EXPLORING LEXICAL SENSITIVITIES IN WORD PREDICTION MODELS: A CASE STUDY ON BERT

by

Kanishka Misra

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the Degree of

Master of Science

Department of Computer and Information Technology

West Lafayette, Indiana

December 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Julia Taylor Rayz, Chair

Department of Computer and Information Technology

Dr. John A. Springer

Department of Computer and Information Technology

Dr. Victor Raskin

Department of English and Linguistics Program

Approved by:

Dr. John A. Springer

Chair of the Graduate Education Committee

Dedicated to my parents:

Kruti and Laxminarayan Misra

and both pairs of grandparents:

Hasumati and Bhupendra Shah

Pramila and Dibyanarayan Misra

ACKNOWLEDGMENTS

I'd like to give heartfelt thanks to my advisor Julia Rayz, for her invaluable advice and support for the past four years of my research career. This work was exceptionally facilitated by my collaboration with Prof. Allyson Ettinger, who has been—and continues to be—a wonderful mentor and an excellent source of key discussion and valuable feedback. I would also like to express my sincere thanks to committee members – Profs. Victor Raskin and John Springer, for agreeing to serve on my thesis committee. Special thanks to Prof. Dawn Laux for giving me the opportunity to discover myself as a teacher, and funding me for three semesters. I am grateful to the CIT graduate program assistant Kari Ludwig and the CIT department head Prof. Tom Hacker for being accommodating with last minute administrative work.

A bulk of the work presented in this thesis was performed during the COVID-19 pandemic, and I would not have been able to finish it if it weren't for the fine company of my dear friends with whom I shared a living space at some point during this year – Ananya Sheth, Parul Verma, Akancha Pandey, Janav Udani, Krishna Pillutla, and Kaushal Jain. I thank them for their valuable career and life advice, (not so) funny jokes, and most importantly, delicious food as I worked on this thesis. I'd also like to thank my exceptional lab mates at the AKRaNLU lab, especially Tatiana Ringenberg and Hemanth Devarapalli, with whom I collaborated extensively prior to starting my thesis research. I've been lucky to be friends with some of the best people. My high-school friends, who are still in touch — Vasudev Vohra, Gautam Wadhwa, Anshul Mathur, Nishank Saini, Vaishnavi Singh, Sarah Hussain, and Arjun Roy. Friends I made during my undergraduate degree, all of who are pursuing excellent careers — Mikaylah Gross, Lakshmi Achutanand, Aditya Gupta, Ridhi Adyanthaya, Vishnu Tadimeti, Joey Zedan, Ashley Jerue, and Sanjna Vinze. I thank you all for your company over these years.

Last but not the least, this thesis has been a testament to the support, love, and blessings I have received from my parents, Kruti and Laxminarayan Misra. Thank you for all that you have done for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Estimating word probabilities in context is a fundamental mechanism underlying the training of neural network-based language processing models. Models pre-trained using this mechanism tend to learn task independent representations that exhibit a variety of semantic regularities that are desirable for language processing. While prediction based tasks have become an important component for these models, much is unknown about what kinds of information the models draw from context to inform word probabilities. The present work aims to advance the understanding of word prediction models by integrating perspectives from the psycholinguistic phenomenon of semantic priming, and presents a case study analyzing the lexical properties of the pretrained BERT model. Using stimuli that cause priming in humans, this thesis relates BERT's sensitivity towards lexical cues with predictive contextual constraints and finer-grained lexical relations. To augment the empirical methodology utilized to behaviorally analyze BERT, this thesis draws on the knowledge-rich paradigm of Ontological Semantics and fuzzy-inferences supported by its practical realization, the Ontological Semantics Technology, to qualitatively relate BERT's predictive mechanisms to meaning interpretation in context. The findings establish the importance of considering predictive constraint effects of context in studies that behaviorally analyze language processing models, and highlight possible parallels with human processing.

CHAPTER 1. INTRODUCTION

The field of natural language processing (NLP) has seen significant progress in the past decade. This progress is evident not only in highly targeted applications such as machine translation for high-resource languages, but also in applications that evaluate general purpose language understanding (Wang et al., 2019, 2018). Computational advances have been led in-part by the emergence of highly parameterized neural networks, which are trained on large amounts of text using “pretraining” (Howard & Ruder, 2018), a process that typically involves estimating word probabilities in context. Improvements made to neural network-based models, either by altering underlying architectures or by training on larger texts, often result in increased complexities, thereby coming at the cost of our understanding of the system. For instance, GPT-3 (Brown et al., 2020), the latest in the line of large neural network models, was trained on $\approx$ 570GB of heterogeneous text and contains a total of 175 billion parameters. This trade-off between complexity and understanding in neural network-based language processing points to the need for principled paradigms that analyze and interpret such models — what is it that they learn? This thesis presents a case study on one such pretrained model, BERT (Devlin et al., 2019), by developing methodology inspired from the psycholinguistic phenomenon of semantic priming (McNamara, 2005; Meyer & Schvaneveldt, 1971).

1.1 Pretrained Language Models and their Analyses

A popular method of pretraining language processing neural networks is by using the language model (LM) objective — estimation of word probabilities in context. A neural language model treats sentences as sequences of words, and computes the following measure:

$$\prod_{t=0}^{T} P(w_t \mid h_t, \Theta_t), \tag{1.1}$$

where w_t is the t^{th} word in the sequence, and h_t is the hidden state of the model which represents the sequence context at time-step t, and Θ_t is the collection of learnable parameters of the model. The principle behind such a process is that by constantly updating Θ in optimizing the estimation of word probabilities in context, the models produce representations for tokens that can then be "fine-tuned" for specific tasks. These representations are often called word embeddings, or word vectors. Pretraining using the LM objective is advantageous since it affords supervision to the models without any labelled data — the input and the labels are both provided by the same source, the text. That is, pretraining facilitates *self-supervised* learning.

As the paradigm in state-of-the-art NLP modeling shifts towards using LM-based pretraining, it has become increasingly relevant to fundamentally understand the kinds of linguistic competencies that word prediction in context confers upon such models (and what it does not). This paradigm shift has given birth to an entire research program (Alishahi, Chrupała, & Linzen, 2019) within NLP that aims to decipher how and whether linguistic knowledge is organized in such models, and what makes them perform impressively well on a variety of tasks. In light of the inherent processes that comprise of training LMs, this thesis divides the general class of LM analyses methods into two broad distinctions: the first class of methods evaluate LMs by providing stimuli and testing on word prediction in context, these are termed as "behavioral" analyses in this thesis. For instance, what is the probability of the word, *are* in example (1.1) and is it greater than the probability of the word, *is*:

(1.1) The keys to the cabinet ___ .

The second class of methods aim to recover linguistic properties from pretrained LMs by using supervised classifiers (known as *probes*) that operate over word representations formed in these models. In this thesis, these methods are collectively referred under "diagnostic classification," or *probing*. Consider the example (1.2), where 1.2a is the surface form of the sentence, and 1.2b is the sequence of part of speech (POS) tags corresponding to the sentence's words. If a probing classifier that is trained to recover POS tags of a given sequence successfully predicts the sequence (1.2b) when provided the sentence (1.2a), then it indicates that the pretrained LM encodes POS information[1].

(1.2) a. The dog chased the cat.

b. DET NN VBD DET NN

[1] the debate about whether the information is encoded completely in the LM or whether the probe learns this information is further explored in (Hewitt & Liang, 2019) and is out of scope for this thesis.

In assessing targeted linguistic phenomena, the class of behavioral analysis methods has three *prima facie* advantages: (1) it allows the testing of LMs in their natural environment, without involving any task specific fine-tuning; (2) it helps ask targeted questions by directly borrowing carefully constructed items in related research fields such as cognitive science, psycholinguistics, or theoretical linguistics; and (3) it provides empirical evidence whether LM pretraining is well-suited for learning the specific linguistic phenomena that is being analyzed — for instance, Ettinger (2020) shows BERT's strong insensitivity to negated sentences, suggesting that LM based pretraining is not cut out for learning message level inferences such as negation.

Although the literature directly focused on the analysis of neural networks has advanced the understanding of the linguistic capacities that pretraining confers upon the models, much of it is dominated by the testing of syntactic phenomena. Indeed, syntactic analyses provide useful methodological frameworks to survey the linguistic structures learned by LMs, which represent sentences as sequences, and do not induce any hierarchical knowledge (for example, the tree-structure representation of English grammar). However, most natural language understanding tasks require competence in the sub-field of linguistics that is responsible for the meaning of everything in language that has meaning (Raskin & Weiser, 1987), semantics. For instance, in example (1.3), if a model trained to perform the task of Natural Language Inference (NLI) (Bowman, Angeli, Potts, & Manning, 2015) or Recognizing Textual Entailment (RTE) (Dagan, Roth, Sammons, & Zanzotto, 2013) predicts that (1.3a) logically entails (1.3b), then the model should, in theory, have learnt the lexical association between bird and robin, i.e., a robin is a type of bird.

(1.3) a. A *robin* is flying.
b. A *bird* is flying.

To specifically interpret whether or not linguistic meaning manifests in pretrained LMs requires evaluation methodologies that ask and test a wide variety of questions pertaining to semantic capacities. This thesis contributes to the development of such methodologies by providing an account of how lexical associations (such as the one described in example (1.3)) are leveraged by one such pretrained model, BERT (Devlin et al., 2019) to inform word probabilities in context. For example, if a word like *airplane* is prepended to (1.4a), to what extent does this increase the BERT's probability for the word *pilot* in the blank position in (1.4b)?

(1.4) a. I want to become a ___ .

b. *airplane.* I want to become a ___ .

This question is particularly relevant because human brains show a robust phenomenon of *semantic priming* (McNamara, 2005), in which the presence of a word such as "airplane" will give rise to faster reactions to a related word like "pilot", than when unrelated word like "table" is present. Empirical observations about priming in humans reveal the organization of lexical knowledge in human brains, in the form of strength of association between lexical items, as reflected by the magnitude of the speed-up of their reaction times. This thesis explores whether the same lexical relations that show priming in humans will also be utilized by BERT to influence word predictions in context. By analogy, doing so will reveal potential insights about BERT's organization of semantic content and how it is used when it performs its primary task (on which it was trained on) of predicting words in context. Priming behavior in BERT, then, is defined as an increase in the model's expectation for a target word (or a lack thereof) in a given sentence context in the presence of a semantically related word as compared to an unrelated one. In casting BERT's priming as a test of lexical sensitivities, it is important to take into account the predictive bias of the context that exists independent of the prime words. This predictive bias manifests within human psycholinguistic experiments in the form of the "constraint" of a sentence (Federmeier & Kutas, 1999; Schwanenflugel & LaCount, 1988) and corresponds to how predictable a sentence with a missing word is. For instance, the missing word in the sentence, *"John kept his gym clothes in the ___ "* is easier to predict than in *"In the valley, there were three small ___ "* (Schwanenflugel, 1991). Taking this into account, this thesis focuses on the lexical sensitivities of BERT's word prediction in context capabilities based on how it is modulated by the predictive constraints of the context. Augmenting the fundamentally empirical endeavor, this thesis borrows from the school of Ontological Semantics (Nirenburg & Raskin, 2004), a meaning-first approach to knowledge representation and reasoning, and presents a qualitative account of the semantic constraints imposed by the sentence context on the missing word. This qualitative analysis is primarily conducted by using the fuzzy-inferences (Zadeh, 1965) facilitated by the Ontological Semantic Technology (OST) (Hempelmann, Taylor, & Raskin, 2010; Raskin, Hempelmann, & Taylor, 2010; J. M. Taylor, Hempelmann, & Raskin, 2010; J. M. Taylor & Raskin, 2010, 2016), the latest practical realization of Ontological Semantics.

In summary, this thesis introduces a methodology for fine-grained exploration of lexical cue sensitivity in language models, grounded in lexical relation phenomena observed in humans during semantic priming. It further considers contextual constraints of cloze-contexts (W. L. Taylor, 1953) by deconstructing the process of predicting a missing word using fuzzy inferences facilitated by OST, a knowledge representation and reasoning system that is purely meaning-based.

1.2 Research Questions

This thesis addresses the following research questions:

1. To what extent does BERT show a sensitivity to lexical cues that cause priming in Humans?
2. Does BERT show similar patterns of lexical sensitivities across different lexical relations between words?
3. How does BERT's lexical sensitivity in context get affected by constraints imposed by the input context?

1.3 Assumptions

The assumptions of this thesis are primarily related to the datasets and the model investigated in the methodology chapter (Chapter 3). Briefly, this thesis makes the following assumptions:

- Priming measures such as response times of participants provide a behavioral account of how humans represent lexical relations.
- The Semantic Priming Project (Hutchison, Balota, Neely, Cortese, Cohen-Shikora, et al., 2013), a dataset described in Chapter 3, comprehensively accounts for lexical relations in the English language.
- The BERT model (Devlin et al., 2019) is representative of the family of models that utilize the masked language modelling procedure.

1.4 Scope and Limitations

The work presented in subsequent chapters of this thesis investigates how a specific language processing model, BERT, adapts from lexical cues in context to inform its word probabilities. The analyses rely heavily on the use of lexical stimuli from semantic priming that elicit priming in humans to inform how lexical relations are utilized during word prediction in context. However, this study only analogously compares the priming behavior of BERT to that of humans. Since the structure of the experiments is formulated differently than standard priming setups, the methodology presented in this thesis cannot simulate or model human priming behavior, and does not offer a cognitive account of semantic priming.

Pretraining by the LM-based objectives has only recently become standard practice in the field of NLP (Howard & Ruder, 2018). Leading this charge was the BERT model (Devlin et al., 2019), which is a bidirectional transformer model, optimized in part to use context information to predict masked words. Due to the underlying architecture of the BERT model, the specific text stimuli presented in this work are unable to address the lexical of interest in unidirectional, incremental models in which the context always appears to the left of the missing token. This thesis also only considers the lexical sensitivities of word prediction in context for BERT trained on English data, which casts doubt about the generalization of the results across different languages. However, the described methodology is flexible enough to be extended to all language models and languages for which enough representative data is available. This will require a careful reconstruction of the stimuli — sampled from a large corpora in the desired language — which places the entire context to the left of the missing token. Finally, the empirical methodology presented in this work is complimented by a qualitative reinterpretation of the BERT's training task, using a meaning-first approach to knowledge representation. However, this account is constrained to only qualitatively providing the information needed to perform word prediction in context, and while it presents a weak quantitative notion using fuzzy-sets, it lacks robust quantitative summaries, mainly due to the absence of standard data-driven metrics in the field of ontological semantics (Nirenburg & Raskin, 2004).

1.5 Thesis Contributions

This thesis investigates pretrained language processing models that are trained by estimating word probabilities in context and sheds light on the lexical dynamics that occur in these models by presenting a case-study on one such model, BERT. By borrowing from the phenomenon of semantic priming, and by considering the nature of sentence contexts in terms of semantic constraints, this thesis builds on a growing precedent of using psycholinguistics-inspired tests which focus on underlying mechanisms and linguistic competence of neural network based models, and how closely they approximate language processing phenomena observed in humans. This thesis also provides a qualitative account of contextual constraints and model behavior using a meaning-first approach to representing knowledge. The resulting original contributions are:

1. A methodology for fine-grained exploration of lexical cue sensitivity in word prediction models, grounded in lexical relation phenomena observed in humans. This analytical framework empirically establishes the importance of considering contextual-semantic constraints that affect analyses that behaviorally probe predictive models. Dissemination – (Misra, Ettinger, & Rayz, 2020a, 2020b):

   - Misra, K., Ettinger A., Rayz, J.T. (2020). Exploring Lexical Relations in BERT using Semantic Priming. In *42nd Annual Virtual Meeting of the Cognitive Science Society.* (Poster Presentation).
   - Misra, K., Ettinger A., Rayz, J.T. (2020). Exploring BERT's Sensitivity to Lexical Cues using Tests from Semantic Priming. *Findings of ACL: EMNLP 2020.* (long paper)

2. A qualitative analysis of the BERT's word prediction in context capability that is cast as "guessing the meaning of an unknown word in context," a task introduced in a series of papers in the school of Ontological Semantics, a meaning-first approach to represent knowledge. This work analyzes BERT's behavior by devising a mechanism that leverages fuzzy inference through Ontological Semantic Technology (OST), and descriptively deconstructs sentence stimuli into event representations which provide an exposition of the various semantic constraints posed by the context. Dissemination:

- Misra, K., & Rayz, J.T. (2020). An Approximate Perspective on Word Prediction in Context: Ontological Semantics meets BERT. In: *2020 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS).* (Regular Paper; forthcoming)

1.6 Organization of this Thesis

The current chapter establishes the primary motivation underlying the work presented in this thesis by situating it in a growing precedent of formulating principled evaluation methods that target specific phenomena captured by neural networks, which are largely black-box in nature.

Chapter 2 constitutes a synthesis of prior work relevant to the justification and motivation of the proposed analysis. It first describes the framework of language modelling, which constitutes the core task of pretrained language processing models. This section covers the broad spectrum of language models starting from count based n-gram models to neural network based predictive language models. The next two sections review the impacts of pretraining and introduce the BERT architecture, the main subject of the experiments. It then summarizes the landscape of analysis methods by distinguishing them into two main types – *probing* and *behavioral* analysis and discusses their findings about language models. Next, it briefly introduces Ontological Semantics and OST along with a concise discussion about fuzzy sets and how fuzziness manifests within OST. Finally, the chapter summarizes the cognitive phenomena of semantic priming and the contextual effects in sentence processing, which form the primary underpinnings of the core set of methodological developments in this thesis.

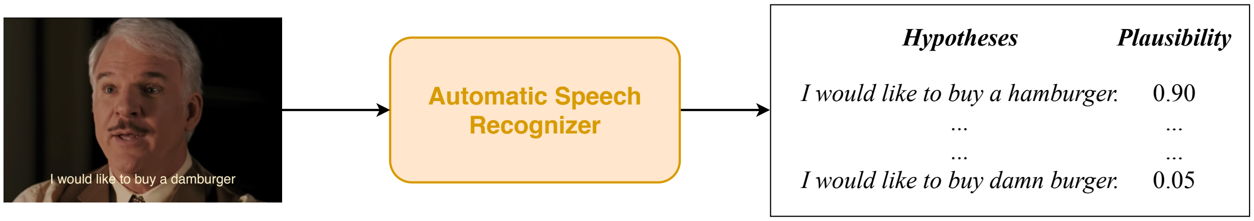
Chapter 3 establishes the methodological contributions of this thesis. It first discusses the nature of the stimuli used in terms of its similarities to stimuli used in cloze tasks (W. L. Taylor, 1953) and the kinds of knowledge one can extract from them. Next, it outlines the analysis of the stimuli from the perspective of Ontological Semantics, which qualitatively describe the semantic constraint imposed within them. This chapter then delves into the core set of methodological considerations for extending semantic priming to BERT, the computation of predictive constraint and its relation to information theory, and the various measures used in subsequent analyses. Finally, it describes the experiments performed to answer the research questions presented in Chapter 1.

Chapter 4 presents the results of the empirical experiments outlined in the Chapter 3, their interpretation in the context of the model, and an investigation into important anomalous patterns. It is concluded by a detailed discussion and implication of the empirical findings. Chapter 5 concludes the thesis and recommends future pathways of research.

CHAPTER 2. BACKGROUND

2.1 Language Modeling

How does one distinguish “good” sentences from bad ones? A native speaker of English can probably tell you that *“The cat sat on the mat.”* is an acceptable English sentence while *“The cat mat on the sat.”* is not. The acceptability of a sentence can be characterized based on whether it is meaningful, grammatical, and able to communicate some sort of message. These ideas are extremely relevant for the field of NLP and artificial intelligence as a whole. For instance, consider an “intelligent” system that has to process Inspector Jacques Clouseau’s dialogue about buying hamburgers (shown in Figure 2.1), due to the noise in the speaker’s verbal input, the system maintains several hypotheses about what he is trying to say — “I would like to buy damn burger” vs “I would like to buy a hamburger.” To successfully decode the speaker’s verbal input, the system will have to rely on a mechanism that ranks these potential hypotheses and selects the most plausible one out of those. The same case could be applied in machine translation, where due to the presence of ambiguity, a sentence can potentially mean multiple closely related (or hilariously different) things. These examples motivate the notion of a mechanism in such intelligent systems that assesses the plausibility of sentences. This section describes one such mechanism that has occupied a central role in the current methodological paradigm within NLP (Eisenstein, 2019; Goldberg, 2017; Jurafsky & Martin, 2020): Language Models.



Hypotheses	*Plausibility*
I would like to buy a hamburger.	0.90
...	...
...	...
I would like to buy damn burger.	0.05

Figure 2.1. A hypothetical scenario where sentence plausibility could play a major role. Screenshot of the dialogue scene taken from `https://www.youtube.com/watch?v=Z6oeAdemFZw`

Language models formalize the intuition of assigning probabilities to sentences in a natural language. Given a sequence of tokens, $s = (w_1, w_2, ..., w_N)^1$, a language model estimates the following probability:

$$P(s) = P(w_1, w_2, ..., w_N). \tag{2.1}$$

The computation of such a probability assumes a discrete set of tokens belonging to the language, often known as the language's vocabulary, $\mathcal{V}$, such that $\forall i,\ w_i \in \mathcal{V}$. For instance, the vocabulary of the English language can be represented as the set: $\mathcal{V}_{\text{ENGLISH}} = \{\textit{aardvark}, ..., \textit{Zyzzyva}\}$. Using the chain rule of probability, Equation 2.1 can be rewritten as:

$$\begin{align}
P(S) &= P(w_1) \times P(w_2 \mid w_1) \times P(w_3 \mid w_1, w_2) \times ... \times P(w_N \mid w_1^{N-1}) \tag{2.2}\\
&= P(w_1) \prod_{i=1}^{N} P(w_i \mid w_1^{i-1}). \tag{2.3}
\end{align}$$

The quality of a language model can be assessed using two broad classes of evaluation methods. The first class of methods are collectively referred to as extrinsic methods, wherein the language model is evaluated as a component of a system for its effectiveness in a higher-level task. For instance, one could measure the change in translation quality of a machine translation system when language model A is replaced by language model B. The second class of evaluation methods are known as intrinsic and, as the name suggests, measure quality of language models in their natural setting, independent of any external application. The dominant approach to assess language models intrinsically is to measure a model's *perplexity* over unseen/new language constructions.

A typical approach to train or build language models is to first select a corpus that is representative of the language for which the model is being constructed—this is known as the training set. The language modeler then defines and estimates the desired probabilities such that given a sentence as an input, the language model produces its estimated probability. If the language model is indeed intrinsically of good quality, it should assign high probabilities to sentences that are valid but aren't observed in the training set. That is, it should generalize well to unseen instances present in the testing set. The perplexity measure on the test set operationalizes the notion of a language

[1] for brevity, let the symbol w_j^k denote $\{w_j, w_{j+1}, ..., w_k\}$. Hence, s can be represented as w_1^N.

model's generalization power. Formally, let S be the testing set, consisting of unseen sentences $\{s_1, s_2, ..., s_m\}$ and containing a total of M words. A robust language model should maximize the overall probability of the test set, $P(S)$:

$$P(S) = \prod_{i=1}^{m} P(s_i), \tag{2.4}$$

The perplexity of a language model LM is then defined as the inverse probability of the test set, normalized by the number of words, M:

$$\begin{aligned} \textit{ppl} &= \left(\prod_{i=1}^{m} P(s_i)\right)^{-\frac{1}{M}} && (2.5)\\ \log_2(\textit{ppl}) &= -\frac{1}{M}\log_2 \prod_{i=1}^{m} P(s_i) && (2.6)\\ &= -\frac{1}{M}\sum_{i=1}^{m} \log_2 P(s_i) && (2.7)\\ \textit{ppl} &= 2^{-\left(\frac{1}{M}\sum_{i=1}^{m} \log_2 P(s_i)\right)} && (2.8) \end{aligned}$$

Therefore, more robust language models produce lower perplexity values—are less "perplexed"—in encountering unseen sentences. It is important to note that the perplexity measure is strongly tied to the corpus that the language modeler/researcher has chosen to be the training and testing set, and so two language models can only be faithfully compared when they are trained on the same corpus (Eisenstein, 2019; Jurafsky & Martin, 2020). Furthermore, if the corpus is not a very good representative collection of sentences in the given language, then it is expected that the language model will not generalize well to test sets.

2.1.1 n-gram Language Models

As represented in equation (2.2), language models estimate the probability of a word in context by conditioning on all words prior to the given word. Conditioning on long sequences requires storing probabilities for every possible sequence of tokens from the vocabulary, making the training of language models a difficult combinatorial problem. n-gram language models alleviate this computational issue by making the *markov assumption*. A k-th order markov assumption assumes that the next word in a sequence/sentence only depends on the k previous words that occur before it (Goldberg, 2017). The task of language modeling under this assumption now requires approximating $P(w_i \mid w_1^{i-1})$ with $P(w_i \mid w_{i-k}^{i-1})$, thus reducing the number of parameters for estimating the probability for any given sequence. This gives rise to n-gram language models, which compute maximum likelihood estimates (MLE) of probabilities derived from corpus counts of sequences.

An n-gram model follows the $(n-1)$-order markov assumption, where the MLE estimate is given as:

$$\hat{P}_{\text{MLE}}(w_i \mid w_{i-(n-1)}^{i-1}) = \frac{n(w_{i-(n-1)}^{i})}{n(w_{i-(n-1)}^{i-1})}. \tag{2.9}$$

At n = 1 we have a unigram model, at n = 2 we have a bigram model, at n = 3 we have a trigram model, at n = 4 we have a "four"-gram, and so on. The unigram model makes an extreme assumption of not conditioning a word on any of its preceding words in a sequence, i.e., the probability of a word in a sequence is independent of any other word prior to it. Unsurprisingly, this results in absurd (non-) sentences such as *"<s> the the the ... "* getting extremely high probabilities, since the word *the* is the most frequent word in most English corpora. This strong independence assumption is weakened in higher order n-gram models, which compute and store more information about word sequences. While n-gram models make it straightforward and easy to model word sequences, they have several shortcomings, summarized below:

- **Failure to model long-range dependencies:** n-gram models only estimate the probabilities of limited sequential histories. This results in a failure to model crucial linguistic phenomena. As an example, a trigram model will fail to show subject-verb-agreement (Linzen, Dupoux, & Goldberg, 2016) in sequences such as *"The keys to the cabinet are...,"* preferring *is* as opposed to *are* as the completion due to the trigram *the cabinet is* being more frequent than *the cabinet are* An obvious solution to this is to increase the value of n to generalize to longer sequences, however this leads to the next problem:

- **Sparsity and Memory based problems:** n-gram models estimate probabilities using counts from a fixed corpus. Therefore, if an n-gram never occurs in the training set, then it is assigned a probability of 0. Given a vocabulary V of size v, there are v^n possible n-grams, and most of them tend to never occur in the corpus, leading to a greater number of 0 counts in the model's parameters. Furthermore, as n increases, the number of possible n-grams increases v-fold, causing storage problems.

- **Failure to show semantic generalization:** n-gram language models are trained by computing the relative frequencies of n-length sequences observed in the training corpus, thus maximizing the probability of the training set. This is in essence of the principle of maximum likelihood estimation. However, this method of training language models does not allow them to generalize across plausible sequences. For instance, i.e. if the bigrams *blue cup* and *red cup* are observed in the training set, and *orange cup* is not, then the model assigns 0 probability to it regardless of it being a semantically plausible occurrence, thus limiting the model's ability to learn semantic similarities beyond those represented in the corpus.

A number of solutions have been proposed to alleviate n-gram models of their aforementioned issues. These techniques range from *smoothing*—the addition of a small amount of noise to the count of each word in the vocabulary, to *interpolation*—estimating the probability of a sequence using multiple language models (e.g. using bigrams and trigrams), to *backoff*—estimating an n-gram's probability using an $(n-1)$-gram; or a combination of the three. However, none of these solutions help solve the issue of the kinds of generalization that align well with the semantic inclination of this thesis.

2.1.2 Neural Language Models

This section presents an alternate method of training language models where instead of counting sequences from corpora, models are constructed to estimate word probabilities by encoding context information using dense representations, known as "embeddings." The core component of these models is a neural-network, and the models are trained using a class of methods known as backpropagation (Rumelhart, Hinton, & Williams, 1986). This section glosses over important training and optimization details of these models and narrows down its focus on language modelling in general, and how architectural decisions affect word prediction.

Feed-forward network Language Models: To tackle the issues of sparsity and semantic generalization faced by n-gram language models, Bengio et al. (2003) proposed the "Neural probabilistic language model." This model accepts as input a k-gram of word representations, and using a feed-forward neural network[2] to predict the probability of the $(k+1)$-th word. The k-gram input, $\mathbf{x}$, comprises of a concatenation of the embeddings of each of the k words, with the embedding of word w in the model's vocabulary $\mathcal{V}$, is represented as a d dimensional vector $e(w) = \mathbf{E}_{[w]}$, where $\mathbf{E}$ is an embedding matrix, $\mathbf{E} \in \mathbb{R}^{|\mathcal{V}| \times d}$. For a k-gram input, $\mathbf{x} \in \mathbb{R}^{1 \times kd}$. The following set of equations describe how the model processes the concatenation of the word embeddings, and produces the output probability distribution over the vocabulary of the model:

$$\begin{aligned}
\mathbf{x} &= [e(w_1); e(w_2); \ldots; e(w_k)], \\
\mathbf{h}_1 &= \sigma(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1), \\
\mathbf{h}_2 &= \mathbf{h}_1\mathbf{W}_2 + \mathbf{x}\mathbf{W}_3 + \mathbf{b}_3, \\
P(w_{k-1}) &= \text{softmax}(\mathbf{h}_2),
\end{aligned} \tag{2.10}$$

where h_1 and h_2 are hidden states of the network, σ is a non-linear activation function, and the softmax function converts its input into a probability distribution, like so:

$$\text{softmax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{d} e^{x_j}} \tag{2.11}$$

The collection, $\Theta = \{\mathbf{E}, \mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3, \mathbf{b}_1, \mathbf{b}_2\}$ represents the collection of trainable parameters of the language model, where $\mathbf{W}_1 \in \mathbb{R}^{kd \times h}, \mathbf{W}_2 \in \mathbb{R}^{h \times |\mathcal{V}|}, \mathbf{W}_3 \in \mathbb{R}^{kd \times |\mathcal{V}|}, \mathbf{b}_1 \in \mathbb{R}^{1 \times h}, \mathbf{b}_2 \in \mathbb{R}^{1 \times |\mathcal{V}|}$. The collection of parameters is referred to as "trainable-parameters" since they get updated during training, i.e. at every instance, the model produces its prediction for the next word [3], which is compared to the ground-truth using a loss function that returns the error in the model's prediction. This error signal is then propagated back to each individual trainable parameter using backpropagation. Using the backpropagation algorithm, the trainable parameters get updated such

[2] also known as a multi-layer perceptron (MLP).

[3] the word whose probability is the greatest in the model's output

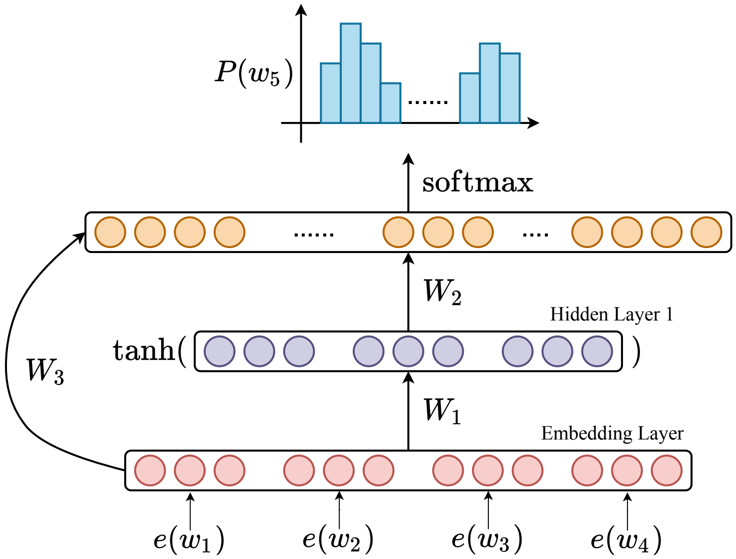


Figure 2.2. A Feed-forward neural network language model architecture, like the one presented in Bengio et al. (2003).

that the overall loss of the model is minimized. The model is trained through the entire training corpus a number of times, with each run referred to as a single “epoch.”

The output of the model is determined using a dot-product between the hidden state matrix and the dense representations of words, which has a notion of similarity in high-dimension space — a word’s fit to the context is being measured using similarity as opposed to counting. Due to this, the model exhibits semantically desirable properties such as generalizing to assigning high probability to sentences such as *“The dog chased the mice”* given that it has only encountered *“The cat chased the mouse”* during training The k-gram input representation can be considered as the predicted word’s context, suggesting that post-training, the embeddings of a word represent information about the kinds of contexts it occurs in, leading to properties akin to representations that embody the distributional hypothesis (Firth, 1957; Harris, 1954), the idea that “You shall know a word by the company it keeps.” as mentioned in Firth (1957). This is the essence of neural network based representations of words such as word2vec (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), GloVe (Pennington, Socher, & Manning, 2014), and fasttext (Bojanowski, Grave, Joulin, & Mikolov, 2017), where the vectors of words that occur in similar contexts are similar in vector space. For example, the top=10 closest words to the vector of the word *thesis* are *dissertation, doctoral_thesis, master_thesis, doctoral_dissertation,* *theses, hypothesis, theory, essay, theology,* and *essays*. Here, similarity of two vector is calculated using the cosine similarity metric, which measures the angle between two input vectors.

Although neural network-based distributional representations of words learned using architectures similar to the one described earlier show semantic tendencies, the language models themselves are still lacking in demonstrating broad-range linguistic qualities. A key bottleneck in this family of language models is the fixed value of k, tying back to the problems associated with the markov assumption, thus failing to model long-range dependencies. Another problem with these models from the point-of-view of linguistics is that these vector representations of words are largely uninterpretable, they do show lexical association from patterns of similarity, but lack any sort of structure to their organization in vector-space, preventing any explicit lexical semantic analysis. The class of models do not alleviate this deeper linguistics based issue—arguably they complicate things even further—but do attempt to tackle the issue of long-range dependency tracking in assigning probabilities to sequences.

Recurrent Neural Network Language Models: Originally proposed as a cognitive model of incremental language processing (Elman, 1990), a Recurrent Neural Network (RNN) model inputs that take the form of sequences. RNN language models (Mikolov, Kombrink, Burget, Černockỳ, & Khudanpur, 2011) have the desirable property of not being constrained by the markov assumption, and can theoretically model sequences of any length due to the notion of a "memory" component in their architecture which preserves information from previous time-steps. An RNN can be viewed as a feed-forward network with dynamic hidden layers, where the hidden state at a given time-step h_t is constructed using the current time-step input x_t, a dense representation of the word at that time-step, as well as the previous time-step's hidden state h_{t-1}. Therefore, the decision of the model at the given time-step is jointly informed by the "memory" from previous time-steps as provided by the hidden layer, and the input from the current times-step, enabling RNNs to represent, in theory, long sequences of free text. Formally, an RNN language model is represented by the following set of equations:

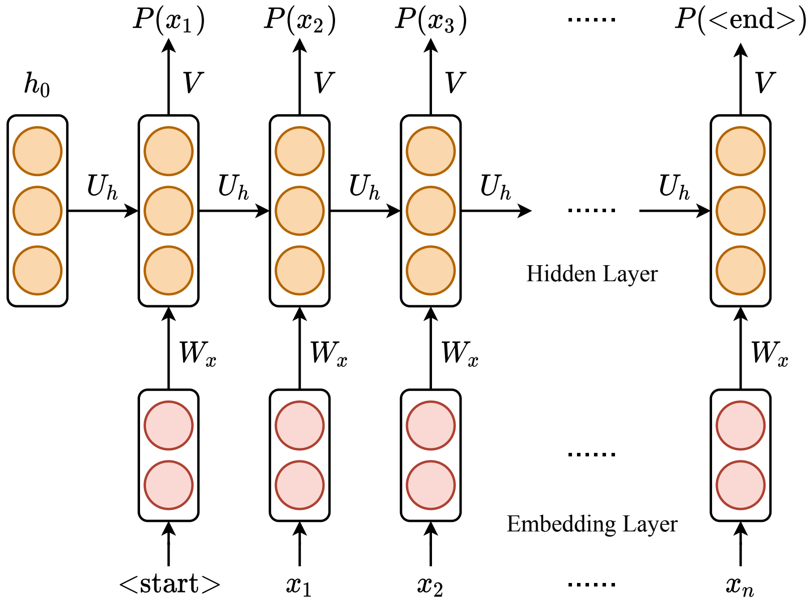


Figure 2.3. A schematic of an RNN language model (Elman, 1990).

$$\begin{aligned} \mathbf{x}_t &= e(w_t) \\ \mathbf{h}_t &= \sigma(\mathbf{x}_t \mathbf{W}_x + \mathbf{h}_{t-1} \mathbf{U}_h + \mathbf{b}_h), \\ \mathbf{y}_t &= \mathbf{h}_t \mathbf{V} + \mathbf{b}_v, \\ P(w_{t+1}) &= \text{softmax}(\mathbf{y}_t). \end{aligned} \tag{2.12}$$

Here, $\mathbf{W}_x$ is a matrix that projects every input embedding into the same high-dimensional space, $\mathbf{U}_h$ is a matrix that processes the previous hidden state to connect it to the current input, and $\mathbf{V}$ is a matrix that processes the current hidden state to produce scores, $\mathbf{y}_t$, which are then softmax-transformed to reflect probabilities of words given the previous context. The function σ is a non-linear activation function (usually $\tanh$), and $\mathbf{b}_h$ and $\mathbf{b}_v$ are the bias vectors. As with other forms of neural networks, RNN language models are also trained using backpropagation.

A common issue in optimizing RNNs for long sequences is the issue of *vanishing* or *eploding gradients*. During training, the hidden state $\mathbf{h}$ keeps getting multiplied by the weight matrix, $\mathbf{U}_h$. During backpropagation, the gradients of each hidden state at every time-step with respect to the loss get multiplied by the same quantity repeatedly, proportional to the length of the sequence, causing them to either become very large (explode) or driven towards zero (vanish), which causes problems in learning or representing the sequence. Despite their engineering issues, RNNs are core mechanisms in several cognitive models of language processing that simulate syntactic and semantic processing of sentences (Brouwer, Crocker, Venhuizen, & Hoeks, 2017; John & McClelland, 1990; Rabovsky, Hansen, & McClelland, 2018). This issue of vanishing gradients motivated the development of complex RNN structures such as the long short term memory network (LSTM) (Hochreiter & Schmidhuber, 1997) and the gated recurrent unit (GRU) (Cho et al., 2014). LSTMs and GRUs attempt to mitigate the problems posed by exploding and vanishing gradients by introducing the concept of "gating." A gate is a mechanism introduced in the LSTM and the GRU hidden layer that controls the flow of information from previous time-steps by allowing certain amount of gradients to flow during backpropagation unchanged, thus dealing with the problem of vanishing gradients. The process of gating involves a linear transformation of the input like in a feed-forward layer, followed by a sigmoid activation that pushes the values either to 0 (deletion of information) or 1 (passed through unchanged), followed by a pointwise multiplication with the layer that is being gated. This gating mechanism is used several times in an LSTM[4] unit, which maintains two forms of memory—the cell state $\mathbf{c}_t$ and a hidden state $\mathbf{h}_t$. Specifically, LSTMs consist of three gates: (1) the forget gate, which is responsible for deletion of information from the context that is not needed in the next time-step; (2) the input gate, which selectively adds information about the current context; and (3) output gate, which selects the information required for the output of the current time-step (such as the probability of a word, given the context, like in a language model). Mathematically, an LSTM is described by the following set of equations:

$$\begin{aligned}
\mathbf{f}_t &= \sigma(\mathbf{x}_t\mathbf{W}_f + \mathbf{h}_{t-1}\mathbf{U}_f + \mathbf{b}_f),\\
\mathbf{i}_t &= \sigma(\mathbf{x}_t\mathbf{W}_i + \mathbf{h}_{t-1}\mathbf{U}_i + \mathbf{b}_i),\\
\mathbf{o}_t &= \sigma(\mathbf{x}_t\mathbf{W}_o + \mathbf{h}_{t-1}\mathbf{U}_o + \mathbf{b}_o),\\
\tilde{\mathbf{c}}_t &= \tanh(\mathbf{x}_t\mathbf{W}_c + \mathbf{h}_{t-1}\mathbf{U}_c + \mathbf{b}_c),\\
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t,\\
\mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{c}_t).
\end{aligned} \tag{2.13}$$

[4] GRUs are very similar to LSTMs and are not discussed in detail within this section, see Cho et al. (2014) for a comprehensive introduction to GRUs.

The RNN model and its variants can be used to encode sentences in both directions — by training two separate models which encode the sequence in left to right (forward RNN) direction and in the reversed direction (backward RNN). These architectures are called *bi-directional RNNs*, or bi-RNNs (bi-LSTM, or bi-GRU for the gated versions). Training a biRNN further relaxes the assumptions of an RNN based language model allow it to look arbitrarily into a word's past and its future in order to compute its hidden state (usually a concatenation of the forward and backward hidden state). The bi-RNN hidden states of a word encode the information about its entire surrounding or "context," and are hence referred to as a word's "contextualized embedding." A contextualized word representation differs from the kinds of representations learned in regular window-based approaches because it completely depends on the local context a word occurs in, as opposed to the static nature of window-based word representations such as word2vec, where for example, the vector of *book* is the same when it occurs in *"I want to book an appointment."* and *"I enjoyed reading that book."* Contextualized word embeddings can then be used in other downstream tasks, and generally tend to improve task performance as compared to their static counterparts. This was shown by Peters et al. (2018), who introduced ELMo (**E**mbeddings from **L**anguage **Mo**dels), a two-layer bi-LSTM that produces a contextualized representation consisting of a word's static embedding, combined with its hidden states from layers 1 and 2.

A key computational issue in training RNN-based models is the complete lack of parallelization. This drawback stems from the fact that the hidden state at any given time step is dependent on all the previous hidden states, i.e., computing a hidden state requires the sequential computation of all the hidden states preceding it. This proves to be an unavoidable bottleneck in training such models and precludes parallelization. The next section describes an approach that avoids this issue by doing away with sequential recurrence and yet allows the learning of long term dependencies in a sentence in an entirely parallel fashion.

Transformer Language Models: A Transformer (Vaswani et al., 2017) is a neural network model that stemmed from the concept of *attention* (Bahdanau, Cho, & Bengio, 2014). Attention is a mechanism that was inherent in a majority of sequence modelling approaches (Sutskever, Vinyals, & Le, 2014) that were dominant in NLP prior to the advent of transformers and can be thought of a way to model a word's representation in a sequence by "attending" to other parts of the sequence. Attention resides at the core of transformers, which completely eliminate the need for sequential

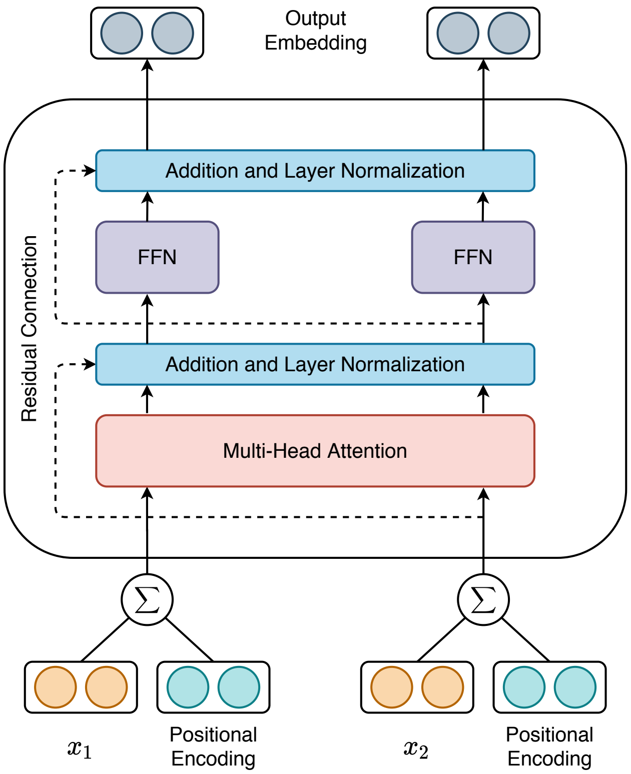


Figure 2.4. A schematic of a Transformer block (Vaswani et al., 2017). Typically, a transformer language model is made up of multiple such blocks, followed by a softmax layer to generate output word probabilities. The acronym FFN stands for "Feed-Forward Network."

recurrence in order to represent sequences of text. While the transformer network was initially proposed as an encoder-decoder model for neural machine translation, they can also be trained as language models. In particular, a Transformer model combines the following ideas to model sequences:

- **Multi-head Self-attention:** Self-attention is a component within transformers that relates different positions of the sequence and composes them together in order to form the representation of a specific position, allowing the model to learn long-range dependencies at every layer. Self-attention is described as a mapping between a token's query representation to a set of key-value representation pairs that store information about other tokens. Specifically, each token's input representation is projected to three individual vectors: queries that represent the focal word being operated on, packed together into a matrix Q; keys that represent the word that the query is being related to, packed together into a matrix K; and values that correspond to words that the keys refer to, packed together into a matrix V. The attention of a token, then, is a weighted sum of the value vectors, where the weight is calculated as a softmax transformation of the scaled dot-product between the queries and the keys. Mathematically, the output representations (packed as a matrix) formed by a single computation of self-attention are calculated as follows:

$$\text{Attended-Output}(Q, K, V) = \text{softmax}(\frac{QK^\top}{\sqrt{d_K}})V, \tag{2.14}$$

where the term 'softmax$(\frac{QK^\top}{\sqrt{d_K}})$' is the attention distribution between the tokens in the sequence. The module that performs the above computation is referred to as an "attention head." A single block of transformer typically tends to have multiple such heads which attend to the sequence independently, allowing the model to learn different relations between parts of the sequence (captured by different heads).

- **Positional Encodings:** Transformers inject information about the order in which words appear in sequences by adding embeddings that exclusively contain information about word positions. These embeddings are kept constant and their size is the same as the token/word embeddings that are used to represent the input of the sentence, allowing the two embeddings to be summed together to jointly denote a token/word and its position in the sequence.

By combining the two ideas of self-attention and positional encodings, Transformers circumvent the sequential bottleneck faced during training RNNs since the hidden state of at every position in the sequence can be computed simultaneously, allowing for massive parallelization and enabling faster computation. The general computation of a transformer involves the calculation of an input embedding, which is computed as the sum of the token and the positional embeddings. The input embedding is then passed to the multi-head attention layer, where multiple attended outputs are computed in parallel for the entire sequence, this output (represented as a concatenation of the outputs of the multi-head attention layer) is then added to the input embedding, creating a residual connection [5] that allows the passing of positional information unchanged. This is followed by layer normalization (Ba, Kiros, & Hinton, 2016), and then a feed-forward layer whose outputs are again

[5] in general, every transformer block involves this residual connection between the input to the block and the attended-output.

subject to concatenation by the previous layer normalized output, followed by another round of layer normalization, leading to the output of the computation performed by a single transformer block. To train transformers as an incremental language model, where sentences are processed one word at a time, Radford et al. (2019) proposed the GPT model. In GPT, the input is passed as an entire sentence, but the attention weights to the right to every word during its processing are zeroed out, preventing the model from accessing information from the future to predict the next word.

2.2 Pre-training by Language Modeling

The ideas discussed in the previous section describe a family of models that use supervision from freely occurring text in order to learn the probabilistic properties of language — what sequences are likely, what word follows a given context, etc. When neural networks are used to perform this task, they tend to produce representations for words that show semantic regularities in the way they are encoded, in the form of vectors, or embeddings (Mikolov, Sutskever, et al., 2013; Pennington et al., 2014). These word vectors serve as ideal initial states for models that perform higher level natural language processing tasks (Collobert & Weston, 2008) such as Question Answering (Rajpurkar, Zhang, Lopyrev, & Liang, 2016), Natural Language Inference (Bowman et al., 2015; Williams, Nangia, & Bowman, 2018), etc. Instead of starting from scratch, these models are now able to learn from already-trained representations which provide them with basic "prior" knowledge about the distributional semantic properties of inputs. The above account of extracting already trained, or "pre-trained" word representations and applying them as initial states in applications can be collectively termed under "feature-based" approaches, where the vector of a word corresponds to a collection of its features that encode some information.

Building upon the successes of using pretrained word embeddings in higher-level tasks, Howard and Ruder (2018) generalized the feature-based approach to "transfer" knowledge using entire models instead of a single embedding layer. Their general framework has three components: (1) Language model pre-training, which trains a neural language model on a large corpus, e.g. Wikipedia; (2) Language Model fine-tuning, where the neural language model adapts to the domain relevant to the task, e.g. movie reviews; and (3) classifier fine-tuning, where the representations are finally updated in order to perform the actual classification task, e.g. predicting the sentiment of movie-reviews. Therefore, instead of transferring pre-trained knowledge only at the initial layer of the model in the form of word embeddings, an entire language model is used to first learn generalized representations of the language and then fine-tuned to adapt to the domain of the task. The first of the three steps — language model pre-training, produces "contextualized" word representations of the kind described in the previous section, as a by-product. The language model pre-training followed by fine-tuning framework yielded state of the art results across a wide variety of NLP tasks at the time of their publication (Howard & Ruder, 2018; Radford, Narasimhan, Salimans, & Sutskever, 2018), leading to the development of BERT (Devlin et al., 2019), a model that brought about a radical shift in the field. The advent of BERT made fine-tuning based approaches a staple component of the modeling process.

2.3 BERT - Bidirectional Encoder Representations from Transformers

BERT is a pretrained language processing model that uses a transformer-encoder (Vaswani et al., 2017) as its core representation learning mechanism. BERT's transformer layers allow it to form contextualized word representations by informing the representation of a word the "attended" parts of its context, thus modeling complex relationships between words in a sequence. While the inquiry into the kinds of relations that are actually captured by the attention mechanism in BERT and its successors remains an active research endeavor, there is initial evidence that the models' attention mechanism actively contributes to the models' learning of linguistic syntax in an unsupervised manner (Clark, Khandelwal, Levy, & Manning, 2019; Manning, Clark, Hewitt, Khandelwal, & Levy, 2020). The BERT model is trained on pairs of sentences from the concatenation of Wikipedia and BOOKCORPUS (Zhu et al., 2015), comprising a total of 3.3 billion words, to perform the following tasks:

- **Masked Language Modeling (MLM):** A reformulation of the language modeling task where a certain percent of words in context are "masked" out and the model's task is to maximize the probability of these words in place of the masked positions. This task is inspired by cloze tasks (W. L. Taylor, 1953). Specifically in BERT, 15% of tokens in each input are masked—by replacing the missing word with a [MASK] token. This component of the modeling process imposes bidirectional properties within BERT, which in tandem with

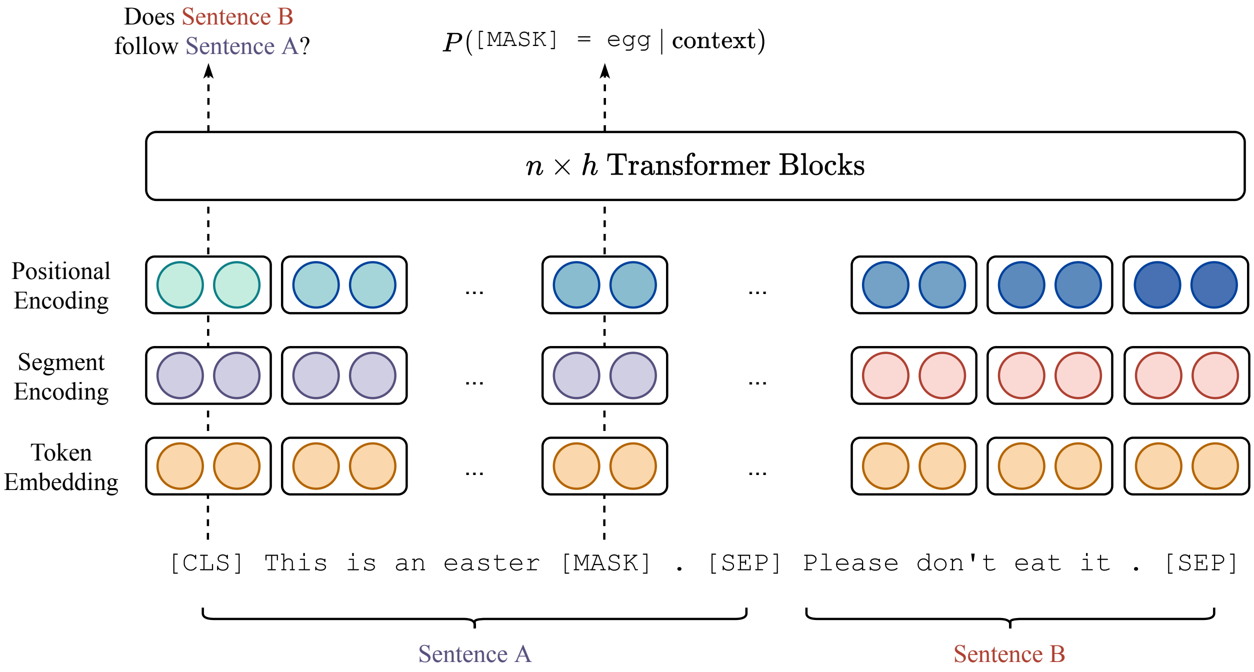

Does Sentence B
follow Sentence A?
P([MASK] = egg | context)
$n \times h$ Transformer Blocks
Positional Encoding
Segment Encoding
Token Embedding
...
[CLS] This is an easter [MASK] . [SEP] Please don't eat it . [SEP]
Sentence A
Sentence B

Figure 2.5. General setup of the BERT model (Devlin et al., 2019). Here, n refers to the number of transformer layers, and h refers to the number of attention heads.

the attention mechanism afforded by the scores of transformer blocks condition the model to rely on a word's entire context in order to produce its representation.

- **Next Sentence Prediction (NSP):** Primarily motivated by the fact that several NLP tasks require the modeling of the relationship between two sentences (such as Natural Language Inference and Reading Comprehension), Devlin et al. (2019) decided to jointly train BERT to also estimate whether the second sentence in its input pair follows the first one. While the exact explanation for how this component contributes to the semantic understanding of sentence pair relationship remains a mystery, the authors empirically showed that it leads to performance gains in downstream tasks.

During training, BERT accepts sequences of the form:

[CLS] This is sentence A. [SEP] This is sentence B. [SEP],

where [CLS] is a special token whose representation is used for fine-tuning on classification tasks, and [SEP] is a special token that indicates sentence separation or termination. BERT uses a word-piece[6] tokenization scheme (Wu et al., 2016) for converting the input sequence into tokens, with a vocabulary size of 30,522. In addition to static token embeddings, BERT uses a similar positional encoding as used by Vaswani et al. (2017), and an additional pair of embeddings to denote whether a word is in the first or the second sequence of the input sequence pair—this is referred to as "segment encodings". The final representation that is used as in input to the stack of transformers is the addition of the token, segment, and positional encodings. The BERT model comes in two variants, differing in the number of transformer blocks used after the input embedding layer: BERT-base: 12 transformer layers with 12 attention heads in each, with embedding and hiddent state size = 768, totalling to 110 million parameters, and BERT-large: 24 transformer layers with 16 attention heads in each, with embedding and hidden state size = 1024, totalling to 340 million parameters. Both models were trained by the authors using a batch size of 256 sequence-pairs for 40 epochs over 4 days on 4 Cloud TPUs (BERT-base), and 16 Cloud TPUs (BERT-large).

2.4 Linguistic Analysis of Pre-trained Language Models

The rapid progress in NLP has seen numerous instances which establish the importance and the benefits of using learned representations from language models to tackle a wide variety of natural language applications. However it is still unclear what language models "learn" during (pre) training. This lack of clarity primarily manifests in the black-box nature of their underlying neural network architectures, which used parameters and weights that are virtually uninterpretable. This has led to the emergence of a variety of methods that aim to improve the understanding of what linguistic knowledge does the language modeling objective impart to models.

This section summarizes at a broad level the various techniques used to analyze the linguistic capacities of neural language models, and reports on their general findings. As mentioned in Chapter 1, this thesis broadly classifies the prevailing analysis methods of investigating neu-

[6] wordpieces are arbitrarily segmented sub-word units, for instance, the word *playing* is segmented into `play` and `##ing`. The development of segmentation schemes corresponding to true morphological segmentation with good generalization on out-of-vocabulary words is an active area of research.

ral network based language processing models into two categories: (1) *probing* methods; and (2) behavioral methods.

2.4.1 Probing methods

Probing refers to the class of methods that investigate whether a specific linguistic or informational property can be decoded from the representations of a neural network using a supervised classifier known as a *probe* (Adi, Kermany, Belinkov, Lavi, & Goldberg, 2017; Ettinger, Elgohary, & Resnik, 2016; Veldhoen, Hupkes, Zuidema, et al., 2016). The concept of Probing is primarily motivated by the framework of Multivariate Pattern Analysis (MVPA) proposed by (Haxby, 2012) that uses a machine learning classification model to extract information from vectors derived from human brain imaging data.

In this paradigm, the black-box system (brain for MVPA, NLP model for probing) accepts inputs that differ in certain linguistic properties (for instance, the animacy of a noun) and produces intermediate representations of the input. In humans, these intermediate representations take the form of fMRI recordings of their activity patterns, while in NLP models they exist as vectors or embeddings. Whether or not the linguistic property is encoded in the representation is discerned using a supervised classification framework — this is akin to asking the question, "can property-containing inputs be consistently distinguished from non-property-containing ones?" If the classifier is able to distinguish between the representations that do and do not encode the property with a sufficiently high accuracy then the researcher concludes the presence of the property of focus within the representation of the black-box system. A caveat here is to ensure the classifier indeed extracts the property or is able to distinguish between inputs that demonstrate the property from ones that do not. This is ensured by using experimental designs common in machine learning literature — the idea of non-overlapping training and testing sets, where the model is trained on a specific set that contains various inputs, and then is tested on an independent, unseen set to measure whether the model achieves a performance on par with that observed during training. The input vectors are kept "frozen" or constant during training.

Probing was independently proposed by a number of researchers: Ettinger et al. (2016) proposed the use of classifiers to analyze sentence embeddings (state of the art representations at that time) for information about the semantic roles of AGENT and PATIENT. As preliminary experiments, they conducted sanity checks for the sentence encoders by probing them for the presence of lexical content (given the embeddings for sentences containing and not containing a given word, classify them into “has-word” and “doesn’t-have-word”) as well as shallow level semantic role information. They then extended this proposal to conduct large scale probing experiments for semantic role information on various carefully generated constructions (Ettinger, Elgohary, Phillips, & Resnik, 2018). Probing was also separately pursued by Veldhoen et al. (2016), who used supervised classification on the units of simple RNNs and GRUs to investigate their generalization for hierarchical structure for and artificially generated arithmetic language. They referred to probes as “diagnostic classifiers” and found the two linear processing architectures (simple RNNs and GRUs) to be competitive in performance to models that explicitly encoded tree-structures (Tree RNNs). Independently, Adi et al. (2017) proposed probes as “auxiliary classifiers” and analyzed sentence embeddings for information present within the sentence’s constituents such as word content (similar to Ettinger et al. (2016)), word order, and sentence length. Their findings suggested trivially that LSTM based sequence models encoded word content and word order better than bag-of-word representations formed by averaging the embeddings of individual words in the sentence.

Since then, a number of probing experiments have been conducted to compare and evaluate the linguistic knowledge of language models and their representations. Specifically, Tenney et al. (2019) extend the probing setup from investigating individual sentence and word embeddings to the concept of “edge-probing” to enable the analysis of span representations within sentences. Using edge-probing, the authors compared four kinds of contextual embeddings available at the time: CoVe (McCann, Bradbury, Xiong, & Socher, 2017), ELMo, OpenAI GPT (Radford et al., 2018), and BERT to their non-contextual counterparts word2vec (Mikolov, Sutskever, et al., 2013) and GloVe (Pennington et al., 2014). The two families of embeddings were compared across eight different sequence-based tasks: Part of Speech Tagging, Constituent Labeling, Dependency Labeling, Named Entity Recognition, Semantic Role Labeling, Coreference Resolution, Semantic Proto-role Labeling, and Relation classification. To this end they borrowed train and test sets from existing datasets suited for the above tasks. Their findings suggested that in general, contextualized embeddings encoded better information about each of the tasks than their non-contextualized counterparts; but this gain was more prominent in syntactic tasks than those of a semantic nature.

This general method of edge probing was also used by Shwartz and Dagan (2019) to conduct a comprehensive study of how lexical composition manifested within contextualized and non-contextualized models where they studied the phenomena of *meaning shift*, where the meaning of a phrase deviates from the individual meanings of its constituents, and *implicit meaning*, where the composition of lexical items introduces an implicit encoding of meaning that requires world-knowledge to understand. Contextualized word embeddings were found to better encode meaning shift than their non-contextual counterparts, whereas both embedding classes struggled almost equally in the more challenging task of detecting implicit meaning tasks, suggesting the need to better incorporate world knowledge in representation learning models. Hewitt and Manning (2019) extend probing and cast it as a regression problem by training a supervised linear transformation of the embeddings of two popular contextualized word embeddings, BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018) to assess their encoding of syntactic trees. Using this "structural probe" the authors provide evidence for the existence of syntactic trees within the embeddings which can be easily recovered from the L2-space of the subsequent learned linear transformation. Similarly, probing has been used to study the understanding of morphology (Belinkov, Durrani, Dalvi, Sajjad, & Glass, 2017), systematic compositionality (Hupkes, Dankers, Mul, & Bruni, 2020), and numeracy judgements (Wallace, Wang, Li, Singh, & Gardner, 2019), *inter alia*.

2.4.2 Behavioral methods

Deviating from using classifier models to investigate linguistic knowledge of pretrained language models, behavioral methods aim to analyze models in their natural setting — via controlled tests involving word probability estimation given some context. The underlying assumption of this method is that information encoded within language models is reflected in their behavior which can be measured using targeted stimuli. Due to the lack of supervision in these tasks, the question about whether the methodology does the heavy-lifting of learning a linguistic property that is prevalent in the probing literature simply does not arise. Instead of implicit analysis of language model representations, behavioral tasks aim to directly query language models in the form of carefully constructed stimuli accompanied by an explicit set of inferences that can be drawn from the responses made for them by the model. This formulation of the analysis method allows the researcher to test a diverse set of stimuli, often directly borrowed from related fields such as cognitive science or psycholinguistics (Ettinger, 2020; Futrell et al., 2019) to assess linguistic patterns within the models.

The general setup of a behavioral analysis is as follows: first, the researcher identifies the lexical phenomena of interest that may be captured by a model. For instance, Gulordava, Bojanowski, Grave, Linzen, and Baroni (2018) sought out to analyze LSTMs using the task of subject-verb agreement, since the model's success was primarily attributed to its learning of long-range dependencies in sentences. The researcher then constructs or samples stimuli that represent the phenomenon; in their paper, Gulordava et al. (2018) automatically extracted long-distance dependencies, where the agreement terms (the syntactic subject and its head verb) are separated by an arbitrary amount of words, using a large dependency treebank (Nivre et al., 2016). A simplified example is as follows:

(2.1) The **keys** to the **cabinet**..

Here, **"keys"** is the syntactic-subject, and **"cabinet"** is an attractor [7]. The test of whether the language model captures subject-verb agreement is then performed by measuring the probabilities of the next word. If the model assigns a greater probability to *are* than *is* for a significant majority of such stimuli, then there is an indication that the model has successfully picked up on patterns corresponding to subject-verb agreement during training. Language model probabilities play a central role in behavioral tasks, revealing insights into the explicit surface level patterns captured by the model's complex architecture. Common analysis methods use some form of a probability measure (for instance, word surprisals or sentence log-probabilities) to investigate language models.

The paradigm of examining the output probabilities language models as a proxy for their behavior in response to targeted stimuli has been dominated by syntactic analysis. Following the results reported by Gulordava et al. (2018), Marvin and Linzen (2018) manually construct a set of stimuli target a wider set of syntactic dependencies. In their analysis, RNNs performed well on local agreement dependencies but made errors whose frequency increased with an increase in the number of attractors. Furthermore, RNNs showed a lack of negative polarity item licensing effects in stimuli with simple as well as reflective NPI licensing. The performance here was measured

[7] a word with the same POS as the cue noun but different number.

by comparing the overall RNN probability of a sentence in grammatical vs ungrammatical constructions. Apart from the thorough analysis of agreement-based generalization, researchers have also shown empirical evidence about LSTMs and RNNs representing information (albeit imperfectly) about filler-gap dependencies (Chowdhury & Zamparelli, 2018; Wilcox, Levy, & Futrell, 2019; Wilcox, Levy, Morita, & Futrell, 2018), subordinate clauses, and simple garden-path effects (Futrell et al., 2019). In each of these assessments, the syntactic information is probed by comparing word-level surprisals (negative log probability of a word given context). This suggests that incremental language models such as LSTMs and RNNs are able to generalize to hierarchical syntax despite their linear processing of linguistic input.

In the realm of semantics-guided behavioral judgements of language models, Ettinger (2020) borrows linguistic stimuli used in psycholinguistics experiments of human sentence processing to assess BERT's knowledge of commonsense and world knowledge, negation, and semantic roles. In particular, these stimuli reflect cases that show a divergence between: (1) cloze probability — the proportion of humans that predict a particular word in context; and (2) brain responses such as the N400 (Kutas & Hillyard, 1980) that are reduced in amplitude to anomalous sentence endings, missing out on key contextual information present in the stimuli. The divergence between N400 responses and cloze probability measures suggests that human brains are insensitive to certain predictive information in the sentence while responding to linguistic input, and therefore these stimuli might present a challenge to predictive models that draw on information from the context to compute word expectation. Through an extensive analysis of three such small but informative sets of stimuli (Chow, Smith, Lau, & Phillips, 2016; Federmeier & Kutas, 1999; Fischler, Bloom, Childers, Roucos, & Perry Jr, 1983), Ettinger (2020) reports that BERT is able to distinguish between good and bad completions in cases involving (1) commonsense and pragmatic reasoning — for example, it assigns greater probability to *lipstick* as opposed to *mascara* or *bracelet* in the sentence input *“He complained that after she kissed him, he couldn't get the red color off his face. He finally just asked her to stop wearing that ___.”*; and (2) attributing nouns to their hypernyms — it predicts a noun's hypernym in the context, *“A* `[noun]` *is a ___.”* 100% of the time in its top-5 predictions. BERT however struggled in cases involving: (1) event knowledge — the models did prefer good completions by using the noun position in instances involving role-reversals, for instance — they predicted *served* in the place of ___ in *“the restaurant* *owner forgot which customer the waitress had* ___ *"* with higher probability than in *"the restaurant owner forgot which customer the customer had* ___ *"*, but were less sensitive than humans, and failed to match the types of predictions on which humans converged; and more prominently, (2) negation — the model shows complete insensitivity to the presence of negation in context, e.g. it predicts *bird* as the top completion in the sentence *"A robin is not a* ___ *,"* showing a complete inability to prefer true over false statements about the noun-hypernym relation. This suggests that the model is able to utilize information from its context in simple cases where it has to attribute nouns to their hypernyms or when the context contains words (*kissed, red, wearing*) whose features are predictive of a certain completion (*lipstick*), reflecting patterns corresponding to commonsense and world-knowledge. However in cases that reverse the truth of the hypernymy relation or involve event-knowledge, BERT shows a glaring lack of "understanding" that is shown by humans. Further evidence about BERT's recall of factual or world-knowledge was shown in a large scale analysis conducted by Petroni et al. (2019), who cast knowledge base triples such as `(dante, born-in, florence)` into cloze sentences, *"Dante was born in* ___ *."* and evaluated the model's word probabilities for the missing position. Their study revealed BERT to perform competitively with supervised Q/A models that had access to an entity-linking oracle — suggesting the potential of language models to retrieve factual knowledge without any supervision or fine-tuning. This study was further expanded by Kassner and Schütze (2020), who introduced the "negated" and "mispriming" probes. The negated probe converted the queries used by Petroni et al. (2019) into negated instances, and further corroborated evidence that pre-trained word prediction models are strongly insensitive to negation, similar to Ettinger (2020). Using the mispriming probe, the authors showed BERT to be easily distracted by misprimes—words chosen to be prepended to cloze-like sentences. For instance, BERT-large predicted *Cicero* as the completion in place of the correct answer, *Plato*, when the previous query is modified to "*Cicero? Platonism is named after [MASK].*"

2.4.3 Probing vs. Behavioral Methods, in brief

A key advantage that both probing and behavioral methods hold are that they are grounded in neuroscience (MVPA in the case of *probing*) and psycholinguistics (behavioral studies such as cloze tasks and human ratings) literature. Furthermore, the two broad classes of methods facilitate different types of investigation into the models. For instance, the knowledge of POS or NER is challenging to test using word probability estimates, while the knowledge of event structure is non-trivial to extract using probing classifiers. Both sets of methods involve crucial experimental design decisions to make — probing analyses require that the testing set does not contain any elements of the train set in order to make faithful conclusions about the model's encoding of information; within behavioral methods, there is considerable amount of work required to carefully construct stimuli that explicitly target the phenomenon of interest. Finally, when the test of the presence of a phenomena can be expressed within both sets of methods, their analyses lead to slightly different conclusions depending on the type of method used (Warstadt et al., 2019), suggesting the lack of proper generalization within these methods, signaling the community to continuously invest time and resources to develop better and more reliable analysis methods.

2.5 Ontological Semantics Technology (OST)

The previous sections have described empirical research in NLP predominantly rooted within statistical learning techniques. This empirical paradigm is enabled by the existence of large corpora and manual annotation that allow the modeling of language by picking up patterns largely based on co-occurrence statistics. It can be argued that this class of methods only allows for the modelling of surface-level usage patterns rather than true textual meaning. This is further substantiated by the fact that the kind of computational models that exist in this paradigm have no in-built mechanism to explicitly represent the knowledge about the world. It is hypothesized that the models pick up world knowledge and semantics somehow during their training process but the study of whether or not this is true remains an active area of research. Nevertheless, this general statistical paradigm for NLP is often referred to as "knowledge-lean" (McShane, 2017). This section briefly describes a strikingly different approach to knowledge representation, based purely on meaning as opposed to patterns derived from large corpora. This approach is known as Ontological Semantic Technology (Hempelmann et al., 2010; Raskin et al., 2010; J. M. Taylor, Hempelmann, & Raskin, 2010), and it belongs to the school of Ontological Semantics (Nirenburg & Raskin, 2004).

The goal of the Ontological Semantics paradigm is to construct a model of the world (an ontology) that can facilitate the representation, extraction, and reasoning of meaning present in text. OST is a practical realization of these goals, and is built to reflect human reasoning and world knowledge. The OST-system comprises of the following components:

- A language independent ontology, which is a graph data structure whose vertices are concepts and edges are the properties that define relations between each concept. Concepts within the ontology are hierarchically organized, with the most common relation being subsumption, or the IS-A relation—for instance, a CHAIR IS-A FURNITURE.
- A lexicon per supported language, that defines word senses of a language and maps them to appropriate concepts or properties in the ontology.
- A set of linguistic modules, consisting of morphological analyzers, and parsers for syntax and semantic dependencies that enable processing of input by mapping it to the conceptual structure as permitted by the ontology.
- An information repository that stores processed input in representation form for use in reasoning/further applications.

Language processing within the OST is event-driven. At its core, OST represents linguistic inputs (usually sentences) along the event (usually a verb) that is being described within it, generating a text meaning representation (TMR). Text meaning representations embody the various concepts (OBJECTS and EVENTS) present in the ontology, and the inter-relationships between them (through properties).

The process of generating TMRs first involves the process of semantic disambiguation of the given text—the mapping of lexemes to their possible senses (or interpretations). Each interpreted concept is accompanied by its properties as permitted by the ontology (for instance, the IS-A relation) as well as those defined in the text. The VALUE of a property links the current concept with the concepts present in the linguistic input. This basic representation is further augmented by incorporating external knowledge that allows the access of information beyond the input text. This external knowledge augmentation is done by defining what Nirenburg and Raskin (2004) term as "permissible facets" of a property. The permissible facets (denoted as just facets hereon) of a property describe its semantic constraints which take the form of a set of fillers, which are usually concepts. Note that facets are always defined for every property, but not necessarily filled. The following are some common facets that exist for properties.

- SEM: the fillers of this facet serve as a selection restriction constraint for the property, i.e., what concepts are allowed to satisfy the given property.
- REL-TO: the fillers of this facet describe the extent to which the restrictions defined by the fillers of SEM can be violated. This is usually seen in non-literal language such as metaphor or metaphor, e.g. *"She bought gifts for the house,"* where *house* metonymically refers to its residents.
- DEFAULT: the filler of DEFAULT is the most expected constraint of the property in a given concept. The DEFAULT of a property is usually the unspoken piece of information that is required to understand the meaning of a text (J. M. Taylor, Raskin, Hempelmann, & Attardo, 2010). In general, if $\mathcal{D}$ is the set of DEFAULT fillers of an event, and $\mathcal{S}$ is the set of of SEM fillers of an event, then $\mathcal{D} \subset \mathcal{S}$.
- NOT: The fillers of this facet are ones that should be excluded from being fillers of the property. It is possible that the fillers of NOT are a subset of fillers introduced by other facets.

Figure 2.6 depicts a minimal example of the collection of TMRs (Nirenburg & Raskin, 2004) for the sentence *"The cat ate the mouse."* It is assumed that OST has successfully disambiguated the sense of the event EAT as described by (Hempelmann et al., 2010).

2.5.1 On the Manifestation of Fuzziness within OST

In classical set theory, the individuals in a given universe of discourse (Ω) are defined to either be members or non-members of a set. These elements have clear, well-defined boundaries that distinguishes them, and therefore they are referred to be *crisp*. For example, the set of all integers, $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$. However, there are examples of items in the world that do not possess such sharp boundaries between them. For instance, the concept of height — when does

```
EAT
    AGENT:
        VALUE: CAT
        REL-TO: SOCIAL-OBJECT
        SEM: ANIMATE
    THEME
        VALUE: RODENT
        REL-TO: ANIMAL, PLANT
        SEM: FOOD
CAT
    AGENT-OF: EAT
RODENT
    THEME-OF: EAT
```

Figure 2.6. TMR for *"The cat ate the mouse."*

someone be classified as tall or short? Ideally, the "tallness" of a person should increase with an increase in the value of their measured height, and their "shortness" should decrease. But the boundary that separates the two is vague. Fuzzy sets (Zadeh, 1965) embody this notion of vagueness by formulating a generalization of crisp sets, by assigning intermediate degrees of membership for elements that do not completely fall under the crisp notion of the set. Here, the membership of 1 signifies perfect membership, while 0 signifies non-membership. Mathematically, a fuzzy set A is defined as:

$$
\begin{aligned}
&A = \{(x, \mu_A(x)), x \in \Omega\}, \\
&\mu : \Omega \rightarrow [0, 1],
\end{aligned}
$$

where x is an element in the universe of discourse, Ω, and $\mu_A(x)$ is defined as the degree of membership of x for the set A. Fuzzy sets are connected to classical crisp sets through the notion of α-cuts. A fuzzy set's α-cut is defined as a crisp set, ${}^{\alpha}A$, whose elements' membership is greater than or equal to α, i.e., ${}^{\alpha}A = \{x : \mu_A(x) \geqslant \alpha, x \in \Omega\}$. An α-cut of a fuzzy set allows a fuzzy set to be decomposed into a potentially infinite number of crisp sets. For a fuzzy set A, its *support* is defined as the set of all members belonging to Ω with non-zero membership values, i.e.,

$\{(x, \mu_A(x)), x \in \Omega$ and $\mu_A(x) > 0\}$, and its *core* is defined as the set of all members belonging to Ω with perfect membership values, i.e., $\{(x, \mu_A(x)), x \in \Omega$ and $\mu_A(x) = 1\}$. Given two fuzzy sets A and B, $A \subseteq B$ iff. $\forall x \in \Omega, \mu_A(x) \leqslant \mu_B(x)$. The standard operations involving fuzzy sets were defined by Zadeh (1965) as follows:

$$\textbf{Complement: } \mu_{\neg A}(x) = 1 - \mu_A(x)$$

$$\textbf{Union: } \mu_{A \cap B}(x) = \min\{\mu_A(x), \mu_B(x)\}$$

$$\textbf{Intersection: } \mu_{A \cup B}(x) = \max\{\mu_A(x), \mu_B(x)\}$$

The the most recent version of OST incorporates the notion of fuzziness through its property facets (Raskin & Taylor, 2009; J. M. Taylor & Raskin, 2010, 2016), which allows it to quantify the uncertainty in the selection restriction constraints imposed on the property. Within a fuzzy instance of the OST, each facet is treated as a fuzzy set, whose fillers denote concepts with varying degrees of memberships. J. M. Taylor and Raskin (2016) discuss the most recent advances in using fuzzy-logic within OST. In general, for an event with some property p, and fillers of facets NOT (N), REL-TO (R), SEM (S), DEFAULT (D), and a function desc(), which denotes "descendant-of," the memberships of concept x for p follow the following pattern:

$$\mu_p(N) = 0 < \mu_p(R) < \mu_p(S) < \mu_p(\texttt{desc}(D)) < \mu_p(D) = 1, \tag{2.15}$$

Furthermore, each facet of a property defines an α-cut, with $\alpha_{\text{DEFAULT}} = 1, \alpha_{\text{SEM}} = 0.75$, and $\alpha_{\text{REL-TO}} = 0.05$ (J. M. Taylor & Raskin, 2010, 2016). The membership of a concept is then defined based on the hierarchy within which it lies, as a function of its path-length ϕ between the various facet concepts within the ontology hierarchy (with root node ALL). In its latest formulation the membership of a concept x as a filler for a given property p was defined initially by J. M. Taylor and Raskin (2010) and later modified by J. M. Taylor and Raskin (2016) to follow the following function:

$$\mu_p(x) = \begin{cases} 1 & x = D \\ 1 - \frac{(1-\alpha_{\text{SEM}})\times\phi(D,x)}{\phi(S,D)} & x \in D \text{ or } x \in S \\ \mu(S) - \frac{(\mu(S)-\alpha_{\text{REL-TO}})\times\phi(S,x)}{\phi(R,S)} & x \in R \text{ and } x \notin S \\ \mu(R) - \frac{\mu(R)\times\phi(R,x)}{\phi(R,\text{ALL})} & x \notin R \text{ and } x \notin N \\ 0 & x \in N \end{cases} \tag{2.16}$$

Equation (2.16) essentially suggests that when a property's DEFAULT is defined, all its descendants are assigned decreasing membership values, as opposed to being the same as the DEFAULT, as was the case in previous iterations and applications of Fuzzy-OST (J. M. Taylor & Raskin, 2010, 2011; J. M. Taylor, Raskin, & Hempelmann, 2011). The membership of the descendant decreases until the SEM value is reached, and the decrease is proportional to the distance between the descendant and the default in the ontology graph, signifying a triangular-like membership function. The following example uses the above notions to demonstrate the calculation of memberships for the concepts, CAT and GOVERNESS, as fillers for the property AGENT of the event TEACH, which is depicted in Figure 2.7. Following J. M. Taylor and Raskin (2010, 2016), the memberships of the DEFAULT (TEACHER), SEM (HUMAN), and REL-TO (ANIMATE) are set to 1, 0.75, and 0.05 respectively, equivalent to their α-cuts.

```
TEACH
    AGENT:
        REL-TO: ANIMATE
        SEM: HUMAN
        DEFAULT: TEACHER
```

Figure 2.7. Event representation for TEACH

For the same example, assume that the following are the path-lengths, $\phi(X, Y)$, between the various concepts[8]:

$$\begin{aligned}\phi(\text{GOVERNESS}, \text{TEACHER}) &= 1\\ \phi(\text{TEACHER}, \text{HUMAN}) &= 13\\ \phi(\text{HUMAN}, \text{ANIMATE}) &= 9\\ \phi(\text{CAT}, \text{HUMAN}) &= 6\end{aligned}$$

Using equation (2.16), the membership of CAT as the AGENT of TEACH is calculated as:

$$\mu_{\text{AGENT}}(\text{CAT}) = \mu_{\text{AGENT}}(\text{HUMAN}) - \frac{(\mu_{\text{AGENT}}(\text{HUMAN}) - \alpha_{\text{REL-TO}}) \times \phi(\text{CAT}, \text{HUMAN})}{\phi(\text{HUMAN}, \text{ANIMATE})}$$

$$\begin{aligned}\mu_{\text{AGENT}}(\text{CAT}) &= 0.75 - \frac{(0.75 - 0.05) \times 6}{9}\\ \mu_{\text{AGENT}}(\text{CAT}) &= 0.283\end{aligned}$$

Similarly, the degree of membership for GOVERNESS is given by:

$$\begin{aligned}&\mu_{\text{AGENT}}(\text{GOVERNESS}) =\\ &\qquad \mu_{\text{AGENT}}(\text{TEACHER}) - \frac{(\mu_{\text{AGENT}}(\text{TEACHER}) - \alpha_{\text{REL-TO}}) \times \phi(\text{GOVERNESS}, \text{TEACHER})}{\phi(\text{TEACHER}, \text{HUMAN})}\end{aligned}$$

$$\begin{aligned}\mu_{\text{AGENT}}(\text{GOVERNESS}) &= 1 - \frac{(1 - 0.75) \times 1}{13}\\ \mu_{\text{AGENT}}(\text{GOVERNESS}) &= 0.964\end{aligned}$$

Using the concepts discussed in the previous subsections, J. M. Taylor and Raskin (2011) and J. M. Taylor et al. (2011) proposed a computational approach towards guessing the meaning of an unknown word (unattested input) in context by formulating it as a cloze-task (W. L. Taylor, 1953). In such a formulation, the functional details in the unknown word's context determine the basis of understanding the meaning of the unknown word. The process involves first decom-

[8] for the purposes of this demonstration, the path-lengths have been gathered from WordNET (Miller, 1995).

posing the event that affects the unknown word (often represented as *zzz*) using a TMR. Then, depending upon the interpretation of the event, the unknown word takes the position of the possible property-fillers that are anchored along each interpretation. The approximate meaning of the unattested concept is then narrowed-down by accessing each property's facets that supply semantic membership values to concepts which give cues to potential interpretations.

2.6 Semantic Priming

This section describes the cognitive phenomenon of semantic priming, followed by two major explanatory accounts that attempt to explain how semantic priming relates to the organization of semantic memory in human minds. The section ends with a brief discussion on how the phenomenon of semantic priming and its inferences about lexical association form the primary motivation behind the experiments proposed in the next chapter.

Semantic Priming refers to the phenomenon in which language comprehenders tend to show a speed up in response to a word when the word is preceded by a semantically related stimulus relative to a semantically unrelated stimulus (McNamara, 2005; Meyer & Schvaneveldt, 1971). For example, in a cognitive task, the time taken to respond to a word like *dog* is faster when it is preceded by a word like *cat* as compared to when it is preceded by a word like *table* — when this happens then it is said that "*dog* facilitates or *primes* the activation of *cat*." The word to which the response is made is referred to as the *target* and the preceding stimuli are called *primes* (either related or unrelated). Levels of priming are evaluated based on participants' response times (RT), which are measured after the target has been shown to them. The RTs to priming stimuli are typically measured using two very common cognitive or perceptual tasks — Lexical decision and Naming. In standard priming methodology, participants are shown the prime stimulus and then asked to perform the task. The prime stimulus and the stimuli for the task are both shown using a visual medium, either with words or pictures of the concepts the word-stimulus refers to. In a lexical decision task, participants are instructed to decide and respond about whether the presented string of characters is a word or a non-word, whereas in a naming task, participants are instructed to rapidly pronounce the word out loud. In both of these tasks, the RTs to target words are often compared between the related prime word and an unrelated prime word as a baseline comparison. The strength of association between the target and the related word is then reflected in the difference between the two RT measures, which also termed as the semantic priming effect (Hutchison, Balota, Neely, Cortese, Cohen-Shikora, et al., 2013; McNamara, 2005).

Of all explanatory models about semantic priming, there are currently two major competing accounts of how priming occurs. In the first account, lexical facilitation occurs as a result of the activation of concepts within a semantic network using a process known as "spreading activation" (Collins & Loftus, 1975; Quillian, 1967). In this theory, semantic memory is conceived as a network of concepts connected to each other by links that represent association. During the cognitive task during the priming experiment, word processing activates its anchored concept within the semantic network and this activation spreads through to related concepts which make the processing of related words easier. The activation of a concept is usually inversely proportional to the path length within the network (Collins & Loftus, 1975) — the number of links traversed. In the second account, concepts in semantic memory are connected not through symbolic links but rather through feature overlaps, i.e., they are represented as densely connected units or "distributed representations" (McClelland & Rumelhart, 1986; McRae, Cree, Seidenberg, & McNorgan, 2005; Plaut, 1995) which can be adjusted during acquisition and correction. During the priming experiment when a related word precedes the word that is being processed, the features of all concepts close to the related word get activated, causing the ease of retrieval of the target concept. This account gave rise to *connectionist models* (McClelland & Rumelhart, 1986), the precursors to modern day neural networks in which inputs are represented as dense vectors and "learning" corresponds to an adjustment of weights. Regardless of the theories, the semantic priming paradigm provides interesting and useful insight about the organization of semantic memory in the human brain (Hutchison, 2003), especially due to the non-semantic nature of the tasks that priming participants are asked to perform.

2.7 Role of Sentence Constraints in Lexical Facilitation

Lexical facilitation has also been studied in humans for sentence inputs using the priming tasks discussed in the previous section, as well as by investigating N400 amplitudes from brain recordings. The N400 (Kutas & Hillyard, 1980) is an ERP component, extracted from EEG recordings of the human brain during sentence comprehension. It is a negative-going deflection elicited in response to words that peaks at about 400ms after word-onset in the presence of a semantically anomalous component in the sentence. For instance, an N400 peak is observed in sentences such as *“I take coffee with cream and dog.”* Its amplitude largely shows strong correlations with cloze probabilities of words in sentence contexts (Kutas & Hillyard, 1984)[9]. In most of these studies, a primary factor that is often considered and taken into account as a control is the nature of the sentence context in terms of the constraint it imposes on the word-position of interest (Schwanenflugel, 1991). The constraint imposed by a sentence context on any given position is often been defined in terms of the cloze probability of the most expected word in the word position. For example, the context *“He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of* ___ *.”* is considered high constraint by Federmeier and Kutas (1999) since the most expected word, *football* has a cloze-probability that lies between .784 and 1.0. In contrast, for the context *“Fred went to the pantry and got out the homemade jelly his grandmother had brought. Fifteen minutes later, however, he was still struggling to open the* ___ *.”* the most expected completion is *jar*, with a cloze-probability that lies between .17 and .784, and is thereforet considered to be low constraint[10].

Experimentally, the general consensus with respect to sentential or contextual constraints is that high constraint sentences typically only show facilitation for the most expected word in context and have a narrow scope of lexical access, as opposed to low constraint sentences, which show a wider scope of facilitation that extends to less predictable lexical items that fit in the context. This is evidenced in lexical decision experiments involving sentence contexts as primes (Schwanenflugel & LaCount, 1988; Schwanenflugel & Shoben, 1985) where low constraint items yielded significant facilitation for unexpected completions while high constraint items only facilitated expected words. Similar results have also been shown in naming experiments (McClelland & O’Regan, 1981; Stanovich & West, 1983). It is important to note that while unexpected completions are facilitated by low constraint contexts, it is only when the completions are semantically related to

[9] Although there are instances when the N400 deviates from this widely hypothesized pattern, see Fischler, Childers, Achariyapaopan, and Perry Jr (1985) and Nieuwland and Kuperberg (2008)

[10] actual probabilities not revealed by Federmeier and Kutas (1999). However the high-low distinction was done using a median split of the most expected word.

the actual expectation of human participants, i.e., semantically unrelated words that do not fit into the context are not facilitated at all.

2.8 Summary

The prior work covered in this chapter focuses on several different concepts within the extant literature of NLP that guide and motivate the methodological and analytical framework discussed in the next chapter. The concept of word prediction and its development from count-based probabilistic models to fundamental blocks of modern-day language representation learning mechanisms motivates its analyses — this thesis treats one such word prediction model, BERT, as the primary subject of investigation. The discussion on the two primarily used tools for the analysis of language models and their findings sheds light on what we know about the kinds of information encoded by the models, and the various perspectives each class of methods brings with it. Behavioral analyses that are guided by semantics reveal the ability of LMs to inform their word probabilities primarily using lexical, as opposed to event level, knowledge encoded in the sentence. Of all the findings discussed, this can be evidenced simply by examining the results of BERT's word prediction in negated sentences (Ettinger, 2020; Kassner & Schütze, 2020). Consider the example of the cloze sentences with an without negation of the same message — *A robin is not a ___ ."* and *"A robin is a ___ ."* The observation that BERT predicts a lexical associate (*bird*) in both sentences with relatively high probability (as compared to other words) alludes to the fact that BERT focuses (or perhaps "attends") more on lexical associates than message level constituents of the sentence.

Semantic priming and its primary use as a fundamental method to reveal the organization of complex lexical semantic concepts within the human brain inspires the methodological contributions of this work. The phenomenon's experimental nature helps narrows-in on the choice of using behavioral methods as opposed to *probing* to explore the lexical relations between BERT. While semantic priming serves as the primary motivation behind the kinds of word-level stimuli used in subsequent experiments, the actual properties of the sentence stimuli are characterized with the help of research conducted in understanding lexical facilitation within sentence context. The specific property that concerns the methodology presented in the next chapter is that of a sentence's predictive constraint. Prior evidence lets us infer that the lexical semantic associates in context inform word prediction model's output in a non-trivial manner. However, there is still insufficient understanding about exactly how these models use lexical relations and how their output is modulated based on the nature of the context itself. To fill this gap, the methodology proposed in this thesis borrows from an experimental paradigm that investigates the organization of semantic memory in human brains *(semantic priming)*, and how this organization interacts with varying levels of the incoming linguistic input's predictive constraints.

This thesis also ties in the concept of Ontological Semantics and its latest product, the OST, both of which lie at a radically opposite spectrum within NLP methodological paradigms. OST's meaning-first approach to knowledge representations, its functionality in decomposing a sentence by representing its event interpretation, and its inherent mechanism of representing semantic constraints through the incorporation of fuzziness within its property facets lends itself to be a suitable analytical framework for cloze-contexts — sentences that bear close resemblance to the stimuli used in the methods. This is elegantly shown in previous work that casts cloze tasks as "guessing the meaning of an unknown word" (J. M. Taylor et al., 2011), and is later borrowed to qualitatively analyze the stimuli in this thesis.

CHAPTER 3. METHODOLOGY

This chapter describes in detail the various methodologies and techniques required to answer the research questions posed in Chapter 1. The chapter borrows heavily from the methodological discussions in the original contributions mentioned in Section 1.5 (Misra et al., 2020a, 2020b). They are restated below:

- Misra, K., Ettinger A., Rayz, J.T. (2020). Exploring Lexical Relations in BERT using Semantic Priming. In *42nd Annual Virtual Meeting of the Cognitive Science Society.* (Poster Presentation).

- Misra, K., Ettinger A., Rayz, J.T. (2020). Exploring BERT's Sensitivity to Lexical Cues using Tests from Semantic Priming. *Findings of ACL: EMNLP 2020.* (Long Paper)

- Misra, K., & Rayz, J.T. (2020). An Approximate Perspective on Word Prediction in Context: Ontological Semantics meets BERT. In: *2020 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS).* (Regular Paper; forthcoming)

3.1 Nature of Stimuli

The stimuli used in subsequent analyses take the form of a sentence with an omitted word (see example (3.1)), and the model has to rely on the context to infer what word best completes the sentence.

(3.1) I was reading a ___ .

These sentences bear resemblance to the stimuli prevalent in cloze tasks (W. L. Taylor, 1953) — tasks where participants are presented with incomplete sentences and are asked to fill in the blank by relying on the context surrounding the missing word — and therefore are referred to as "cloze contexts."

Such stimuli are compatible with BERT's masked language modeling procedure, briefly described in Section 2.3. These stimuli offer insight into the kinds of knowledge that, in principle, can be extracted from the context surrounding the blank position in order to complete the sentence. The following Section expands qualitatively on how a system that demonstrates natural language understanding can approach this task.

3.2 Analyzing cloze contexts through the lens of Ontological Semantics

This Section offers a perspective on the specific semantic knowledge needed to "guess" the best completion to cloze contexts through the fuzzy-inferences supported by the OST system. This perspective is highly qualitative, and only relies on the rather quantitative nature of fuzzy membership functions to provide relative comparisons to the probabilistic outputs of a language model. This method directly follows the one discussed in Section 2.5.1, i.e., works that proposed a fuzzy approach to understanding the meaning of an unknown word (in a sentence-context) using the OST system (J. M. Taylor & Raskin, 2011; J. M. Taylor et al., 2011).

Within the OST framework, a given cloze context is decomposed along the event, E, that affects the missing word. This event is represented in a notation which lists its various properties, along with their inferred facet values. The properties and facets are inferred based on the functional elements in the cloze-context. An example event notation is shown in Figure (3.1).

```
E
    PROPERTY-1
        REL-TO:
        SEM:
        DEFAULT:
    PROPERTY-2:
        ...
```

Figure 3.1. Minimal representation notation for events described by cloze-contexts.

As described in Section 2.5.1, the facets (NOT, REL-TO, SEM, DEFAULT) of a property represent its semantic constraints and indicate concept memberships for properties that are endowed to the event, E. In this qualitative interpretation of cloze-context, the membership of a concept, $\mu(C)$ denotes the extent to which it satisfies the semantic constraints imposed by the specific property of the event, and follows the same pattern as described in Equation (2.15)

The decomposition of cloze-contexts into events with facet memberships allows discerning what concepts are evoked by a context, and qualitatively comparing against language model outputs. It also allows relative comparison of the semantic constraint (more formally described in Sections 3.3 and 3.4) applied on the missing word-position. This relative comparison can be demonstrated using the following example contexts, each denoting the event, WASH:

(3.2) a. I washed my ___ .

b. I used laundry detergent to wash my ___ .

```
WASH
    AGENT: HUMAN
    THEME: ??
        SEM: PHYSICAL-OBJECT
```

(a) Event representation for example (3.2a).

```
WASH
    AGENT: HUMAN
    INSTRUMENT: LAUNDRY-DETERGENT
    THEME: ??
        SEM: PHYSICAL-OBJECT
        DEFAULT: CLOTH-ITEM
```

(b) Event representation for example (3.2b).

Figure 3.2. Event representation of WASH in example (3.2).

In both examples, the missing position is interpreted as the THEME of the event WASH. Without any other properties afforded to WASH in example (3.2a), its THEME stipulates any PHYSICAL-OBJECT to serve as the minimum selection restriction (SEM). Hence any instance or descendant of the concept PHYSICAL-OBJECT would receive equal membership. When the same WASH event is endowed with a property, such as an INSTRUMENT with value LAUNDRY-DETERGENT like in example (3.2b), its THEME now carries a DEFAULT facet of CLOTH-ITEM. This causes its membership values to readjust, with $\mu_{\text{THEME}}(\text{WASH}, \text{CLOTH-ITEM}) = 1$. At the same time, the membership for every descendent of CLOTH-ITEM, such as COAT, TROUSER, etc., is now greater than that of other physical objects (UTENSILS, BODY-PART, etc.). This allows a direct behavioral comparison between LMs and OST's fuzzy inferences, where LM probabilities for completions can be matched up against lexical instances of the concepts evoked within the OST framework for the given cloze-context. The membership pattern change in example (3.2) also indicates a clear distinction in the semantic constraint of the two cloze contexts — example (3.2b) is more constraining because of the additional INSTRUMENT property, which is absent from (3.2a).

The above example discusses the dynamics of property fillers when properties are explicitly endowed to an event, for instance, WASH with INSTRUMENT = LAUNDRY-DETERGENT causes certain concepts to get promoted in the membership rankings, while also actively demoting others. These dynamics can also implicitly arise when the event is "primed" with a certain concept. Priming in the context of OST and its inbuilt fuzzy mechanisms can be considered akin to an unstructured endowment of properties to events. They are similar in the sense that both of them cause a shift in the membership values of certain fillers, in both directions. This shift in membership values can be seen as weighted activations to certain concepts — similar to those in the spreading activation model of priming behavior (Collins & Loftus, 1975; Quillian, 1967). While supplying properties is more widely discussed in the Ontological Semantics literature, priming fits better within the context concerning the direct goals of this thesis. For example, let's consider example (3.2a) again, now with an additional element of a "prime" concept — SPOON. The concept of SPOON activates nearby concepts in its hierarchy and memberships of concepts such as CUTLERY (immediate ancestor) or shared category members such as FORK, for the THEME property are increased as a result. If this spreading activation is assumed to be proportional to the path length, $\phi(C_1, C_2)$ between concepts in the ontology, then the memberships to valid fillers of THEME-OF WASH such as any descendent of CLOTH-ITEM will be demoted in comparison. Hence, although FORK and SHIRT are assigned equal membership as the THEME-OF the "unprimed" instance of example (3.2a) since they are both descendants of PHYSICAL-OBJECT, priming by SPOON causes an irregular shift in the sub-hierarchy of PHYSICAL-OBJECT, raising the value of $\mu_{\text{THEME}}(\text{WASH}, \text{FORK})$ in comparison to $\mu_{\text{THEME}}(\text{WASH}, \text{SHIRT})$. A key limitation of this approach is that it is constrained by decomposing an event along its main verb (or noun, when it is a noun-event), and cannot address priming or any other source of membership dynamics when the event is the word that is missing in context.

Regardless of the property-filler dynamics and their modulation by explicit and implicit factors, the event decomposition of cloze-contexts to understand semantic constraints using fuzzy-inferences is taken into qualitative consideration in the next section, which advances towards the primary goals of this thesis.

3.3 Extending Semantic Priming to BERT - Considerations

In humans, semantic priming occurs due to the presence of a lexical associate that affects the speed of response to a stimulus. Analogously, this thesis is interested in learning how BERT's behavior (defined as a change in its output word probability) is affected by a lexical cue present in its input context. In that regard, the methodology in this thesis is primarily centered around the use of cloze contexts, as described in Section 3.1. While cloze contexts can be constructed by randomly omitting words from sentences, in the current study, only a single word is removed from individual cloze sentences at a time. This removal is systematic in nature, to test the behavior of the model when particular "target" words are removed.

In this work, semantic priming in pretrained LMs is defined as an increase in the model's expectation for a target word (or a lack thereof) in a given context in the presence of a semantically related word as compared to an unrelated one. Structurally, priming is simulated by the addition (or prepending) of a "prime" word before a cloze context. Consider the following example:

(3.3) a. I want to become a ___ .

b. *airplane.* I want to become a ___ .

c. *table.* I want to become a ___ .

If the probability of the target word, *pilot* is greater in (3.3b) as compared to that in (3.3c), then it results in an interpretation that the related word (*airplane*) primes BERT more than the unrelated word (*table*) does, for the target *pilot* in the context (3.3a). Such a test ensures that the only difference in BERT's output for the blank position in both cases is due to the swapping of the prime words, allowing one to infer the degree to which BERT relies on single word cues to inform its probability for the target word. Importantly, this setup does not allow for direct comparisons to human word prediction in context—the structure of the tests is adapted for BERT's conventional usage by placing words in context, and thus deviates from standard priming structure.

As established in Chapter 2, the processing of cloze stimuli of the form shown in (3.3a) often involve an effect of the amount of constraint placed on the missing position. Adducing the motivations presented in the aforementioned chapter, the analysis technique developed in this chapter focuses on measuring BERT's sensitivity to individual prime words under varying contextual constraints. For example, consider the following two stimuli where the target word is *key*:

(3.4) a. He lost his ___ yesterday.

b. She opened the door using a ___ .

In (3.4a), the blank position can be any word that denotes the concept which satisfies the semantic role THEME-OF for the event LOSE. Arguably, the blank position is far more constrained in (3.4b), where it can be a word that denotes the concept which satisfies the semantic role INSTRUMENT-OF for the event UNLOCK-DOOR. The subject (*She*) can open the door using a *key, lock-pick,* or perhaps a *screwdriver*, and semantically the sentence is highly constraining towards predicting any word denoting those three concepts or their relatives. Borrowing from Section 2.5.1 and Section 3.2, the constraint of the two sentences can be described by treating the properties THEME-OF and INSTRUMENT-OF as fuzzy sets with their facets representing concept membership. For the event LOSE as described by the sentence in (3.4a), its THEME is relatively unconstrained, and therefore does not have a well-defined DEFAULT. The typical THEME of LOSE can be semantically constrained to any PHYSICAL-OBJECT, which materializes as the SEM facet here. The concept of KEY is a type of PHYSICAL-OBJECT and therefore it receives a membership, $\mu_{\text{THEME}}(\text{LOSE}, \text{KEY}) \leqslant 0.75$. The INSTRUMENT property of the UNLOCK-DOOR event in (3.4b) on the other hand, has a DEFAULT filler of KEY (J. M. Taylor & Raskin, 2011), and therefore receives the highest possible membership, $\mu_{\text{INSTRUMENT}}(\text{UNLOCK-DOOR}, \text{KEY}) = 1$, signifying greater constraint for (3.4b) than for (3.4a). BERT matches this pattern of membership, with its probability for *key* being far higher in (3.4b) than in (3.4a). These analysis is summarized in table 3.1.

Table 3.1. A brief analysis of the semantic constraints imposed by sentences in Example (3.4).

Sentence	*He lost his ___ yesterday.*	*She unlocked the door using a ___ .*
Event Representation	LOST AGENT: HUMAN GENDER: MALE THEME: ???	UNLOCK-DOOR AGENT: HUMAN GENDER: FEMALE INSTRUMENT: ???
Concept Memberships	$\mu_{\text{THEME}}(\text{LOST}, key)$ $<$	$\mu_{\text{INSTRUMENT}}(\text{UNLOCK-DOOR}, key) = 1$
BERT probabilities	$\text{P}_{\text{BERT-large}}(\text{key}) = 0.005$	$\text{P}_{\text{BERT-large}}(\text{key}) = 0.977$

Focusing on how the semantic constraints affects the notion of semantic priming established earlier offers two main advantages. First, it facilitates the exploration of the dynamics of information provided by lexical-cues as primes as compared to words already present in the sentence. This corresponds to exploring how much more information about the target word (*key*) does prepending a related word like *lock* provide in a high-constraint context such as (3.4b), beyond *open* and *door*, which are already present in the sentence, and have association with the target. This can be compared to when the related word, *lock*, is prepended to (3.4a), which imposes fewer constraints on the blank position. Second, it allows comparisons against priming behavior and lexical response observed under high and low constraints in humans. As a caveat, due to the differences in the two setups, this comparison can only be at the behavioral outcome level. The current setup of tests does not allow direct comparisons to human priming under a sentence context, which would ideally require a simulation model of some sort, and is out of scope for this thesis.

3.4 Data Setup and Stimulus Construction

Human priming data used as ground truth for lexical relations are derived from the Semantic Priming Project (SPP) (Hutchison, Balota, Neely, Cortese, Cohen-Shikora, et al., 2013), which is currently the largest resource of its kind. The SPP has previously been used to evaluate word embedding models, such as word2vec (Mikolov, Sutskever, et al., 2013) and GloVe (Pennington et al., 2014), in a number of studies (Auguste, Rey, & Favre, 2017; Ettinger & Linzen, 2016; Mandera, Keuleers, & Brysbaert, 2017). These evaluations have been carried out by measuring the amount of variance in priming response times, derived from the SPP, explained by cosine similarity between word vectors of the target and the prime words as a predictor. The SPP consists of priming data for 768 human subjects across 3322 priming instances which are of the form ($\mathcal{T}$, $\mathcal{R}$, $\mathcal{U}$), where $\mathcal{T}$ is the target word, and $\mathcal{R}$ and $\mathcal{U}$ are the related and unrelated primes, respectively. It also consists of a number of different measures, such as response times (RTs) across different tasks assigned to the participants. In addition, the SPP also provides annotations for the various kinds of lexical relations that exist between the related prime and the target word, these are briefly summarized in Table 3.2. To enable fair comparison, target words that do not occur in BERT's vocabulary, as well as instances where some of the RTs were missing are filtered out, resulting in 92% of the total triples (n = 3058) left for further preprocessing.

Table 3.2. Relations covered by the SPP (Hutchison, Balota, Neely, Cortese, Cohen-Shikora, et al., 2013), their total counts, and example pairs.

Relation	*N*	*Target, Related*
Synonym	418	*anger, fury*
Forward Phrasal Associate	263	*ache, stomach*
Category	164	*bed, sofa*
Antonym	153	*deep, shallow*
Backward Phrasal Associate	151	*cause, effect*
Supraordinate	131	*spaghetti, pasta*
Script	124	*judge, court*
Perceptual property	90	*leaf, tree*
Functional property	73	*bell, ring*
Instrument	35	*bow, arrow*

In addition to the SPP triples, the experiments in this thesis also introduce a context, $\mathcal{C}$, which is a naturally-occurring sentence originally containing the target word, $\mathcal{T}$, now with $\mathcal{T}$ replaced by the "[MASK]" token [1]. The aim of the priming experiments is to test the model's expectation for $\mathcal{T}$ in the masked position when $\mathcal{C}$ is preceded by a related prime, $\mathcal{R}$, and compare it to when it is preceded by an unrelated prime $\mathcal{U}$, denoted as $(\mathcal{R}, \mathcal{C})$ and $(\mathcal{U}, \mathcal{C})$, respectively. $\mathcal{T}$ is embedded in $\mathcal{C}$ in order to better simulate BERT's conventional usage—to predict words in

[1] BERT uses the [MASK] token as its representation of a missing/blank word.

sentence contexts. The contexts $\mathcal{C}$ are chosen to be naturally-occurring sentences, since BERT is trained on well-formed sentences that affect its word level expectation. The target contexts are sampled from the concatenation of the ROCstories Corpus (Mostafazadeh et al., 2016), and the train and test sets used in the "Story Cloze Test" task (Mostafazadeh, Roth, Louis, Chambers, & Allen, 2017), primarily due to the simplistic nature of the sentences contained in those corpora.

Two scenarios are considered for the prime contexts, $\mathcal{R}$ and $\mathcal{U}$: (a) WORD: where the prime word, followed by a period, '.' is prepended to the target context, and (b) SENTENCE: where a neutral context, *"the next word is "* followed by the prime word and a '.', is prepended to the target context. To be compatible with the input format BERT has operated over during training and following previous studies using a similar setup (Ettinger, 2020; Goldberg, 2019; Petroni et al., 2019), the [CLS] and [SEP] tokens are added at the beginning and the end of each stimulus, respectively. Table 3.3 shows full example items from these different settings. The prime words are embedded either as single word or neutral sentence contexts because any naturalistic sentence containing $\mathcal{R}$ would be different from that containing $\mathcal{U}$, thus adding imbalanced noise from the non-prime words. The context $\mathcal{C}$ for the target, by contrast, will remain constant given that the target is constant (for any pair of primes).

Table 3.3. Example Stimuli, with prime contexts in italics. Here, $\mathcal{T}$ = ***pilot***, $\mathcal{R}$ = ***airplane***, and $\mathcal{U}$ = ***table***.

Scenario	Prime Context	Stimulus
WORD	$\mathcal{R}$	[CLS] ***airplane***. I wanted to become a [MASK]. [SEP]
	$\mathcal{U}$	[CLS] ***table***. I wanted to become a [MASK]. [SEP]
SENTENCE	$\mathcal{R}$	[CLS] *The next word is **airplane**.* I wanted to become a [MASK]. [SEP]
	$\mathcal{U}$	[CLS] *The next word is **table**.* I wanted to become a [MASK]. [SEP]

The present study focuses on analyzing BERT's reliance on single-word lexical cues (prime words) to inform its target word probability under the predictive constraints on the [MASK] token. Doing so first requires a quantitative measure of how constraining the un-primed context, $\mathcal{C}$ is. In previous work on analyzing lexical processing in sentence contexts (Federmeier & Kutas, 1999), sentence constraint was measured by using cloze probabilities (by human participants) and performing a median split on the probabilities of the best predicted words to get a binarized set of low and high constraint sentences. Therefore, highly predictable contexts count as a high-constraint contexts, and contexts that are not very predictable count as low-constraint contexts. Similar to Federmeier and Kutas (1999), contextual constraint in this study is measured using the probability of the most expected word by BERT, averaged for the BERT-base and BERT-large models. Mathematically, the constraint of a context ($\mathcal{C}$) is defined as:

$$\text{constraint}(\mathcal{C}) = \max_{x \in \mathcal{V}} \frac{1}{2} \sum_{m \in \{b,l\}} P_m([\textsc{mask}] = x \mid \mathcal{C}), \tag{3.1}$$

where P_m represents the probability distribution for [MASK] in the output of the BERT model, either base (*b*) or large (*l*), and x is a token belonging to BERT's vocabulary, $\mathcal{V}$. The constraint function is thus bounded by $[0, 1]$. In contrast to Federmeier and Kutas (1999), this thesis considers a continuous measure of constraint (as opposed to a binary high/low split) by uniformly binning contexts into n equal bins with increasing constraint scores, to create a graded measure. In principle, n can be any arbitrary number, and larger n corresponds to a more graded measure of the constraint. In the present study, n is set to 10. This allows one to study priming behavior as a function of roughly continuous constraint scores, which was previously not possible with the binarized version. Sentences from the source corpus that contain the target word are grouped into 10 equal bins of a constraint score width of 0.1 each, i.e, a constraint score of 0.38 would be in bin 4. Additionally, as a control, a synthetic and unconstraining target context [2] that is referred to as "neutral" is also used, these neutral context items appear as follows:

[CLS] the last word of this sentence is [MASK]. [SEP]

Intuitively, the neutral context contains no information about what [MASK] can be—it can be filled by any word in BERT's vocabulary. Therefore, it provides the least constraint on the [MASK] token. Using Equation 3.1, the empirical constraint for this context was found to be $\approx$ 0.02, thus confirming this intuition.

[2] The choice of neutral context follows Schwanenflugel and LaCount (1988).

To make robust conclusions about the effect of constraint, only the triples that have at least one target context in each of the 10 bins were sampled. Polysemy issues were encountered for 72 target words, where the sense of the target in the originally sampled $\mathcal{C}$ did not fit the lexical relation with the primes—these were manually corrected by re-selecting appropriate contexts from the corpus, however, this issue could not be resolved for 28 items, which were discarded. This further reduces the number of unique triples to 2112 (69% of the valid instances), with each triple being associated with 11 (10 bins and a neutral context) stimuli. Figure 3.3 displays the average constraint score within each constraint bin.

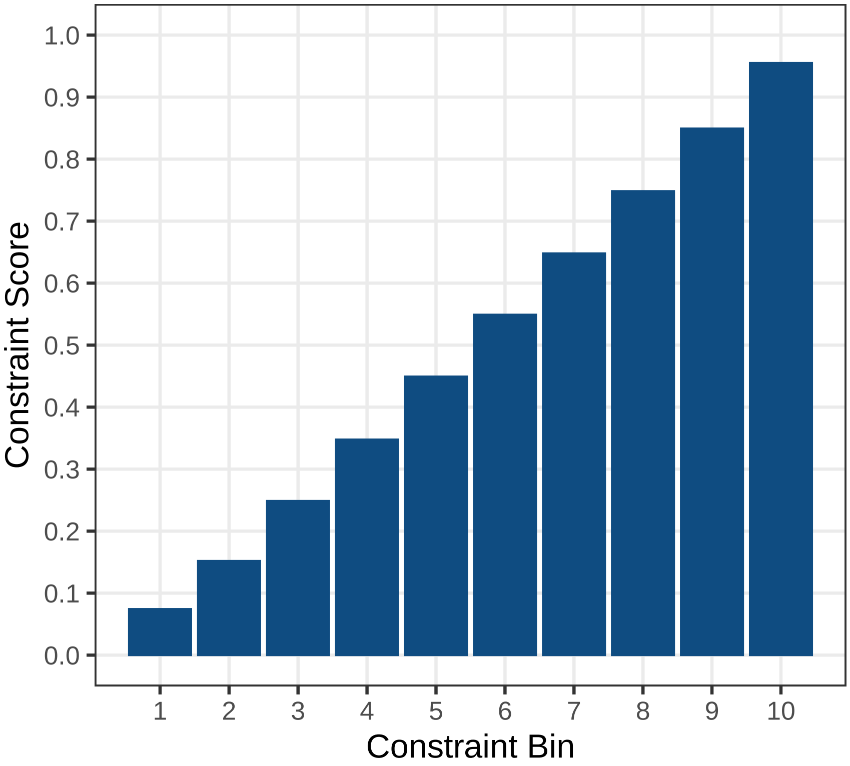


Figure 3.3. Average constraint score within each constraint bin.

To further understand the numerical properties of the notion of constraint adopted in this thesis, constraint scores were compared against the entropy of BERT's probability distribution for the [MASK] token, over its vocabulary. Entropy (Shannon, 1948) is an information theoretic measure that quantifies the average amount of information or uncertainty encoded in a random variable's possible outcomes. For a discrete random variable X, with possible outcomes

$\{x_1, x_2, ..., x_n\}$, that occur with probabilities $\{p(x_1), p(x_2), ..., p(x_n)\}$, its entropy $H(X)$ is mathematically defined as:

$$H(X) = -\sum_{i}^{n} p(x_i) \log p(x_i). \tag{3.2}$$

In the context of the research goals of this thesis, the entropy of BERT's output distribution for a cloze-context can be considered to denote a form of constraint on the [MASK] token, i.e. in terms of the uncertainty about the filler for [MASK]. Here, low constraint contexts would represent high uncertainty, and therefore a high entropy value. Similarly, lower entropy values would be observed for high constraint contexts, where the model is To compare the proposed constraint scores with this new entropy-based notion of constraint, the following quantity is computed for every context ($\mathcal{C}$) used in subsequent experiments, where the entropies of the outputs for both BERT models (*b, l*) are averaged:

$$H_{\text{constraint}}(\mathcal{C}) = -\frac{1}{2} \sum_{m \in \{b,l\}} \sum_{x \in \mathcal{V}} P_m(x \mid \mathcal{C}) \log P_m(x \mid \mathcal{C}) \tag{3.3}$$

The Pearson correlation between constraint$(\mathcal{C})$ and $H_{\text{constraint}}(\mathcal{C})$ was found to be -0.89, indicating strong empirical relationship between constraint measured as the probability of the best completion and entropy of the predicted distribution. This relationship is also represented by Figure 3.4.

3.5 Measuring Priming in BERT

For a probabilistic model of word prediction in context, such as BERT, its expectation for a word w in context c can be measured using the model's **surprisal**, or the negative log likelihood of w, given c or hidden state representation h_c (in case of a neural network-based model):

$$\textit{Surp}(w \mid c) = -\log_2 P(w \mid h_c). \tag{3.4}$$

The surprisal of a model represents the amount of "surprise" in encountering a particular word in the given context. Surprisal is an effective linking hypothesis between language model probabilities and measures of human language processing. For instance, surprisal derived from n-gram and

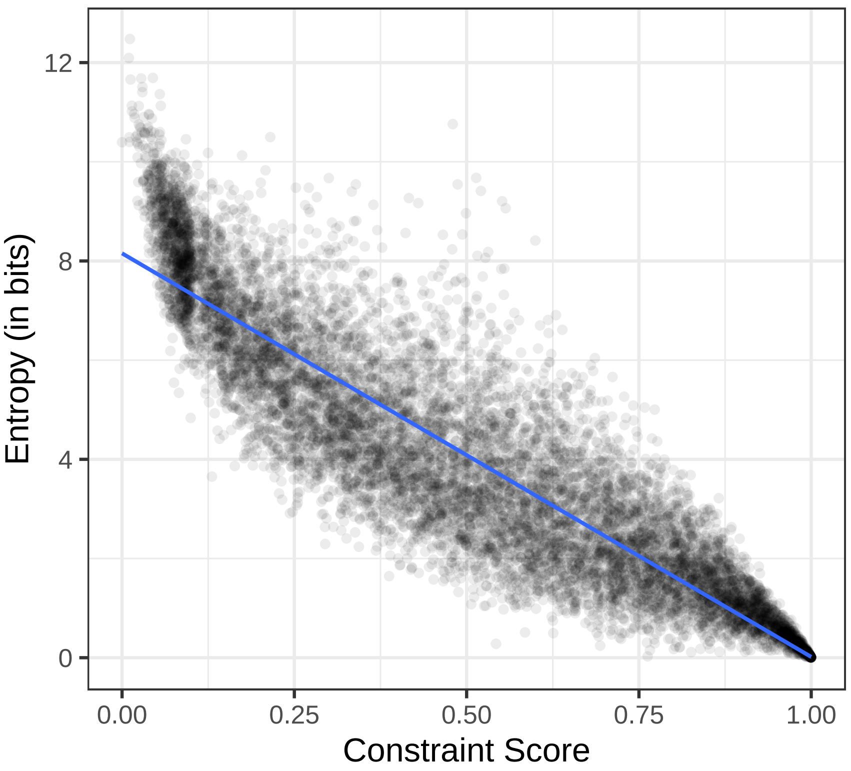


Figure 3.4. Average entropies of contexts (Equation 3.3) plotted against their constraint scores computed using Equation 3.1.

RNN based LMs was shown to be a significant predictor of self-paced reading times, a measure of cognitive load incurred during sentence comprehension in humans (Hale, 2001; Levy, 2008; Smith & Levy, 2013). Word surprisals have also found to be predictive of the amplitude of the N400 event related potential (ERP) (Kutas & Federmeier, 2011; Kutas & Hillyard, 1980).

The stimuli in this thesis are sentences prepended by prime contexts (word or sentence) that differ minimally in the prime word ($\mathcal{R}$ or $\mathcal{U}$), thus keeping the main sentence context $\mathcal{C}$ constant. This nullifies the effect of syntactic/structural differences on BERT's surprisal for $\mathcal{T}$. Therefore, measuring the difference in BERT's surprisals for the target word $\mathcal{T}$ between the two minimal pairs quantifies the degree to which the model gets influenced by one isolated prime word over the other. This can be considered analogous to how the difference in response times corresponds to the strength of lexical association between the prime and target words, as used by the human language processing faculty. Following the definition of priming in the context of word prediction models mentioned in Section 3.3, priming in BERT can be quantified by this difference between surprisals of $\mathcal{T}$ in the unrelated and related contexts. This quantity is referred to as "**facilitation**", $\mathbb{F}$, and is mathematically described as:

$$\mathbb{F} = \textit{Surp}(\mathcal{T} \mid \mathcal{U}, \mathcal{C}) - \textit{Surp}(\mathcal{T} \mid \mathcal{R}, \mathcal{C}). \tag{3.5}$$

If $\mathcal{R}$ primes BERT in predicting $\mathcal{T}$ in place of [MASK], more than $\mathcal{U}$ does, then BERT should show less "surprise"—i.e., produce lower probability—in its expectation for $\mathcal{T}$ in the context $(\mathcal{R}, \mathcal{C})$, than in $(\mathcal{U}, \mathcal{C})$. In such cases, $\mathbb{F}$ will be positive.

Measuring the number of cases where $\mathbb{F}$ is positive allows one to discern general patterns of a model's sensitivity to lexical associations between the target and the related prime word. For n samples of minimal-pair priming stimuli as described before, the proportion of cases which show priming by the related word is measured by:

$$\textit{Priming} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\mathbb{F}_i > 0}, \tag{3.6}$$

where $\mathbb{1}$ is an indicator function which returns 1 when $\mathbb{F}_i > 0$ and 0 otherwise. The cases where priming by related words is observed are termed as "primed instances."

In summary, the quantities that measure priming in BERT stem from the model's probabilities of target words in context, and denote the level of sensitivities that the model shows to isolated lexical cues (related primes and unrelated primes). The first quantity, facilitation ($\mathbb{F}$), denotes the level of surprise shown by BERT in encountering the target word in presence of an unrelated prime relative to that in the presence of a related prime, and the second quantity denotes the aggregate level of facilitation across a range of stimuli. The next Section describes the tests designed to answer the research questions presented in Section 1.2, and how the quantities described in this Section are utilized to form conclusions about BERT's priming behavior.

3.6 Description of Empirical Tests

The tests described in this thesis target the extent to which BERT shows sensitivities towards isolated lexical cues that appear in cloze stimuli as "primes." These sensitivities are especially observed under varying levels of contextual constraint, a measure grounded in psycholinguistics literature that indicates the degree to which a cloze-context is predictable. Patterns of sensitivities and contextual constraints are analyzed using two empirical tests.

The first test investigates the relation between BERT's overall priming behavior and the level of contextual constraint imposed on missing tokens in the cloze-stimuli. It focuses on the exact nature of the interaction between the strength of lexical relation between words during word prediction and the constraint imposed by the context. This test is conducted by measuring BERT's facilitation as a function of the constraint scores established in Equation 3.1. Specifically, constraint scores are used as fixed effects in a linear mixed-effects model (Baayen, Davidson, & Bates, 2008) with random intercepts for target words. The relation between facilitation and constraint is quantified by the estimated coefficient for the constraint score in the model — i.e., a negative estimate would suggest a negative relationship. To establish statistical significance, this model is compared to a baseline model without the fixed effect component of the constraint using a likelihood ratio test. This test is also augmented by the analysis of the relation between the proportion of "primed instances" and contextual constraint scores, revealing aggregated patterns where BERT positively uses the presence of a related word in context to inform its word probabilities.

The second test performs a finer-grain investigation into the consistency of the patterns discovered in the results of the first test across various lexical relations represented in the SPP dataset. This test uses the same general methodology of assessing the relation between facilitation and constraint scores, but specifically measures the interaction for stimuli representing various lexical relations between words, revealing the extent to which BERT is attuned towards them. Specifically, the test focuses its analysis on the stimuli belonging to the top-10 most frequent lexical relations in the SPP dataset: *synonym, forward phrasal associate, category, antonym, backward phrasal associate, supraordinate, script, perceptual property, functional property,* and *instrument*.

Apart from the two tests which form the core set of analyses, the following chapter presents additional insights into the differences between priming by a word as opposed to a sentence context; between BERT-base and BERT-large, the two models investigated in this thesis, and more importantly, a *post hoc* analysis to further explore potentially anomalous patterns.

CHAPTER 4. RESULTS AND DISCUSSION

This chapter presents the results of the empirical tests proposed in this thesis (and described in section 3.6), and at the end, summarizes them in the broader context of existing literature.

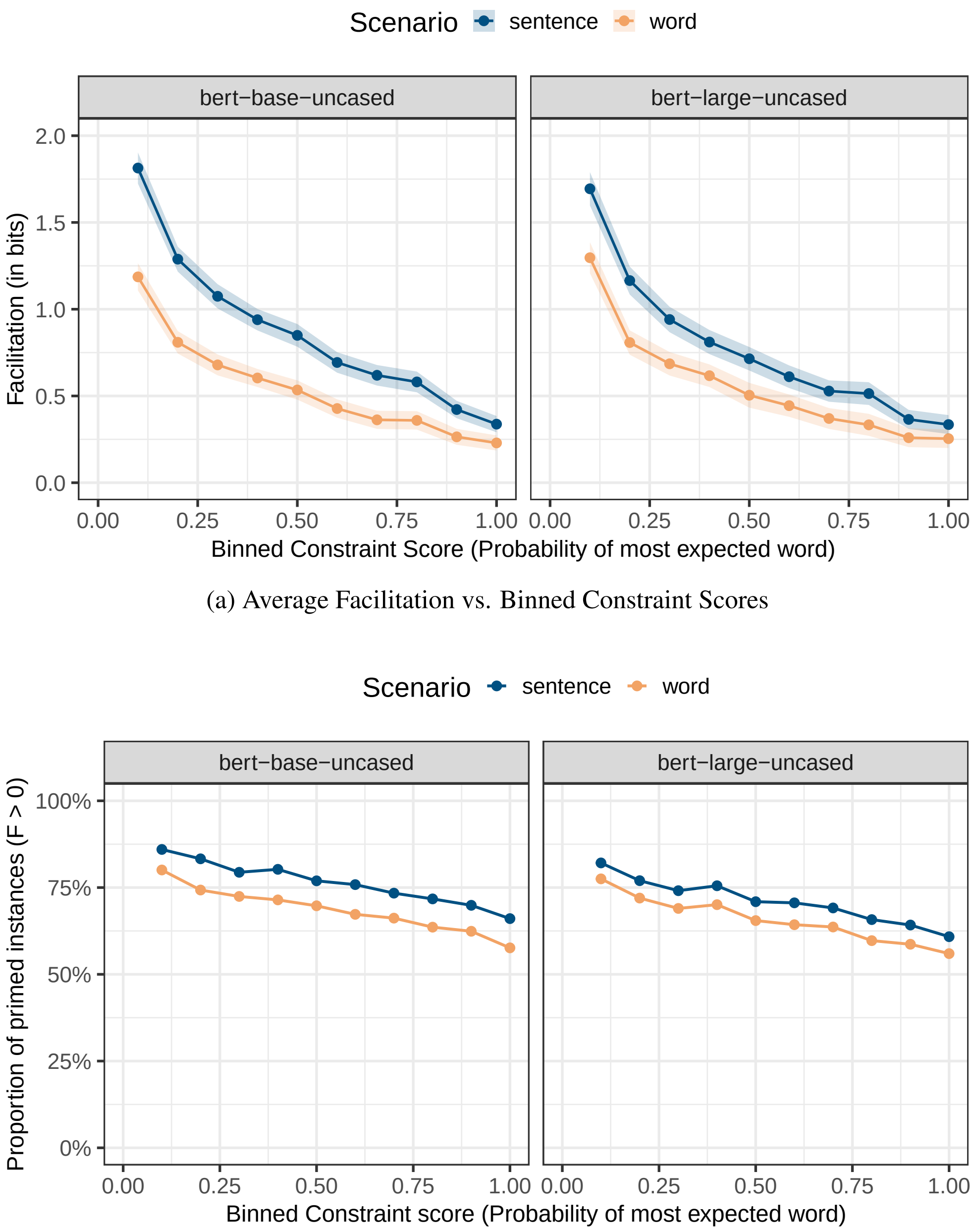
4.1 Relationship between Facilitation in BERT and Contextual Constraint

The aim of the first experiment was to investigate the relationship between BERT's lexical sensitivity to prime words, quantified by the facilitation, and the predictability of the cloze context, quantified by constraint scores. This relationship between BERT's facilitation and the input's contextual constraints is depicted in Figure 4.1a, which shows the average facilitation values for BERT-base and BERT-large across the SENTENCE and WORD scenarios, plotted against binned constraint scores. Additionally, Figure 4.1b shows the total proportion of instances where the two BERT models showed priming behavior ($\mathbb{F} > 0$) across both scenarios. Both measures for the neutral context, "[CLS] *the last word of this sentence is* [MASK]. [SEP]" are shown separately in Table 4.1 since the neutral context was synthetically designed, and deviates from the samples extracted from naturally occurring text for which constraint scores were calculated.

Table 4.1. Average Facilitation (with 95% standard error) and percentage of primed instances for Neutral contexts.

Model	Scenario	$\mathbb{F} \pm$ 95% S.E.	Percent of Primed Instances
BERT-base	SENTENCE	4.10 ± 0.16	88.26%
	WORD	2.69 ± 0.12	85.23%
BERT-large	SENTENCE	5.12 ± 0.16	91.95%
	WORD	5.14 ± 0.16	91.29%

On average, both BERT models show positive facilitation values across all constraint items and prime-context scenarios ($p < .001$ in all cases, as measured by one-sample t-tests). From Figure 4.1a, priming effects in BERT models decreases as the constraint of the cloze-context on [MASK] increases and this is further evidenced by the negative relationship between facilitation and constraint scores as measured by the likelihood-ratio test between a model that estimated facilitation



(a) Average Facilitation vs. Binned Constraint Scores

(b) Proportion of primed instances vs. Binned Constraint Scores.

Figure 4.1. Average facilitation *(a)* and proportion of primed instances, i.e., $\mathbb{F} > 0$ *(b)* vs. binned constraint score. Error bands in *(a)* represent 95% confidence intervals.

using constraint as fixed effects together with random intercepts for target words and a baseline model that only included the random intercepts. This is shown in Table 4.2, where the coefficient of the constraint is estimated to be negative for both models across both scenarios. This indicates that the information provided by the related prime word (relative to the unrelated one)

Table 4.2. Relation between facilitation and constraint (quantified by $\beta_{\text{constraint}}$) indicated by a linear mixed-effects model. Significance calculated using Likelihood Ratio Tests.

Model	Scenario	$\beta_{\text{constraint}}$	$\chi^2(1)$	p-value
BERT-base	SENTENCE	-2.51	3842.76	$< .001$
	WORD	-1.65	2506.84	$< .001$
BERT-large	SENTENCE	-2.90	3995.26	$< .001$
	WORD	-2.76	3495.80	$< .001$

is increasingly outweighed by the information provided by the predictive constraints as the level of constraint increases. This becomes particularly apparent upon comparing facilitation values of naturally occuring cloze contexts shown in Figure 4.1a and those of neutral contexts, shown in Table 4.1. Neutral contexts, where BERT receives almost no context information from non-prime words, show substantially larger facilitation than their naturally occurring counterparts (for instance, 5.12 bits of facilitation is observed in neutral contexts for BERT-large for the SENTENCE prime-context scenario as opposed to 1.69 bits in the next lowest constraint score of 1), suggesting that BERT almost completely relies on the prime words to inform the target word's probability in context. On comparing settings with and without sentence context for the prime word, BERT consistently shows greater facilitation effects when the prime context is a sentence rather than a single word, across every constraint bin ($p < .001$), with the exception of BERT-large for neutral contexts, where the magnitudes of the facilitation are the largest (as shown in Table 4.1), but not significantly different between sentence and word prime contexts ($t(2111) = -0.3402, p = 0.6331$). Overall, this suggests that sentence contexts confer more information that the BERT model is able to utilize to inform its word probabilities.

4.2 Facilitation and Constraint Scores across Lexical Relations

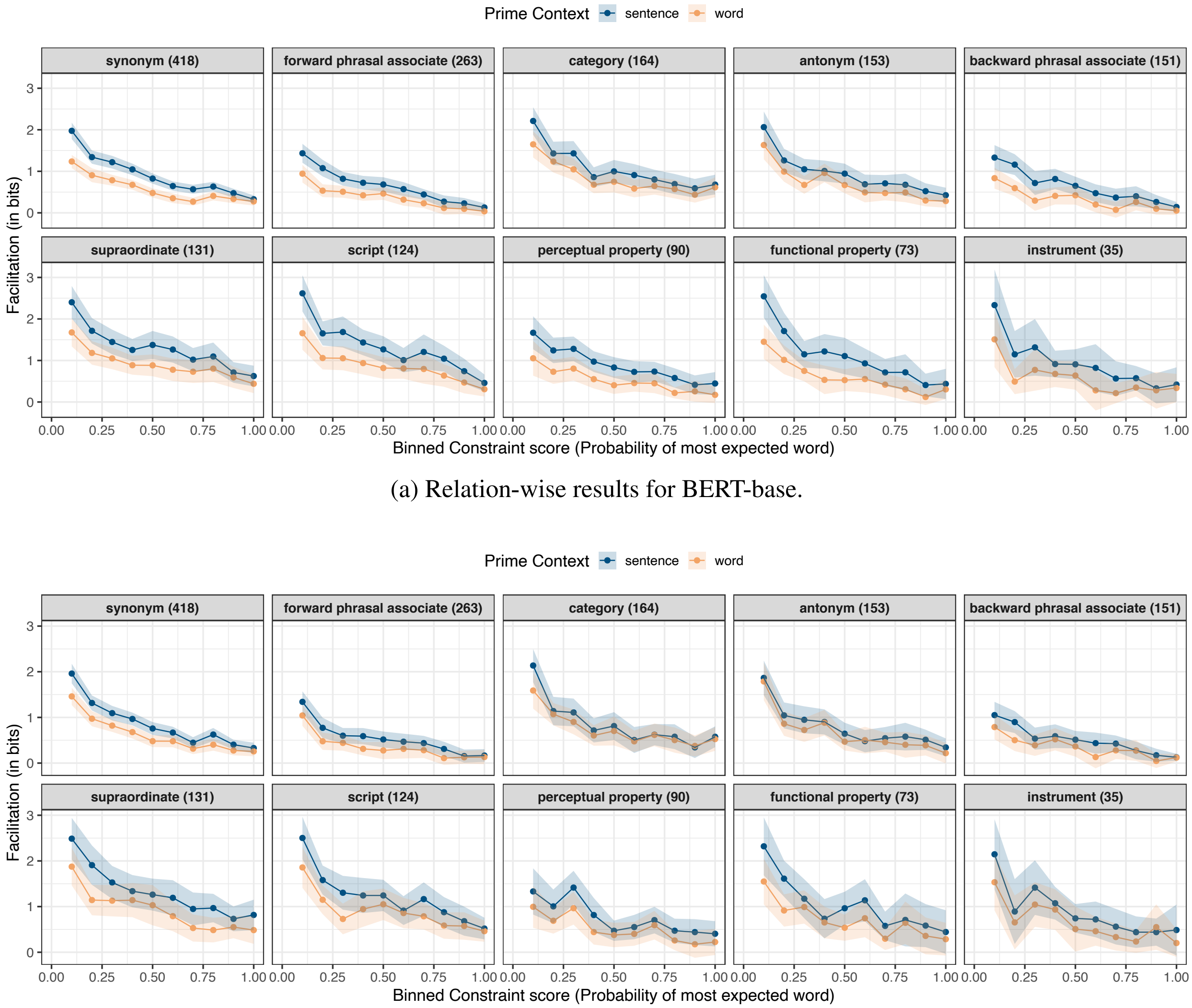
The second experiment follows from the first one and investigates the interaction between BERT's facilitation values and the constraint scores of the cloze-contexts, but now with the fine-grained lens of specific lexical relations between the target and related words. Figure 4.2 depicts average facilitation scores plotted against binned constraint scores across the SENTENCE and WORD scenar-

Table 4.3. Average Facilitation (with 95% standard error) and percentage of primed instances for Neutral contexts.

Relation	N	$\mathbb{F} \pm$ 95% S.E.		Percent of Primed Instances	
		SENTENCE	WORD	SENTENCE	WORD
Synonym	418	5.62 ± 0.23	4.89 ± 0.24	94.50%	93.06%
Forward Phrasal Associate	263	2.97 ± 0.27	2.61 ± 0.26	84.60%	81.75%
Category	164	6.90 ± 0.41	5.46 ± 0.40	96.65%	95.12%
Antonym	153	8.12 ± 0.40	5.82 ± 0.39	97.71%	95.75%
Backward Phrasal Associate	151	3.40 ± 0.38	2.99 ± 0.36	87.42%	84.11%
Supraordinate	131	4.81 ± 0.45	4.30 ± 0.45	90.46%	90.84%
Script	124	4.51 ± 0.45	4.02 ± 0.47	88.71%	87.90%
Perceptual Property	90	3.74 ± 0.50	3.15 ± 0.52	88.33%	82.78%
Functional Property	73	3.75 ± 0.53	3.79 ± 0.54	88.36%	89.04%
Instrument	35	4.00 ± 0.80	3.26 ± 0.77	87.14%	84.29%

ios for BERT-base (Figure 4.2a) and BERT-large (Figure 4.2b). As in the previous section, results for the neutral context are summarized separately in Table 4.3. In addition, Table 4.4 summarizes the results of performing a comparison between the constraint vs. facilitation linear mixed-effects model, and the constraint-less baseline linear mixed-effects model for each of the relations, across both models and prime-context scenarios.

Figure 4.2 and Table 4.3 suggest a somewhat consistent patterning of the average facilitation with the constraint of a cloze-context — positive facilitation scores, on average, for all relations across every constraint bin ($p < .001$ in all cases). Again, neutral contexts show greater facilitation across relations, indicating a consistently similar usage of isolated lexical cue information by BERT, in estimating target word probabilities. Interestingly, the plots shown in Figure 4.2 show somewhat inconsistent patterns local to certain constraint-score regions, where higher constraint scores show more priming effects than do preceding lower constraint scores (for instance, the facilitation increases when going from 0.7 constraint score to 0.8 in stimuli within the synonym relation for both models). This locally inconsistent behavior is more pronounced in relations with low sample sizes, which incidentally also show wider confidence intervals, indicating the need for more samples in order to form stronger conclusions. Overall, the results from the likelihood ratio tests indicated in Table 4.4 appear resistant to the locally-inconsistent patterns within certain constraint scores—showing consistently and significantly negative relations between facilitation and



(a) Relation-wise results for BERT-base.

(b) Relation-wise results for BERT-large.

Figure 4.2. Facilitation effects across top-10 lexical relations for (a) BERT-base and (b) BERT-large. Error bands represent 95% confidence intervals.

constraint scores across all relations and prime-constraint scenarios, for both models ($p < .001$ in every case). Specifically, among lexical relations with a considerably high sample size synonymy, category, and antonymy relations show the most pronounced differences, with both BERT models showing considerably larger facilitation in the neutral context than for other relations (Table 4.3). This suggests that BERT's word predictions in context may be more strongly attuned to relations of synonymy, category membership, and antonymy than to other lexical relations.

Table 4.4. Relation between facilitation and constraint (quantified by $\beta_{\text{constraint}}$) indicated by a linear mixed-effects model, across top-10 relations in SPP (Hutchison, Balota, Neely, Cortese, Cohen-Shikora, et al., 2013). Significance calculated using Likelihood Ratio Tests. $p < .001$ in all cases.

Model	Relation	N	$\chi^2(1)$		$\beta_{\text{constraint}}$	
			SENTENCE	WORD	SENTENCE	WORD
BERT-base	Synonym	418	1172.58	720.12	-3.01	-1.97
	Forward Phrasal Associate	263	374.75	242.30	-1.84	-1.24
	Category	164	358.10	251.01	-3.33	-2.18
	Antonym	153	428.46	354.35	-4.05	-2.69
	Backward Phrasal Associate	151	216.22	126.50	-1.92	-1.20
	Supraordinate	131	223.14	138.85	-2.52	-1.68
	Script	124	225.85	133.26	-2.59	-1.67
	Perceptual Property	90	164.92	93.95	-2.08	-1.35
	Functional Property	73	144.26	111.27	-2.49	-1.65
	Instrument	35	69.28	36.38	-2.41	-1.39
BERT-large	Synonym	418	1147.61	987.75	-3.52	-3.44
	Forward Phrasal Associate	263	399.17	306.24	-2.06	-1.98
	Category	164	396.64	326.75	-4.08	-3.54
	Antonym	153	416.00	365.34	-4.43	-3.74
	Backward Phrasal Associate	151	243.55	202.71	-2.22	-2.17
	Supraordinate	131	220.71	220.05	-3.05	-3.13
	Script	124	227.21	195.60	-2.96	-2.86
	Perceptual Property	90	145.26	129.46	-2.33	-2.33
	Functional Property	73	117.63	138.61	-2.65	-2.87
	Instrument	35	74.37	67.80	-2.77	-2.55

4.3 On Primes and Distractors: a *post hoc* examination

The preceding results show a decrease in number of primed instances as contextual constraint increases. This means that as the constraint imposed by the context increases, there are more instances in which the probability of the target word in presence of the related word is *less* than that in presence of an unrelated word. For example, the first row of Table 4.5 shows an instance for a target, *bacon*, with a constraint score of 0.89 (i.e., the 9th bin). Contrary to priming patterns observed in low-constraint contexts, the probability of *bacon* is quite low when BERT is primed by *pork*, and very high when the unrelated word, *meteorite*, is the prime. Here, the related prime acts

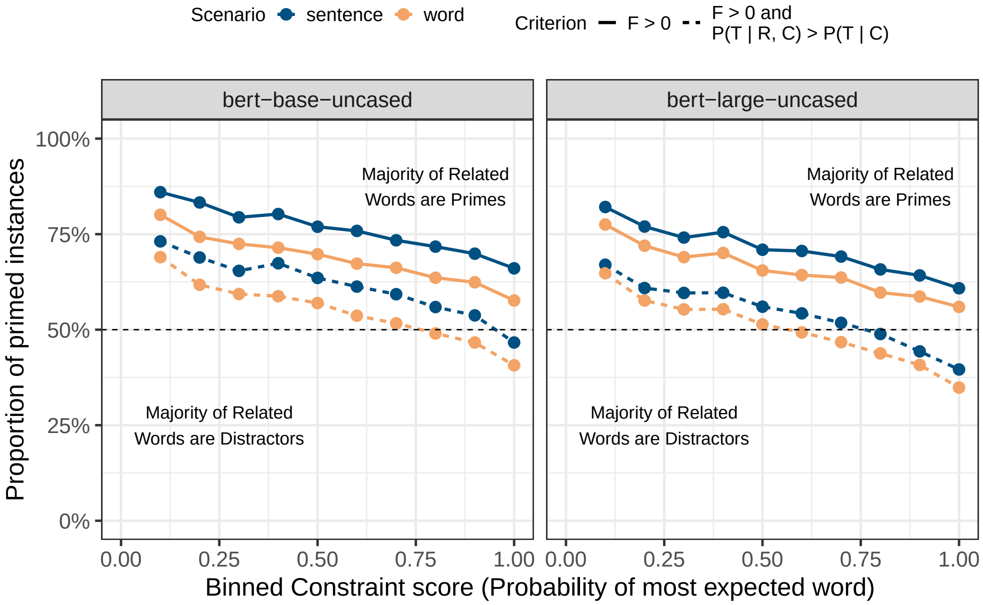


Figure 4.3. Proportion of primed instances under more (dashed) and less (solid) stringent priming criteria.

as a distractor,[1] similar to the mispriming phenomenon reported in Kassner and Schütze (2020). Upon further investigation, the probability of the target word in presence of the related word is in fact also observed to be lower than that in an un-primed context, i.e., $P(\mathcal{T} \mid \mathcal{R}, \mathcal{C}) < P(\mathcal{T} \mid \mathcal{C})$. In such cases, the related word "distracts" rather than "aiding" BERT from reliably getting primed, thereby reducing the probability of the target. To account for such cases, the criterion of what counts as an "primed" instance can be made more stringent — where primed instances show positive facilitation ($\mathbb{F} > 0$) *and* if the presence of the related word increases the probability of the target over that in the un-primed instance ($P(\mathcal{T} \mid \mathcal{R}, \mathcal{C}) > P(\mathcal{T} \mid \mathcal{C})$). These changes are reflected in Figure 4.3 which shows a plots similar to Figure 4.1b, now with the more stringent criterion. In Figure 4.3, the proportion of facilitatory instances is now substantially lower with this more robust notion of priming, but it follows the same pattern observed when only facilitation score was considered. At higher constraint scores, the proportions fall under 50%, giving us thresholds beyond which BERT shows more "distraction" from related prime words than facilitation. For

[1] it is referred to as a distractor rather than a misprime since the target word is not the absolute correct completion for our contexts, as they are not factual like in Kassner and Schütze (2020).

example, starting at the 8th constraint bin, BERT-base shows priming only for 49% or fewer cases in the WORD prime context.

Table 4.5. Example high constraint instances that show "distraction" rather than priming in BERT-large.

Target (Constraint)	$(\mathcal{R},\mathcal{U})$ Context	Top 5 Predicted Words (BERT-large probability)	
		Primed by Related	Primed by Unrelated
bacon (0.89)	(pork/meteorite). she cooked up some eggs, [MASK], and toast.	*eggs (0.20), potatoes (0.04), tea (0.04), pancakes (0.04), cheese (0.03)*	*bacon (0.78), sausage (0.06), ham (0.03), pancakes (0.02) toast (0.02)*
painting (0.75)	(drawing/champagne). dana was a young artist who spent many hours a day [MASK].	*drawing (0.88), painting (0.10), studying (<0.01), writing (<0.01), practicing (<0.01)*	*painting (0.79), drawing (0.06), working (0.03), studying (0.03), teaching (0.01)*

Specific instances of model predictions are qualitatively examined to shed further light on the factors that contribute to BERT's distraction (as opposed to priming) effects. Table 4.5 shows two examples in which such distraction patterns are observed in BERT. In the example with *painting* as the target, BERT's behavior is akin to that discussed in Kassner and Schütze (2020). Here, the presence of a distractor (*drawing*), one that fits as a completion in the [MASK] position, leads BERT to predict the distractor with greater probability than the target (*painting)*. However, the example with *bacon* as the target shows a different kind of distraction: *pork* cannot replace *bacon* here as well as *drawing* can replace *painting* in the previous example, but *bacon* is still demoted in the probability distribution in favor of other foods related to *pork*. By contrast, in both examples the unrelated primes resemble "random misprimes" in Kassner and Schütze (2020): BERT isn't distracted by them—likely due to their degraded relevance to the context—and still predicts the target as the best completion.

4.4 General Discussion

The experiments above show that when using word pairs informed by human semantic priming, the BERT model is reliably sensitive to individual lexical cues in its context—*if* the context is minimally constraining, such that there is little predictive information beyond that lexical cue. As the predictive constraint applied by the context increases, BERT's level of sensitivity to a given lexical cue decreases. These results suggest that BERT uses lexical cues as needed: when informative sentence cues are available, single lexical items are of less value, and so they exert less influence on BERT's expectations for a masked word.

The examination of patterns across different types of lexical relations suggests that this general effect of constraint holds across several relation types, but synonym, category, and antonym relations elicit larger lexical sensitivities in BERT, as compared to other relations (when the context is unconstraining). This suggests that BERT has identified these relations—or the particular words that share these relations—to be reliably predictive. Human brains show facilitation to priming items, likely due to predictive mechanisms sensitive to co-occurrence or feature overlap (Hutchison, 2003). BERT, being a strong predictive model for language, can reasonably be expected to pick up on these patterns too, thus showing strong sensitivities in presence of informative prime words. The informativeness of prime words is likely amplified greatly in the absence of highly related words in low constraint sentences and hence BERT is reliably primed with greater effects in such instances. BERT's priming behavior can therefore be strongly linked to how well it captures co-occurrence statistics about words, and how this leads to it forming higher-order relational associations that inform its relative lexical sensitivities.

While we see that these priming-based lexical relations can have facilitatory effects on BERT's word predictions when the context is otherwise unconstraining, we see conversely that when the context *is* constraining, prime words can actually have a "distractor" effect—actively demoting the target word in the probability distribution. This finding builds on recent evidence of BERT's sensitivity to such distractions when predicting completions to factual queries (Kassner & Schütze, 2020). The analyses presented above show that the nature of this distraction depends critically on the interaction of contextual constraint and the strength of the lexical relation: when the context is unconstraining, the probability of a word is likely to be promoted by a related lexical item more than by an unrelated lexical item. If the context is constraining, a related lexical item may demote the probability of a target word in the predicted distribution, while an unrelated word is likely to have less impact. Judging from the qualitative analyses, it can be speculated that as a context's constraint increases, so does BERT's expectation for an item similar/related to the target – but once that expectation is met by the related prime, BERT no longer expects another similar word in the target position. If this is indeed a generalizable inference, then for any model, its “distraction” effects would depend on its expectation based on context, or sensitivity towards redundancy. This opens up new avenues for future work investigating different families of language models.

The distraction effects observed in BERT are likely to be avoided in the ontological semantic account of priming, which can be seen as a relative change in the membership of a concept to a property’s fillers as briefly described in Section 3.2. Taking the example of the first row in Table 4.5, the sentence is primarily decomposed along the event COOK, and the [MASK] token occupies a position that appears to be a candidate for the THEME of the event, represented as the following:

```
COOK
    AGENT: HUMAN
    THEME: ??
        DEFAULT: UNCOOKED-FOOD
        SEM: FOOD-ITEM
```

Figure 4.4. Event representation for the cloze-context: *“She cooked up some eggs,* [MASK]*, and toast.”*

When primed by PORK, concepts that have a close relation to it (and are subsumed by FOOD-ITEM or UNCOOKED-FOOD) are activated — i.e., they experience an increase in their fuzzy membership values for the property of interest, the magnitudes of which depend on their closeness to the prime. Hence, BACON shows reliably stronger membership to THEME-OF COOK. When the prime is an unrelated concept that is not subsumed by any of the candidate fillers, such as METEORITE, the membership of BACON remains low due to its distance from the prime in the ontological hierarchy. However, due to its limitations as discussed in section 3.2, the Fuzzy-OST method is unable to account for priming effects in cases resembling the example in the second row of Table 4.5, where missing words denote events themselves.

The effectiveness of human priming pairs in influencing BERT’s lexical sensitivities, as well as the impact of contextual constraint on BERT’s use of lexical context cues suggest possible parallels with mechanisms in human language processing. Not only do humans show priming with the same word pairs that this thesis shows to impact BERT’s predictions here, but like BERT, humans also show more limited semantic priming in constraining contexts, and wider scope of priming in low-constraint contexts (Schwanenflugel & LaCount, 1988; Schwanenflugel & Shoben, 1985). This suggests that the mechanisms that dictate BERT's lexical sensitivity may be optimized in a manner—or at least to an outcome—comparable to those underlying priming effects in humans.

In practical terms, the results of these analyses highlight the importance of contextual constraint in the dynamics of word prediction and information usage in the BERT model. Future work studying these dynamics should be mindful of this fact, as any observed prediction dynamics may change with the predictive-ness of the context. This further emphasizes parallels with the study of human processing, as the predictive constraint of context has long been an important consideration and instrument in studying human sentence processing (Federmeier & Kutas, 1999; Schwanenflugel, 1991; Schwanenflugel & LaCount, 1988). This thesis' findings show a similarly important role played by the amount of constraint imposed on a masked word during word probability estimation, which can lead to substantially different outcomes in behavioral analysis of pre-trained models.

CHAPTER 5. CONCLUSION AND FUTURE WORK

5.1 Overview

Models that are pretrained by estimating word probabilities in context have become ubiquitous in natural language processing. By-product representations that are learned as part of the pretraining process have substantively improved the state of the art in several high-level NLP tasks. However, the question of what linguistic properties pretraining confers upon models is by and large a significant research pursuit. This thesis focuses specifically on the behavioral properties of models to infer the degree to which they inform their word probabilities using isolated lexical cues in context. To this end, this thesis presented a case study analyzing the pre-trained BERT model with tests informed by semantic priming. Priming in this thesis is defined by a direct analogy from humans. Just as humans get primed to react or respond faster to stimuli in presence of a related as opposed to an unrelated "prime", this thesis proposed to evaluate whether BERT more strongly forms its expectation for a word in a sentence context in the presence of a related as opposed to an unrelated lexical cue. The experiments in this thesis are based on word pairs with clear, cognitively-based lexical relationships for which one can explore fine-grained relation differences. Further, this thesis integrated perspectives from cognitive science and psycholinguistics and empirically studied BERT's priming dynamics based on how it was modulated by contextual constraint. Apart from empirically defining the constraint of a sentence using metrics grounded in experimental and behavioral studies involving humans, this thesis blended in the paradigm of ontological semantics to qualitatively understand the semantic constraints of sentence contexts.

Overall, this work found BERT to show "priming," predicting a word with greater probability when the context includes a related word versus an unrelated one. This effect decreased as the amount of information provided by the context increased — suggesting parallels with sentence based human priming experiments where low constraint (less predictable, high entropy) sentences showed a greater scope of facilitation as opposed to high-constraint which only elicited facilitation in high constraint sentences. Follow-up analysis showed BERT to be increasingly distracted by related prime words as the context became more informative, assigning *lower* probabilities to related words. The findings of this thesis establish the importance of considering contextual constraint effects when studying word prediction in word prediction models, and highlight possible parallels with human processing.

5.2 Recommendations for Future Work

The following subsections sketch out recommendations for future work that can augment the empirical contributions of this thesis.

5.2.1 Broad-coverage analyses of priming in different language model strategies

This thesis has developed a methodological framework by using lexical stimuli that cause priming in humans to investigate analogously similar effects within a word prediction model — BERT. This framework can currently cover other word prediction models similar to the masked language modelling framework of BERT, but differ from BERT and each other mainly in terms of parameter counts, such as RoBERTa (Liu et al., 2019), ELECTRA (Clark, Luong, Le, & Manning, 2020), *inter alia*, without any changes made in the stimuli setup or construction. It would be worthwhile to extend this framework to language models that process language incrementally, formulating lexical and syntactic hypotheses as they encounter new words/tokens — i.e., models such as RNNs and left-to-right transformer based models such as GPT2 (Radford et al., 2019). However, this would require a drastic change in the stimuli needed as these models come with the restriction of processing context only from the left side of the cloze-position. Furthermore, there can be a radical mismatch between vocabularies of various models (for instance, using GPT-2 with the current set of stimuli would reduce the total sample size by 34%), which can be a non-trivial challenge to circumvent.

Regardless, comparing priming behavior across a broader set word prediction models will reveal novel and general insights about the roles played by parameter counts, training objective, and underlying architectures in influencing the interaction between context predictability and target word facilitation. Such an analysis can be made more robust in the context of forming firm conclusions about neural network architectures by investigating priming behavior in non-neural n-gram LMs, as a baseline analysis.

5.2.2 Priming/adaptation effects using complex semantic units

The general framework of measuring sensitivities in word prediction models can be extended to include more complex semantic units such as phrases or sentences, where two contrasting events are used as "sentence primes," followed by a generic incomplete context with the missing word occupying the rightmost position. Such a setup is similar to the CPRAG-102 dataset, compiled by Ettinger (2020), and will likely target more complex interactions between syntactic, lexical, compositional, and pragmatic elements within the model, which come into play in processing a longer context as input. Adaptation to more complex set of inputs facilitates more broader analyses of the underlying dynamics of a model's primary task of word prediction. This further leads to a better characterization of what properties of language does pretraining with the language modeling objective inject into the model, and what it does not.

5.2.3 Thorough investigation into distraction effects

This thesis reports on the existence of counter-facilitatory or "distraction" effects in the presence of related words. This exploratory finding can be augmented by conducting controlled experiments that aim to discern patterns in priming as discussed in this thesis that lead to distraction rather than priming in models. From the qualitative analysis discussed in the previous chapter, a likely explanation behind distraction can be found by studying the dynamics of the fit of the related primes themselves to the unprimed cloze-context and its interaction with the nature of the context itself. This category of future work is important to potentially understand and develop counter-measures against misrpriming to train better word prediction models.

REFERENCES

Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., & Goldberg, Y. (2017). Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *5th international conference on learning representations, ICLR 2017, toulon, france, april 24-26, 2017, conference track proceedings.* Retrieved from `https://openreview.net/forum?id=BJh6Ztuxl`

Alishahi, A., Chrupała, G., & Linzen, T. (2019). Analyzing and interpreting neural networks for nlp: A report on the first blackboxnlp workshop. *Natural Language Engineering*, *25*(4), 543–557.

Auguste, J., Rey, A., & Favre, B. (2017). Evaluation of word embeddings against cognitive processes: primed reaction times in lexical decision and naming tasks. In *Proceedings of the 2nd workshop on evaluating vector space representations for NLP* (pp. 21–26). Copenhagen, Denmark: Association for Computational Linguistics.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, *59*(4), 390–412.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., & Glass, J. (2017). What do neural machine translation models learn about morphology? In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 861–872).

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, *3*(Feb), 1137–1155.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). 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).

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. (2017). A neurocomputational model of the n400 and the p600 in language processing. *Cognitive science*, *41*, 1318–1352.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... others (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1724–1734).

Chow, W.-Y., Smith, C., Lau, E., & Phillips, C. (2016). A "bag-of-arguments" mechanism for initial verb predictions. *Language, Cognition and Neuroscience*, *31*(5), 577–596.

Chowdhury, S. A., & Zamparelli, R. (2018). Rnn simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th international conference on computational linguistics* (pp. 133–144).

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 acl workshop blackboxnlp: Analyzing and interpreting neural networks for nlp* (pp. 276–286).

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=r1xMH1BtvB

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, *82*(6), 407.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on machine learning* (pp. 160–167).

Dagan, I., Roth, D., Sammons, M., & Zanzotto, F. M. (2013). Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, *6*(4), 1–220.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Eisenstein, J. (2019). *Introduction to natural language processing*. MIT Press.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179–211.

Ettinger, A. (2020). 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.

Ettinger, A., Elgohary, A., Phillips, C., & Resnik, P. (2018). Assessing composition in sentence vector representations. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1790–1801).

Ettinger, A., Elgohary, A., & Resnik, P. (2016). Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st workshop on evaluating vector-space representations for nlp* (pp. 134–139).

Ettinger, A., & Linzen, T. (2016). Evaluating vector space models using human semantic priming results. In *Proceedings of the 1st workshop on evaluating vector-space representations for NLP* (pp. 72–77). Berlin, Germany: Association for Computational Linguistics.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of memory and Language*, *41*(4), 469–495.

Firth, J. R. (1957). *A synopsis of linguistic theory 1930-1955.* Studies in linguistic analysis.

Fischler, I., Bloom, P. A., Childers, D. G., Roucos, S. E., & Perry Jr, N. W. (1983). Brain potentials related to stages of sentence verification. *Psychophysiology*, *20*(4), 400–409.

Fischler, I., Childers, D. G., Achariyapaopan, T., & Perry Jr, N. W. (1985). Brain potentials during sentence verification: Automatic aspects of comprehension. *Biological psychology*, *21*(2), 83–105.

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of NAACL-HLT 2019* (pp. 32–42). Minneapolis, Minnesota: Association for Computational Linguistics.

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, *10*(1), 1–309.

Goldberg, Y. (2019). Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless Green Recurrent Networks Dream Hierarchically. In *Proceedings of NAACL-HLT 2018* (pp. 1195–1205). New Orleans, Louisiana: Association for Computational Linguistics.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Second meeting of the north American chapter of the association for computational linguistics.* Retrieved from https://www.aclweb.org/anthology/N01-1021

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146–162.

Haxby, J. V. (2012). Multivariate pattern analysis of fmri: the early beginnings. *Neuroimage*, *62*(2), 852–855.

Hempelmann, C. F., Taylor, J. M., & Raskin, V. (2010). 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).

Hewitt, J., & Liang, P. (2019). Designing and interpreting probes with control tasks. 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. 2733–2743).

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. 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. 4129–4138).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 328–339). Melbourne, Australia: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P18-1031 doi: 10.18653/v1/P18-1031

Hupkes, D., Dankers, V., Mul, M., & Bruni, E. (2020). Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, *67*, 757–795.

Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap? a microanalytic review. *Psychonomic bulletin & review*, *10*(4), 785–813.

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C.-S., ... Buchanan, E. (2013, December). The semantic priming project. *Behavior Research Methods*, *45*(4), 1099–1114.

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C.-S., ... Buchanan, E. (2013). The semantic priming project. *Behavior Research Methods*, *45*(4), 1099–1114.

John, M. F. S., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial intelligence*, *46*(1-2), 217–257.

Jurafsky, D., & Martin, J. H. (2020). *Speech & Language Processing, 3rd Edition*. Retrieved from `https://web.stanford.edu/~jurafsky/slp3/`

Kassner, N., & Schütze, H. (2020, July). Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7811–7818). Online: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/2020.acl-main.698` doi: 10.18653/v1/2020.acl-main.698

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp). *Annual review of psychology*, *62*, 621–647.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*.

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 1192–1202).

McCann, B., Bradbury, J., Xiong, C., & Socher, R. (2017). Learned in translation: Contextualized word vectors. In *Advances in neural information processing systems* (pp. 6294–6305).

McClelland, J. L., & O'Regan, J. K. (1981). Expectations increase the benefit derived from parafoveal visual information in reading words aloud. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(3), 634.

McClelland, J. L., & Rumelhart, D. E. (1986). Parallel distributed processing. *Explorations in the Microstructure of Cognition*, *2*, 216–271.

McNamara, T. P. (2005). *Semantic priming: Perspectives from memory and word recognition*. Psychology Press.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, *37*(4), 547–559.

McShane, M. (2017). Natural language understanding (nlu, not nlp) in cognitive systems. *AI Magazine*, *38*(4), 43–56.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of experimental psychology*, *90*(2), 227.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Kombrink, S., Burget, L., Černockỳ, J., & Khudanpur, S. (2011). Extensions of recurrent neural network language model. In *2011 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5528–5531).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, *38*(11), 39–41.

Misra, K., Ettinger, A., & Rayz, J. (2020a, November). Exploring BERT's sensitivity to lexical cues using tests from semantic priming. In *Findings of the association for computational linguistics: Emnlp 2020* (pp. 4625–4635). Online: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/2020.findings-emnlp.415`

Misra, K., Ettinger, A., & Rayz, J. (2020b). Exploring lexical relations in bert using semantic priming. In *Proceedings of the 42nd annual conference of the cognitive science society* (p. 1939).

Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., … Allen, J. (2016). A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of NAACL-HLT 2016* (pp. 839–849). San Diego, California: Association for Computational Linguistics.

Mostafazadeh, N., Roth, M., Louis, A., Chambers, N., & Allen, J. (2017). LSDSem 2017 Shared Task: The Story Cloze Test. In *Proceedings of the 2nd workshop on linking models of lexical, sentential and discourse-level semantics* (pp. 46–51).

Nieuwland, M. S., & Kuperberg, G. R. (2008). When the truth is not too hard to handle: An event-related potential study on the pragmatics of negation. *Psychological Science*, *19*(12), 1213–1218.

Nirenburg, S., & Raskin, V. (2004). *Ontological semantics*. MIT Press.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., … others (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the tenth international conference on language resources and evaluation (lrec'16)* (pp. 1659–1666).

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of EMNLP 2014* (pp. 1532–1543).

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. 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. 2227–2237).

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. (2019). 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).

Plaut, D. C. (1995). Semantic and associative priming in a distributed attractor network. In *Proceedings of the 17th annual conference of the cognitive science society* (Vol. 17, pp. 37–42).

Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral science*, *12*(5), 410–430.

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the n400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, *2*(9), 693–705.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training* (Tech. Rep.). Open AI.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 2383–2392).

Raskin, V., Hempelmann, C. F., & Taylor, J. M. (2010). Guessing vs. knowing: The two approaches to semantics in natural language processing. In *Annual international conference dialogue 2010* (pp. 642–650).

Raskin, V., & Taylor, J. M. (2009). 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).

Raskin, V., & Weiser, I. (1987). Chapter 8. In *Language and Writing: Applications of Linguistics to Rhetoric and Composition.* ABLEX Publishing Corporation.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.

Schwanenflugel, P. J. (1991). Contextual constraint and lexical processing. In *Advances in psychology* (Vol. 77, pp. 23–45). Elsevier.

Schwanenflugel, P. J., & LaCount, K. L. (1988). Semantic relatedness and the scope of facilitation for upcoming words in sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(2), 344–354.

Schwanenflugel, P. J., & Shoben, E. J. (1985). The influence of sentence constraint on the scope of facilitation for upcoming words. *Journal of Memory and Language*, *24*(2), 232–252.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Shwartz, V., & Dagan, I. (2019). Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, *7*, 403–419.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Stanovich, K. E., & West, R. F. (1983). On priming by a sentence context. *Journal of Experimental Psychology: General*, *112*(1), 1–36.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (pp. 3104–3112).

Taylor, J. M., Hempelmann, C. F., & Raskin, V. (2010). 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).

Taylor, J. M., & Raskin, V. (2010). Fuzzy ontology for natural language. In *2010 annual meeting of the north american fuzzy information processing society* (pp. 1–6).

Taylor, J. M., & Raskin, V. (2011). Understanding the unknown: Unattested input processing in natural language. In *2011 ieee international conference on fuzzy systems (fuzz-ieee 2011)* (pp. 94–101).

Taylor, J. M., & Raskin, V. (2016). Conceptual defaults in fuzzy ontology. In *2016 annual conference of the north american fuzzy information processing society (nafips)* (pp. 1–6).

Taylor, J. M., Raskin, V., & Hempelmann, C. F. (2011). Towards computational guessing of unknown word meanings: The Ontological Semantic approach. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33).

Taylor, J. M., Raskin, V., Hempelmann, C. F., & Attardo, S. (2010). An unintentional inference and ontological property defaults. In *2010 ieee international conference on systems, man and cybernetics* (pp. 3333–3339).

Taylor, W. L. (1953). “cloze procedure”: A new tool for measuring readability. *Journalism quarterly*, *30*(4), 415–433.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., … others (2019). What do you learn from context? probing for sentence structure in contextualized word representations. In *7th international conference on learning representations, iclr 2019.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Veldhoen, S., Hupkes, D., Zuidema, W., et al. (2016). Diagnostic classifiers: Revealing how neural networks process hierarchical structure. In *Ceur workshop proceedings* (Vol. 1773).

Wallace, E., Wang, Y., Li, S., Singh, S., & Gardner, M. (2019). Do nlp models know numbers? probing numeracy in embeddings. 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. 5310–5318).

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., … Bowman, S. R. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in neural information processing systems* (pp. 3266–3280).

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International conference on learning representations.*

Warstadt, A., Cao, Y., Grosu, I., Peng, W., Blix, H., Nie, Y., … others (2019). Investigating bert's knowledge of language: Five analysis methods with npis. 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. 2870–2880).

Wilcox, E., Levy, R., & Futrell, R. (2019). Hierarchical representation in neural language models: Suppression and recovery of expectations. In *Proceedings of the 2019 acl workshop blackboxnlp: Analyzing and interpreting neural networks for nlp* (pp. 181–190).

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do rnn language models learn about filler–gap dependencies? In *Proceedings of the 2018 emnlp workshop blackboxnlp: Analyzing and interpreting neural networks for nlp* (pp. 211–221).

Williams, A., Nangia, N., & Bowman, S. R. (2018, June). 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). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/N18-1101` doi: 10.18653/v1/N18-1101

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., . . . others (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, *8*(3), 338–353.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 ieee international conference on computer vision (iccv)* (pp. 19–27).