# An Approximate Perspective on Word Prediction in Context: Ontological Semantics meets BERT

Kanishka Misra and Julia Taylor Rayz

**Abstract** This paper presents an analysis of a large neural network model – BERT, by placing its word prediction in context capability under the framework of Ontological Semantics. BERT has reportedly performed well in tasks that require semantic competence without any explicit semantic inductive bias. We posit that word prediction in context can be interpreted as the task of inferring the meaning of an unknown word. This practice has been employed by several papers following the Ontological Semantic Technology (OST) approach to Natural Language Understanding. Using this approach, we deconstruct BERTs output for an example sentence and interpret it using OSTs fuzziness handling mechanisms, revealing the degree to which each output satisfies the sentences constraints.

## 1 Introduction

Recent progress made by deep learning approaches in natural language processing (NLP) have led to the emergence of highly parameterized neural network models that represent a word in its context, collectively known as contextualized word embeddings (CWE). The goal of these embeddings is to adapt to context (described by sentences) for the same word. This means that a word *table* should be represented differently depending whether it is furniture or chart. One such CWE, BERT [1] learns word representations by using a training procedure known as Masked Language Modelling, which is similar to Cloze Tasks [20]. In this task, a word in a sentence (typically called a "cloze sentence") is hidden or "masked and the task is to identify the masked word given the context it occurs in, an example is shown in (1). For this example BERT predicts the word bank in place of the mask with greater than 0.96 probability.

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

(1) I went to the ______ to withdraw some money.

Unfortunately, BERT representations and mapping to the word, while robust and impressive in scale, can be somewhat questionable in quality. In this paper we deconstruct the task of predicting a word in context by borrowing from the school of Ontological Semantics [10], and its latest product, the Ontological Semantic Technology (OST) [5, 11, 15], which is inherently fuzzy in nature [16]. We analyze simple cloze sentences by making fuzzy inferences with the help of the OST system and represent the outputs of BERT by their corresponding concepts that form various solutions to the cloze task depending on their fuzzy membership which is calculated based on the concepts that occur in their context.

## 2 Bidirectional Encoder Representations from Transformers (BERT)

BERT is a language representation neural network model that learns to represent words in sentences by jointly conditioning on words to the left of a target as well as to the right. The representation of a word (a vector) is computed by estimating word probabilities in context, and thus the model produces context-sensitive or contextualized representations of words. Its underlying architecture is based on Transformers [21], which enables it to represent each word as a function of words occurring in its context. The model comes in two variants — BERT-base and BERT-large, differing in the total number of parameters — 110M and 340M respectively. Although the exact nature of these outputs is largely unknown, [13] found BERT's representations of words with similar senses to cluster together in vector space, signaling to some extent that BERT captures sense-specific properties of words through its training mechanism. As a result, BERT advanced the state-of-the-art in NLP (at the time of its publication) by facilitating fine-tuning on a wide variety of language tasks such as Question Answering, Natural Language Inference, etc. The model accepts two sentences as input during each step and is jointly optimized using the following objectives: (1) Masked Language Modelling (MLM), inspired by the cloze task, in which the model uses its context to predict hidden tokens in the input, and (2) Next Sentence Prediction, in which the model predicts whether the second sentence follows the first sentence. Due to its MLM objective, BERT is not considered to be an incremental language model (such as models using Recurrent Neural Networks or its variants) that form sentences by predicting words one by one in a single and fixed direction (left to right) and contain a sequential inductive bias. In our analysis of BERT in this paper, we will analyze the BERT-base model, but this can be extended to any similar language model.

## 2.1 Semantic Capabilities of BERT

Fine tuning BERT has resulted in incremental performance on NLP tasks that require a high linguistic competence. As a result, a myriad of methods have been used to probe BERT for the various linguistic properties that it captures. A majority of such methods have focused on BERT's knowledge of syntactic phenomena such as number agreement [8] and garden-path [14]. While BERT shows syntactic competence on a variety of tasks, it has been found to be less sensitive to an analysis of semantic properties.

Such tasks can be adapted from adjacent disciplines that test human competence. Adapting from the psycholinguistic apparatus of human sentence processing — the N400 experiment — Ettinger [2] developed a suite of tests to analyze BERT's sensitivity for semantic phenomenon observed in humans. BERT was found to show a level of insensitivity in several of these tasks. Specifically, for tests of negation, BERT is unable to assign a lower probability for *bird* as compared to a nonsensical — in the given context — word *tree* in the stimulus: *A robin is not a* ___. Further, it generated non-sequitur situations in tests for commonsense inference, such as predicting words such as *gun* in the stimuli: *The snow had piled up on the drive so high that they couldn't get the car out. When Albert woke up, his father handed him a* ___. However, it showed positive results in attributing nouns to their hypernyms [1] and was sensitive to role-reversal stimuli, such as assigning higher probability to the word *served* in the stimuli - *the restaurant owner forgot which customer the waitress had* ___ as opposed to its role-reversed counterpart, *the restaurant owner forgot which waitress the customer had* ___.

Misra, Ettinger, and Rayz [9] investigated the degree to which BERT borrows from lexical cues in the context of a missing word position. In example (2), when the sentence is preceded by a minimal lexical cue of *delicate*, BERT is able to predict *fragile* in place of the ___ with higher probability as compared to when the sentence is preceded by an unrelated word, *salad*.

(2) a. **delicate.** It was a very ___ tea set.
b. **salad.** It was a very ___ tea set.

# 3 Ontological Semantic Technology

Unlike BERT whose knowledge is based on a corpus, albeit very large, Ontological Semantics is based on human knowledge and the ontology is hand-crafted. Raskin et al. [11] argued that the relative time of acquisition is acceptable for a semantic system and recent views on deep learning [7] agree that a knowledge-based approach could improve deep learning systems. With this in mind, we outline the differences

[1] is-a relationship, for example: *a dog* ***is a*** *mammal*

in results between a very large scale DL architecture and a very small knowledge-based one.

Ontological Semantic Technology [5, 11, 15] is a meaning-based Natural Language Understanding system that consists of several repositories of world and linguistic knowledge. The main static resources consist of: a language independent ontology — a graph whose nodes are concepts and edges are the various relations between them; a lexicon per supported language, that defines word senses of a language by anchoring them with an appropriate concept or property in the ontology. OST processing is event-driven, usually selected from the main verb in the sentence. Once the events sense is disambiguated, a Text Meaning Representation (TMR) is produced, and stored in the Information Repository. This repository is used in processing of further text, depending on the application.

### 3.1 On OST's Fuzzy Nature

OST is fuzzy in nature [16, 12] as most of the processing is driven by so-called facets that represent various membership degrees of a particular event, as described by information in the sentence. The memberships themselves are derived from a location of a concept, recovered from a sentence, in an ontological hierarchy, based on the defined facet and filler combination. While the explicit hierarchical nature is easy to navigate for fillers of OST facets [16] it is worth mentioning that the same procedure can be applied to concepts that can be virtually formed with the help of ontological properties. Such construction of virtual nodes for a crisp ontology was explained in [15]. A crisp ontology, however, always has a membership degree of 1 for every acceptable concept, thus it is worth to address the fuzzy virtual ontology here.

When an ontological event is defined, its semantic roles are filled with concepts, defined in the ontology. For each property, each (facet, filler) pair is a pointer to a concept and its descendants with a membership degree of a pointed concept defined by a facet. OST has four facets: `default,` `sem,` `relaxable-to,` and `not`. `default` has the largest membership degree, 1; `sem` has a smaller membership degree, `relaxable-to` approaches 0, while not membership is exactly 0. The ontological hierarchy, shown in Fig. 1 shows a hierarchy of concepts that can be used in a given event E. The grey concepts will be used explicitly in a definition of a property in the ontology, such as $P(\text{default}(x))(\text{sem}(a))(\text{relaxable-to}(p))$. The membership degrees of other concepts, indicated as single circles, will be calculated according to the formula showed in [18].

Concepts indicated with double circles are virtual – they are not defined by a knowledge engineer but rather taken from a lexical knowledge of language. These are calculated per language and do not have to be stored. Their membership degree is calculated as if they were explicitly defined. In other words, if a knowledge engineer were to place node $z$ into the ontology, the calculation of its membership in an event

$E$ should not change. This gives us flexibility when working with several languages at a time.

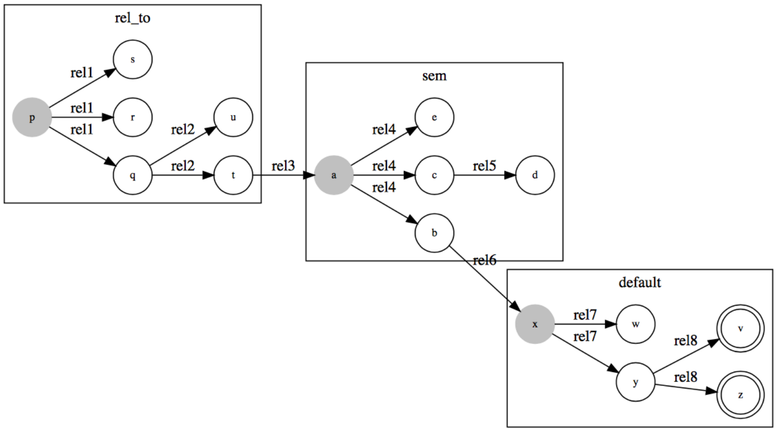

**Fig. 1** Hierarchy of concepts (nodes) with properties (edges) and facets (boxes) and virtual nodes (double circles).

For example, consider a concept WASH. Since any physical object of an appropriate size can be washed, its `sem` facet for a property THEME is likely to be PHYSICAL-OBJECT. However, we may see a sense in some lexicon that restricts a verb anchored in WASH by adding a property to WASH, such as INSTRUMENT with a filler LAUNDRY-DETERGENT. Now suppose an INSTRUMENT of WASH is SOAP, defined as `sem`. SOAP, however, can have different children, such as HAND-SOAP, SHAMPOO, etc. When a lexicon sense of WASHING-WITH-LAUNDRY-DETERGENT is found, its filler, LAUNDRY-DETERGENT, would be placed as a virtual node of a hierarchy, as demonstrated in Fig. 1. This node, defined as INSTRUMENT-OF(WASHING-WITH-LAUNDRY-DETERGENT) will be used whenever appropriate in a calculation of sentence acceptability.

## 4 Masked Word Prediction as Guessing of an Unknown Word's Meaning

Masked word prediction forms the basis of how BERT learns word representations, which are further used in high-level NLP tasks to produce substantial improvements in terms of performance (as reported). Our goal in this work is to analyze BERT's word prediction in context by viewing it from the lens of OST's fuzzy inference capabilities. There can be several ways to infer what will appear in the place of the masked token. From the distributional semantics point of view, words that appear in the same context tend to have similar meanings [4, 3]. An example borrowed from Jurafsky and Martin [6] is presented in (3).

(3) a. *Ongchoi* is delicious sauteed with garlic.
b. *Ongchoi* is superb over rice.
c. *spinach* sauteed with garlic over rice
d. *chard* stems and leaves are delicious
e. *collard greens* and other salty leafy greens.

Since the unknown word *ongchoi* occurs in similar contexts as *spinach, chard,* and *collard greens,* it can be inferred that it is a green leafy vegetable, similar to those mentioned before. While distributional semantics presents a case for statistical approaches, Taylor, Raskin and Hempelmann [17, 19] present a computational semantic approach using OST. Like [1], they formulate the process of acquiring the meaning of an unknown word as a cloze task and produce TMRs by analyzing exemplar contexts consisting of the unknown word. Here, the functional details in the unknown word's context (usually a sentence) determine the basis of understanding the meaning of the unknown word. The example they analyze is a sentence with the verb *rethink* (4a), and the task is to understand the meaning of its direct-object, the new curtains which is replaced with a *zzz* in (4b) to indicate that it is unknown.

(4) a. She decided she would rethink *the new curtains* before buying them for the whole house.
b. She decided she would rethink *zzz* before buying them for the whole house.
c. She decided she would rethink the new ____ before buying them for the whole house.

Based on the TMR representation presented in their paper (shown below), the word *zzz* references the concept that must satisfy certain constraints: (i) it is something that can be rethought, (ii) it carries the semantic role - theme of BUY (iii) it is located in a HOUSE. Note that the paper determines the concept the unknown word evokes as opposed to the word itself. The word could be anything that satisfies those constraints: any kind of furniture - *chair, table, desk, sofa, etc.* or a decorative item such as a *painting*.

```
(DECIDE
    (AGENT(HUMAN(GENDER(FEMALE)))
        (THEME(CONSIDER-INFO(ITERATION(MULTIPLE))
            (AGENT(HUMAN(GENDER(FEMALE)
            (THEME(???))
            (BEFORE(BUY
                (THEME(???(HAS-LOCALE(HOUSE))))))
)))
```

In BERT's case, this instance would be formulated as (4c), where only the word *curtains* has been masked due to BERT's limited capability to only decode one token.

Nevertheless, the example still holds as the only change in the input is an addition of the adjective *new* to describe the object. A selective list of BERT's predictions for (4c) is shown in Table 1. We see BERT assigns high probability to items that can be bought for a house: *paintings, furniture, decorations, etc.*, and even the original masked word, *curtains*. Interestingly, the highest probability is assigned to *clothes*, which is anomalous but could be considered valid if *house* is metonymically referring to the people living in the house, i.e., she is buying clothes for all of them. Whether these predictions are due to purely statistical patterns or something close to true language understanding remains an open research endeavor. We posit that a system that truly understands natural language should assign approximately equal scores to objects that are semantically and syntactically plausible in the sentence. Such a phenomenon is manifested in OST's interpretation of sentences, where the meaning resolution is performed in a structured manner, using TMRs. At the same time, acquiring concepts for the ontology in an accurate manner presents a few challenges, such as an extensive training by a master ontologist.

**Table 1** Selective list of word probabilities for (4c) as estimated by BERT-base

| Rank | Token | Probability | Rank | Token | Probability |
|---|---|---|---|---|---|
| 1 | clothes | 0.1630 | 21 | design | 0.0067 |
| 2 | designs | 0.1320 | 22 | curtains | 0.0063 |
| 13 | paintings | 0.0131 | 23 | gifts | 0.0060 |
| 16 | furniture | 0.0111 | 24 | wardrobe | 0.0057 |
| 17 | pictures | 0.0101 | 25 | products | 0.0049 |
| 18 | books | 0.0096 | 26 | toys | 0.0047 |
| 19 | decorations | 0.0078 | 28 | photos | 0.0041 |
| 20 | arrangements | 0.0070 | 30 | decor | 0.0040 |

## 5 Deconstructing BERT's output using OST and fuzzy inference

In this section, we interpret BERT's output for an example cloze sentence using OST's fuzzy inference mechanism. We first describe our procedure, and then present the interpretation of our example sentence.

**Procedure** Owing to the fact that OST is event-driven, we represent the sentence along the event that affects the missing word, E. The event is represented as a minimal-script where its various case-roles are listed based on the given sentence, as follows:

```
E
    AGENT:
    THEME:
    INSTRUMENT:
    ...
```

Assuming we do not possess a priori knowledge regarding the sense of the event, and so we decompose E into its possible senses $\{$E-v$_1$, E-v$_2$, ..., E-v$_n\}$. For each sense of the concept of E, we compute the fuzzy membership values of the concepts that can occupy the missing position, provided by BERT. This is denoted by $\mu_R(\text{E-v}_i, c)$

where $\mu_R$ is the membership of concept $c$ that participates in the relation $R$ for the i$^{\text{th}}$ sense of the event, E-v$_i$. For the sake of simplicity, we only consider the top-5 words predicted by BERT. Finally, in the same vein as [17], we compute the syntactic and semantic acceptability of the sentence (sent) formed by choosing each of the concept denoted by BERT's prediction as follows:

$$\mu_{\text{syntax}} = \min_{\text{phr} \in \text{sent}} \max_{x,y \in \text{phr}} [\mu_{\text{phr}}(x, y)],$$

$$\mu_{\text{semantics}} = \min_{R \in \text{sent}} \max_{x,y \in R} [\mu_R(x, y)],$$

$$\mu_{\text{acceptability}} = \min[\mu_{\text{syntax}}, \mu_{\text{semantics}}],$$

where $\mu_{\text{acceptability}}$ denotes the overall acceptability membership value, and $\mu_{\text{syntax}}$ and $\mu_{\text{semantics}}$ denote individual membership values for the sentence for its syntax and semantics, respectively. For a detailed analysis of how these values are obtained, please refer to Taylor et al. [17]. We do not choose any final sentence using the acceptability scores, instead, the list of acceptability memberships provide us with relative scores for the concepts evoked by BERT's predictions and help us decipher the extent to which each concept fits into the contextual constraints of the sentence.

**Interpretation Example** Let's consider the following example:

(5) She quickly got dressed and brushed her ___.

**Table 2** Top-5 predicted words for (5) as estimated by BERT-base.

| Rank | Token | Probability |
|---|---|---|
| 1 | teeth | 0.8915 |
| 2 | hair | 0.1073 |
| 3 | face | 0.0002 |
| 4 | ponytail | 0.0002 |
| 5 | dress | 0.0001 |

In its predictions, BERT attributes words that denote concepts that can be the theme of BRUSH (assuming the act-of-cleaning sense of the concept). It predicts *teeth* with the highest probability, alluding to the possibility that a similar sentence describing a person's morning routine has been observed during BERT's training procedure. Following *teeth* are *hair* (the word originally present in the sentence)*, face, ponytail,* and *dress*. Assigning a considerably higher probability to *teeth* as opposed to *hair* can be attributed to BERT's statistical bias which is determined by the corpus it was trained on. While the sentence has two events (DRESS and BRUSH), we will only work with the one that is most concerned with the missing word – BRUSH. This event can be represented as the following minimal-script:

```
BRUSH
    AGENT: HUMAN
        GENDER: FEMALE
    THEME: ___
    INSTRUMENT: NONE
```

Further, consider the following senses of BRUSH (as a verb):

1. Act of cleaning [*brush your teeth*]
2. Rub with brush [*I brushed my clothes*]
3. Remove with brush [*brush dirt off the jacket*]
4. Touch something lightly [*her cheeks brushed against the wind*]
5. ...

Only the first two senses are applicable for the words shown in Table 2. In this analysis, we will interpret the first sense of the event, BRUSH-V1, since the same procedure can be applied to interpret any other sense of the event.

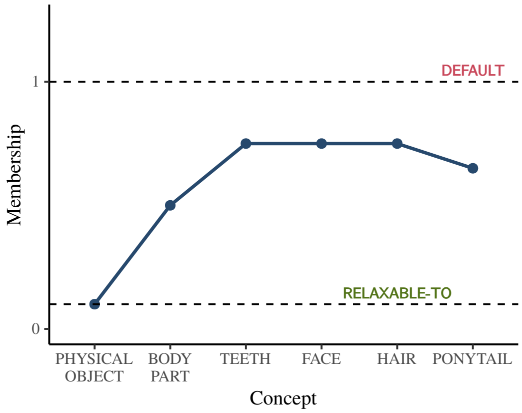

**Fig. 2** Membership values of the various concepts that could be theme of BRUSH-V1. Note that none of the concepts are a `default` but have high membership when the instrument of BRUSH-V1 is not present. Descendants of all such concepts (such as PONYTAIL, which is a child of HAIR) have slightly lower membership. The concept BODY-PART is added to indicate relative position, close to TEETH, etc. and distant from PHYSICAL-OBJECT.

Considering BRUSH-V1, we have four concepts that can have the property, THEME-OF BRUSH: TEETH, HAIR, FACE, PONYTAIL. Notice that PONYTAIL is a descendent of HAIR and its membership for THEME-OF BRUSH-V1 would be slightly lower than that of HAIR. Since the instrument of BRUSH-V1 is missing here, TEETH, HAIR, and FACE have the same membership value as shown in Fig. 2. While all these concepts have high-membership, none of them can be a `default`. We also consider the `relaxable-to` facet here as we want to restrict non-physical objects from being counted as THEME of BRUSH-V1. The memberships of the concepts denoted by these words would be ordered as follows:

$$\mu_{\text{theme}}(\text{TEETH}) = \mu_{\text{theme}}(\text{FACE}) = \mu_{\text{theme}}(\text{HAIR}) > \mu_{\text{theme}}(\text{PONYTAIL})$$

However, consider the following sentences:

(6) a. She quickly got dressed and brushed her ___ with a comb.
b. She quickly got dressed and brushed her ___ with a toothbrush.

These examples further constrain the membership values for the concept that satisfies the THEME-OF BRUSH-V1 relation by adding an INSTRUMENT-OF relation. To account for the instrument, we traverse down the virtual hierarchy of BRUSH-V1, a subset of which is shown in Fig. 3. As mentioned in section 3.1, two new virtual

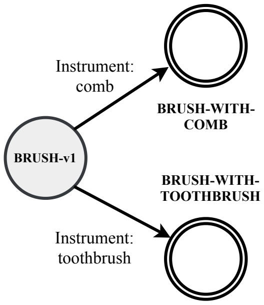

**Fig. 3** Virtual nodes created in the hierarchy of BRUSH-v1 when it is endowed with an INSTRUMENT-OF relation. These nodes alter the membership values for concepts that can satisfy the relation, THEME-OF for BRUSH-v1, and assign new scores to them depending on the value of the INSTRUMENT-OF BRUSH-v1.

nodes are created when BRUSH-v1 is endowed with an instrument (either COMB or TOOTHBRUSH). With this new knowledge, the membership value for certain concepts is elevated to the `default` facet. At the same time, the membership of all other concepts that can no longer be the theme of BRUSH-WITH-[INSTRUMENT] is lowered. For descendants of the `default`, the membership for THEME-OF BRUSH-WITH-[INSTRUMENT] would increase relative to their membership for THEME-OF BRUSH. This can be summarized by the following for concepts HAIR, TEETH, and PONYTAIL:

$$\mu_{\text{theme}}(\text{BRUSH-WITH-TOOTHBRUSH}, \text{TEETH}) = 1$$
$$\mu_{\text{theme}}(\text{BRUSH-WITH-COMB}, \text{HAIR}) = 1$$
$$\mu_{\text{theme}}(\text{BRUSH-WITH-COMB}, \text{PONYTAIL}) > \mu_{\text{theme}}(\text{BRUSH}, \text{PONYTAIL})$$

BERT's outputs for the sentences in example (6) is shown in Table 3.

**Table 3** BERT-base probabilities for words predicted in (5) but with (6a) and (6b) as inputs

| BRUSH-WITH-COMB (6a) | | | BRUSH-WITH-TOOTHBRUSH (6b) | | |
|---|---|---|---|---|---|
| **Rank** | **Token** | **Probability** | **Rank** | **Token** | **Probability** |
| 1 | hair | 0.8704 | 1 | teeth | 0.9922 |
| 2 | teeth | 0.1059 | 2 | hair | 0.0052 |
| 3 | face | 0.0210 | 3 | face | 0.0019 |
| 12 | ponytail | $< 0.0001$ | 31 | ponytail | $< 0.0001$ |
| 27 | dress | $< 0.0001$ | 98 | dress | $< 0.0001$ |

We see the probabilities estimated by BERT for the event of BRUSH with and without an INSTRUMENT-OF relation are considerably different. BERT assigns the highest probability to each event's `defaults`, denoted by words *hair* and *teeth* respectively. Judging from the top-predicted word in both events, BERT follows the same pattern as our interpretation of assigning higher score to the `default`. We compare the rest of the top-5 predicted words by following the discussion about THEME-OF BRUSH-WITH-[INSTRUMENT]. For the event of BRUSH-WITH-COMB, the following acceptability measures emerge:

$$\mu_{\text{syntax}}(\text{THEME-OF BRUSH-WITH-COMB}, \{\text{top-5-predictions}\}) = 1 \tag{1}$$

$$\mu_{\text{semantics}}(\text{THEME-OF BRUSH-WITH-COMB}, \text{HAIR}) = 1 \tag{2}$$

$$\mu_{\text{semantics}}(\text{THEME-OF BRUSH-WITH-COMB}, \{\text{TEETH, FACE, DRESS}\}) < \\ \mu_{\text{semantics}}(\text{BRUSH-WITH-COMB}, descendent\text{-}of(\text{HAIR})) \tag{3}$$

Every prediction in table 3 is considered syntactically acceptable for both examples since all the predicted words are nouns in the example sentences. The difference lies in the overall acceptability scores with respect to semantics. The `default`, HAIR will have the highest score, 1. Following HAIR should be its descendants, in this case, PONYTAIL. Every other word and the concepts evoked by them should be scored considerably lower since they would be considered semantically anomalous as THEME-OF BRUSH-v1. Hence,

$$\mu_{\text{acceptability}}(\text{HAIR}) > \mu_{\text{acceptability}}(\text{PONYTAIL}) = \\ \mu_{\text{acceptability}}(descendent\text{-}of(\text{HAIR})) > \mu_{\text{acceptability}}(\{\text{all-other-predictions}\})$$

From the predictions for BRUSH-WITH-COMB, we find BERT's outputs to deviate from our acceptability measures. PONYTAIL is scored considerably lower than words denoting concepts with much lower membership values for THEME-OF BRUSH-WITH-[INSTRUMENT]. Why such a situation arises is subject to further quantitative evaluation, including the investigation of BERT's internal mechanisms. Our takeaway here is that the statistical patterns learnt by BERT could prevent it from demarcating concepts that satisfy a semantic relation, as exemplified in the above evaluation.

## 6 Conclusion

In this paper, we have adopted an Ontological Semantics [10] approach to analyze and interpret the output of a black-box neural network model, BERT [1]. We evaluate BERT's output for the task of predicting the word occupying a missing position (indicated by a ___) in a sentence, which is in the same vein as "guessing the meaning of an unknown word in context," a task proposed earlier [19]. To this end, we utilize a realization of Ontological Semantics, the OST system and its inherent ability to make fuzzy inferences about the concept occupying the missing position in the sentence. This provides a mechanism for us to quantify the degree to which a concept is evoked by the unknown word in its context based on the functional relationships of its surrounding items [17, 19]. Using an exemplar sentence, we discussed the event of BRUSH [act of cleaning] and modified it by adding an INSTRUMENT-OF relation. While we found BERT to change its top-predicted word when the instrument of the event changed, it was unable to show structural (semantics-wise) phenomena. This was evidenced by BERT scoring a descendent of HAIR: PONYTAIL lower than a nonsensical concept (in the given instance) – TEETH, indicating a hindrance displayed by its training process with respect to learning true semantic understanding.

## References

1. 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 (2019)
2. Ettinger, A.: What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. Transactions of the Association for Computational Linguistics **8**, 34–48 (2020)
3. Firth, J.R.: A synopsis of linguistic theory 1930-1955. Studies in linguistic analysis (1957)
4. Harris, Z.S.: Distributional structure. Word **10**(2-3), 146–162 (1954)
5. Hempelmann, C.F., Taylor, J.M., Raskin, V.: Application-guided ontological engineering. In: ICAI 2010: Proceedings of the 2010 International Conference on Artificial Intelligence (Las Vegas NV, July 12-15, 2010), pp. 843–849 (2010)
6. Jurafsky, D.: Speech & language processing, 3rd Edition (2020)
7. Launchbury, John: A DARPA Perpective on Artificial Intelligence (2019). https://www.darpa.mil/attachments/AIFull.pdf
8. Linzen, T., Dupoux, E., Goldberg, Y.: Assessing the ability of LSTMs to learn syntax-sensitive dependencies. Transactions of the Association for Computational Linguistics **4**, 521–535 (2016)
9. Misra, K., Ettinger, A., Rayz, J.T.: Exploring Lexical Relations in BERT using Semantic Priming. In: Proceedings of the Annual Meeting of the Cognitive Science Society (2020)
10. Nirenburg, S., Raskin, V.: Ontological semantics. MIT Press (2004)
11. Raskin, V., Hempelmann, C.F., Taylor, J.M.: Guessing vs. knowing: The two approaches to semantics in natural language processing. In: Annual International Conference Dialogue 2010, pp. 642–650 (2010)
12. Raskin, V., Taylor, J.M.: The (not so) unbearable fuzziness of natural language: The ontological semantic way of computing with words. In: 2009 Annual Conference of the North American Fuzzy Information Processing Society, pp. 1–6. IEEE (2009)
13. Reif, E., Yuan, A., Wattenberg, M., Viegas, F.B., Coenen, A., Pearce, A., Kim, B.: Visualizing and Measuring the Geometry of BERT. In: Advances in Neural Information Processing Systems, pp. 8592–8600 (2019)
14. van Schijndel, M., Linzen, T.: Modeling garden path effects without explicit hierarchical syntax. In: Proceedings of the Annual Meeting of the Cognitive Science Society (2018)
15. Taylor, J.M., Hempelmann, C.F., Raskin, V.: On an automatic acquisition toolbox for ontologies and lexicons in ontological semantics. In: ICAI 2010: Proceedings of the 2010 International Conference on Artificial Intelligence (Las Vegas NV, July 12-15, 2010), pp. 863–869 (2010)
16. Taylor, J.M., Raskin, V.: Fuzzy ontology for natural language. In: 2010 Annual Meeting of the North American Fuzzy Information Processing Society, pp. 1–6. IEEE (2010)
17. Taylor, J.M., Raskin, V.: Understanding the unknown: Unattested input processing in natural language. In: 2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011), pp. 94–101. IEEE (2011)
18. Taylor, J.M., Raskin, V.: Conceptual defaults in fuzzy ontology. In: 2016 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS), pp. 1–6. IEEE (2016)
19. Taylor, J.M., Raskin, V., Hempelmann, C.F.: Towards computational guessing of unknown word meanings: The Ontological Semantic approach. In: Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 33 (2011)
20. Taylor, W.L.: Cloze procedure: A new tool for measuring readability. Journalism quarterly **30**(4), 415–433 (1953)
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)