and adverbs, like “sorted” and “less”, are often also good predictors of category Green. Similarly, words, like “are”, “to”, “think”, “like”, etc., were found well linked to category Grey, while words, e.g., “text”, “function”, “output”, “equals”, etc., were negatively linked, even though these words are less likely to be used by a human in deciding the categories of the sentences.

Therefore, we concluded that while the classification accuracy of DistilBERT is high, the words utilized in classification are not similar to the words and word sets that humans use in interpreting the sentences. The explainability was low.

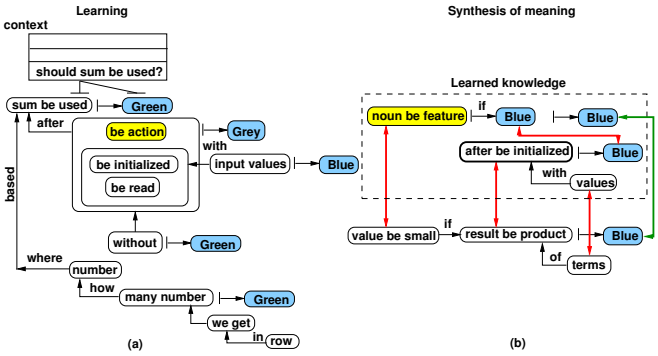

Figure 1: Contextual Prediction Networks.

Proposed Model for Semantic Classification

Figure 1 shows the proposed networked model, called Contextual Prediction Networks (CPNs), to not only semantically classify the sentences but also produce explainability insight that is more similar to human understanding. The left figure shows the learning part of the networked model, and the right figure illustrates the synthesis of the meaning of a sentence needed to classify it using explainable features.

Learning. The networked model (Figure 1(a)) includes three parts: clusters, context, and arcs. Clusters are sets of words with similar interpretation, like “be initialized” and “be read” in the figure. While their word similarity is low, their action similarity (i.e. interpretation) is high. Depending on the word similarity of the word sets assigned to the same cluster, more abstract patterns are found for a cluster, like the pattern “be action” highlighted in yellow in the figure, where “action” represents a verb describing an action. Context is formed of the sentences used in a certain time window before the current sentence. Finally, arcs connect the word sets in a sentence depending on the specific prepositions, like “after”, “with”, “if”, “without”, on so on. Arcs $\mapsto$ show the explainability information, like which words produce the assigned category.

Synthesis of meaning. The learned structures are used to find the category of a new sentence through a matching process. The matched structures are indicated using red arrows in Figure 1(b). For example, the matching of “after be initialized with values” (present in the knowledge structure and associated with category Blue) with the input “result is product of sums”, assigns category Blue to this input fragment too. This is an interpretation of the fragment. Then, matching associates the input fragment “value is small” to the more abstract rule “noun be feature” (shown in yellow), which produces category Blue, if the rule is linked to a word set already interpreted as category Blue. As this interpretation was already produced, the category of the entire input sentence is category Blue.

References

Baxandall, M. (Ed.). (1985). *Patterns of intention. on the historical explanation of pictures*. Yale University Press, New Haven and London.

Devlin, J., Chang, M.-W., & et al. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding* (Tech. Rep. No. 1810.04805). arXiv.

Dickie, G. (1969). Defining art. *American Philosophical Quarterly*, *6*.

Duke, R., & Doboli, A. (2022a). Applications of dialogic system in individual and team-based problem-solving applications. In *Proceedings of the ieee international symposium on smart electronic systems (ises).*

Duke, R., & Doboli, A. (2022b). *dialogic: Non-invasive speaker-focused data acquisition for team behavior modeling* (Tech. Rep. No. 2209.00619). arXiv.

https://github.com/cdpierse/transformers-interpret. (n.d.).

Liu, Y., Ott, M., & et al. (2019). *Roberta: A robustly optimized bert pretraining approach* (Tech. Rep. No. 1907.11692). arXiv.

Sanh, V., Debut, L., & et al. (2020). *Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter* (Tech. Rep. No. 1910.01108). arXiv.