How Relevant Are Selectional Preferences for Transformer-based Language Models?

Eleni Metheniti ♣♦
IRIT (CNRS)♦
Université Toulouse - Paul Sabatier (UT3)
31400 Toulouse, France

Tim Van de Cruys♠
KU Leuven ♠
Faculty of Arts
Department of Linguistics
Leuven.AI institute
B-3000 Leuven, Belgium

Nabil Hathout ♣
CLLE-CNRS ♣
Université Toulouse - Jean Jaurès (UT2J)
Maison de la Recherche
31058 Toulouse, France

firstname.lastname@{univ-tlse2.fr♣,irit.fr♦,kuleuven.be♠}

Abstract

Selectional preference is defined as the tendency of a predicate to favour particular arguments within a certain linguistic context, and likewise, reject others that result in conflicting or implausible meanings. The stellar success of contextual word embedding models such as BERT in NLP tasks has led many to question whether these models have learned linguistic information, but up till now, most research has focused on syntactic information. We investigate whether BERT contains information on the selectional preferences of words, by examining the probability it assigns to the dependent word given the presence of a head word in a sentence. We are using word pairs of head-dependent words in five different syntactic relations from the SP-10K corpus of selectional preference (Zhang et al., 2019b), in sentences from the ukWaC corpus, and we are calculating the correlation of the plausibility score (from SP-10K) and the model probabilities. Our results show that overall, there is no strong positive or negative correlation in any syntactic relation, but we do find that certain head words have a strong correlation, and that masking all words but the head word yields the most positive correlations in most scenarios—which indicates that the semantics of the predicate is indeed an integral and influential factor for the selection of the argument.

1 Introduction

Motivated by their semantics, the vast majority of predicates (such as verbs) have a tendency to favour certain arguments to others. Consider the following examples:

(1) The athlete runs a marathon.

(2) The bassoon runs a banana.

Most native speakers would readily accept example (1) as a well-formed English sentence, while example (2), even though it is syntactically correct, is more likely to be judged as awkward and ill-formed. The first example is semantically felicitous (since it is perfectly normal for athletes to run, and a marathon is something that can be run), while the second example is semantically infelicitous (both a bassoon and a banana are inanimate entities without a literal capability of motion). This preference of predicates for particular arguments is known as *selectional preference* (Katz and Fodor, 1963). A proper understanding of this phenomenon is important within various natural language processing (NLP) applications, and selectional preferences have indeed been used as an additional knowledge source for various NLP tasks, such as word sense disambiguation (McCarthy and Carroll, 2003) and semantic role labeling (Gildea and Jurafsky, 2002).

While language processing architectures prior to the neural network paradigm primarily made use of a sequential NLP pipeline, where designated modules sequentially provide increasingly complex linguistic annotations (such as part of speech tagging and syntactic parsing), more recent approaches tend to tackle NLP problems with a single, overarching neural network architecture: words are modeled as multi-dimensional embeddings that are fed to the neural network architecture, without any additional linguistic

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 28th International Conference on Computational Linguistics, pages 1266–1278
Barcelona, Spain (Online), December 8-13, 2020

annotations; all the linguistic knowledge—that the neural network might exploit for its final decisions—is modeled implicitly throughout the neural network's hidden layers. This line of research has culminated in the recent contextual embedding language architectures, such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), which adapt the individual word embeddings in order to yield a fine-grained representation of the meaning within the context of a specific text fragment, and these embeddings can subsequently be exploited for language prediction tasks.

The implicit nature of the linguistic knowledge contained within these recent approaches opens up a number of questions with regard to the relevance of selectional preferences in the neural network paradigm. Whereas previous NLP approaches generally considered selectional preferences as a stand-alone module that may be easily evaluated *in vitro*, this is not the case for current neural approaches: the neural network functions as a black box, and as such it is impossible to find out exactly what kind of information has led the model to make a final decision, and whether selectional preference information played a significant role in it. It is therefore necessary to design specific experimental settings in order to inspect what information the network actually made use of. Lately, there has been a great deal of interest in comprehending the way contextual language models encode linguistic knowledge. However, most work has focused on syntactic phenomena, such as the identification of constituents and dependencies, and syntactic agreement; investigating how these models encode semantics proves to be a more difficult issue, and research into the question is much less common (Mickus et al., 2020).

In this paper, our goal is to investigate whether Transformer models include selectional preference information in their embeddings and whether they rely on it in order to make their predictions—and if so, to what extent that information is relevant or vital. We focus on the BERT architecture, and consider the influence of selectional preferences for the model's standard *masked language model* pretraining objective. Our sentences consist of pairs of head and dependent words from the SP-10K dataset, annotated with a plausibility score by human judges, and spanning five different types of syntactic relations (Zhang et al., 2019b). We create prompt sentences based on these word pairs and a syntactically annotated version of the ukWaC corpus (Ferraresi et al., 2008) in order to extract full sentences with said pairs. We then feed the prompt sentences masking the dependent word of the word pair to BERT. Then, we are able to observe whether there is a strong correlation between the plausibility score of a word pair (how likely it is for the head and the dependent word to exist in an utterance) and the probability of the dependent word in the context of a sentence, as assigned by BERT. Subsequently, we can make assumptions on how much the head word contributed to the prediction of the dependent word, by applying *attention masks* to the input sequences of the model. Our findings show that, over all sentences and word pairs, there is no strong positive or negative correlation between the plausibility score (as retrieved from SP-10K) and the probability of the dependent word; however, we make note of some head words and phenomena that showed significant positive or negative correlation ($>$0.4) in some syntactic relations. Moreover, our results show that masking all words but the head word yields the most positive correlations in most scenarios, which indicates that the semantics of the predicate is indeed an integral and influential factor for the selection of the argument.

This paper is structured as follows. In Section 2, we provide an overview of related work, both with regard to selectional preference modeling and with regard to interpretability of neural architectures. In Section 3, we outline the methodology of our research. Section 4 provides an overview of the results, including a quantitative comparison of the various experimental settings as well as a qualitative discussion. Section 5 describes some of the challenges that we faced in our research, as well as a general discussion of our results. Finally, Section 6 summarizes our findings and our key points, and provides a number of avenues for future work.

2 Previous Work

2.1 Selectional Preference Induction

A seminal approach to selectional preference induction makes use of posterior probabilities: Resnik (1996) relies on WordNet synsets in order to generate generalized noun clusters. The *selectional preference strength* of a specific verb v in a particular relation is calculated by computing the Kullback–Leibler

divergence between the posterior cluster distribution of the verb and the prior cluster distribution:

$$S_{R(v)} = \sum_{c} \mathrm{p}(c|v) \log \frac{\mathrm{p}(c|v)}{\mathrm{p}(c)} \tag{1}$$

where c stands for a noun cluster, and R stands for a given predicate-argument relation. The *selectional association* of a particular noun cluster is the contribution of that cluster to the verb's preference strength.

$$A_{R(v,c)} = \frac{\mathrm{p}(c|v) \log \frac{\mathrm{p}(c|v)}{\mathrm{p}(c)}}{S_{R(v)}} \tag{2}$$

The model's generalization relies entirely on WordNet, and there is no generalization among the verbs.

A number of other researchers have equally exploited WordNet for generalization (Li and Abe, 1998; Clark and Weir, 2001; Ó Séaghdha and Korhonen, 2012). Most researchers, however, acknowledge the shortcomings of hand-crafted resources, and focus on the acquisition of selectional preferences from corpus data. Rooth et al. (1999) propose an Expectation–Maximization (EM) clustering algorithm for selectional preference acquisition based on a probabilistic latent variable model. In their model, both predicate and argument are generated from a latent variable, where the latent variables represent clusters of tight verb–argument interactions.

$$\mathrm{p}(v, o) = \sum_{c \in C} \mathrm{p}(c, v, o) = \sum_{c \in C} \mathrm{p}(c)\mathrm{p}(v|c)\mathrm{p}(o|c) \tag{3}$$

The use of latent variables allows the model to generalize to predicate–argument tuples that have not been seen during training. The latent variable distribution—and the probabilities of predicates and argument given the latent variables—are automatically induced from data using EM.

Erk (2007) and Erk et al. (2010) describe a method that uses corpus-driven distributional similarity metrics for the induction of selectional preferences. The key idea is that a predicate-argument tuple (v, o) is felicitous if the predicate v appears in the training corpus with arguments o' similar to o, i.e.

$$S(v, o) = \sum_{o' \in O_v} \frac{wt(v, o')}{Z(v)} \cdot sim(o, o') \tag{4}$$

where O_v represents the set of arguments that have been attested with predicate v, $wt(\cdot)$ represents an appropriate weighting function (in its simplest form the frequency of the (v, o') tuple), and Z is a normalization factor.

Van de Cruys (2009) presents a model based on tensor factorization, which is able to model multi-way selectional preferences. Three-way co-occurrences of subjects, verbs, and objects are represented as a three-way tensor (the generalization of a matrix), and a latent tensor factorization model is applied in order to generalize to unseen instances.

A number of researchers presented models that are based on the framework of topic modeling. Ó Séaghdha (2010) describes three models for selectional preference induction based on Latent Dirichlet Allocation, which model the selectional preference of a predicate and a single argument. Ritter et al. (2010) equally present a selectional preference model based on topic modeling, but they tackle multi-way selectional preferences (of transitive predicates, which take two arguments) instead.

More recently, neural network based approaches have equally been used. Van de Cruys (2014) presents an approach based on feed-forward neural networks; predicates and arguments are represented as embeddings, and serve as input to a simple feed-forward neural network architecture, which yields a single selectional preference value. And Zhang et al. (2019a) present multiplex word embeddings for selectional preference modeling. The key idea is to create, for each word, a 'central embedding' that represents the word's global semantics, and several 'relational' embeddings that represent how the word relates to other words within a specific syntactic relation.

For the evaluation of selectional preference models, researchers have exploited two different kinds of evaluation tasks: pseudo-disambiguation (Rooth et al., 1999; Ritter et al., 2010; Van de Cruys, 2014),

and comparison to human judgements (McRae et al., 1998; Zhang et al., 2019b). The goal of the former is to discriminate actually attested selectional preference pairs (extracted from a large corpus) from randomly-constructed, corrupted pairs. For the latter kind, selectional preference judgements of the model are compared to manually labeled datasets of human judgements, using a correlation measure. In this research, we make use of the latter kind of evaluation.

2.2 Interpretability of Neural Models

The recent emergence of contextual embedding models such as BERT, and their subsequent application and improvement they brought on many NLP tasks, has been followed by extensive research on whether these embeddings accommodate an implicit understanding of linguistic and semantic knowledge. A great deal of research has been focused on the syntactic knowledge learned by BERT; Goldberg (2019) has found that BERT (a Transformer model) is more robust in syntactic tasks than a simple LSTM architecture (a Recurrent Neural Network), and with a series of probing tasks on different datasets he proved that there is some syntactic knowledge beyond semantic and contextual relations. BERT was especially more successful in such tasks compared to other contextual embedding models, because of its bi-directional architecture (Wolf, 2019). Further research on learned syntactic information showed that BERT captures phrase-level information in the lower layers, and learns more sophisticated relations in higher layers (Jawahar et al., 2019). Coenen et al. (2019) found that the attention matrices output by `bert-base-uncased` contain syntactic representations, with certain directions in space representing specific relations, and they were also able to locate similar sub-spaces for semantic relations. Petroni et al. (2019) report that BERT contains enough relational knowledge to compete with knowledge-based methods on tasks such as open-type questions, which leads them to the conclusion that the model has acquired a certain level of semantic knowledge. And Ettinger (2020) presents a number of experiments in which in many cases BERT makes good predictions with regard to semantic fit, such as hypernyms and subject-object nouns.

However, McCoy et al. (2019) question the ability of BERT—and similar pretrained models—to truly capture deep linguistic structures and semantic information, as past bibliography has suggested. Tenney et al. (2019) also investigated pretrained models on their performance on both syntactic and semantic phenomena, and concluded that simple syntactic phenomena were successfully identified, but phenomena which mostly relied on semantic relations were not as easily learned. Ettinger (2020) has also pointed out that BERT performance on predictions dropped in cases of true/false statements and negations. Zhang et al. (2019c) created SemBERT, a BERT model with integrated explicit contextual semantics, supporting the fact that external semantic knowledge was more useful than manipulating inherent model knowledge to achieve better results in semantics-related tasks. Mickus et al. (2020) delve further into exploring the embeddings of BERT, and report that it is uncertain whether the embeddings are able to properly represent semantic similarities on a word-base level (as the theory of distributional semantics would suggest), due to the influence of the context sentence on the distributional semantics space (even without meaning correlates).

3 Methodology

3.1 Selectional Preference Corpus

Out of the several datasets of syntactic-semantic relations which have been released throughout the years for linguistics and NLP research, such as MST98 (McRae et al., 1998), F-Inst (Ferretti et al., 2001), P07 (Padó, 2007) and GDS-all (Greenberg et al., 2015), we decided to use SP-10K (Zhang et al., 2019b). SP-10K is the largest dataset available to date for evaluating the selectional preference abilities of natural language processing tasks. It is composed of slightly over 10K pairs of words,[1] evenly split into five different types of syntactic relations: **nsubj** (verb and noun as verb + subject), **dobj** (verb and noun as verb + direct object), **amod** (noun and adjective where the adjective is a modifier to the noun), **nsubj_amod**

[1]The authors had initially created 2,000 word pairs for each category, and later they added 124 more word pairs for the **nsubj_amod** and **dobj_amod** categories, from the Winograd Schema Challenge dataset (Levesque et al., 2012).

(verb and adjective where the adjective is a modifier to a noun which is the verb's subject; the noun remains undefined), and **dobj_amod** (verb and adjective where the adjective is a modifier to a noun which is the verb's direct object; the noun remains undefined). While the first three categories deal with one-hop syntactic relations, the two latter represent higher-level, two-hop dependencies, which the authors claim to also include meaningful semantic connections in certain cases and contexts. The words composing the word pairs are 2,500 frequent words, lemmatized, and all of the word pairs are annotated with a *plausibility score*, i.e. a value between 0 and 10, which is derived from human judgements on how plausible the dependent word (noun or adjective) is as an argument or modifier to the head word (verb or noun).

3.2 Prompt sentence corpus

Our goal is to investigate the relative importance of selectional preference information on BERT's predictions for a masked word within the context of a complete sentence. Therefore, the word pairs of the SP-10K corpus do not suffice; additionally, we need to find appropriate sentences that include the word pairs in the correct syntactic positions. We want grammatical sentences, with varied contexts, in order to examine in which cases selectional preferences have an influence on the prediction of the masked word. We decided not to compose our own prompt sentences, as this would be a very large-scale effort and could have introduced some unwanted biases. On the other hand, existing datasets of prompt sentences were either too small to include a sufficient amount of the SP-10K word pairs (such as the Corpus of Linguistic Acceptability, Warstadt et al. (2018)) or too specialized on semantic relations (such as the LPAQA corpus, Jiang et al. (2020)).

Thus, we decided to use a large corpus and extract sample sentences for each word pair. The ukWaC corpus was created by crawling websites in the .uk domain, and it includes a variety of texts in English (articles, titles, user reviews, etc.) and over 2 billion words (Ferraresi et al., 2008). In order to find the SP-10K word pairs in the ukWaC sentences in the correct syntactic positions and relations, we used a syntactically annotated version of the corpus, which was parsed using the Malt parser (Nivre et al., 2006).

Out of the 85 million sentences in the ukWaC corpus, we looked for short sentences (4 to 15 tokens), in order to stay well under BERT's limit of 512 tokens per sequence, but also in order to ensure that the sentences would not be erroneously composed of multiple sentences (due to segmentation errors), or be composed of multiple clauses, or include complex and distant dependencies. In an effort to eliminate syntactic complexity as an extraneous factor of prediction difficulty, we considered excluding some specific dependency labels, such as *xcomp* (open clausal complement) and *acl:rel* (for relative subclauses), but our selected sentences were already short enough to not broadly include more complex syntactic phenomena. We did not exclude passive clauses, because automatic parsing is rarely able to label passive structures correctly, and we considered that for most of our syntactic relations, their existence would not cause a problem. Also, we determined that the distance between the two words of the pair in the sentence should be between one and five words, allowing enough positions for determiners and modifiers, but not too distant to raise complexity.

At this stage in the collection of sentences, we decided to investigate the quality of the selected sentences. One problem we faced was parsing errors—which inevitably occur in automatic dependency parsing, but when repeated on the same word pair, they potentially produce many false prompt sentences. Additionally, we noticed a number of problems with regard to the quality of word pairs in the SP-10K corpus. For example, we noticed that some of the word pairs, in the context of the specified syntactic relation, should have been assigned with the lowest possible score (0), but were, in several cases, considered felicitous or at least plausible. We are aware that some of the word pairs were intentionally designed to be of low frequency or low plausibility; we are referring to falsely tagged syntactic structures (e.g. "look way" is incorrectly included as a plausible **dobj** word pair, but the dependency between these words is always verb and adverbial modifier). For these reasons, we decided it was imperative to perform a quick and non-exhaustive manual evaluation of the SP-10K word pairs and the resulting extracted sentences. We provide an elaborate discussion on the challenges we faced with our datasets in Section 5.1.

In Table 1, we present the number of sentences, for each type of SP-10K word pairs, for which we

found at least one sentence containing the word pair in the given dependency relation and parts of speech, with the criteria of length and distance that we previously determined, and the final counts of the sentences after our manual evaluation.

Type	*Word pairs in ukWaC*	*Found sents*	*Final sents*	*Avg. plaus. score*
nsubj	958 / 2,000	38,613	30,526	6.64
dobj	980 / 2,000	70,250	56,777	7.39
amod	1,030 / 2,000	29,403	23,110	7.62
nsubj_amod	956 / 2,061	15,265	12,911	5.75
dobj_amod	922 / 2,063	28,336	21,839	6.32
TOTAL	4,846 / 10,124	181,867	145,163	

Table 1: The number of SP-10K word pairs which were found in sentences of the ukWaC corpus (out of the total number of word pairs), the initial number of found sentences and how many of those sentences include the word pairs in the correct syntactic positions (after our evaluation). In the last column is the average value of the plausibility scores over all sentences.

3.3 BERT

BERT (Devlin et al., 2019) is a Transformer-based bi-directional encoder, which is trained by randomly sampling positions in the input sequence and learning to fill the word in the masked position. The pretrained version has been trained on the Toronto BookCorpus (Zhu et al., 2015) and the English edition of Wikipedia, and pretrained models have been made available with 12 layers of attention (`bert-base`) or 24 layers (`bert-large`), and trained on lower-cased corpora (`uncased`) or as is (`cased`).

For our experiments, we used the `bert-base-uncased` model for English, as provided by HuggingFace's `transformers` Python library (Wolf et al., 2019); this model has been used by Goldberg (2019) for syntactic tasks and performed well, and has also been favoured by the researchers probing for the semantic tasks mentioned in Section 2 (e.g. McCoy et al. (2019)). Some preliminary experiments we performed with `bert-large-uncased` did not show significantly different results, thus for the sake of time-conservation we used the computationally-lighter `base` model. We add the special BERT tokens [CLS] (to indicate the start of a sentence) and [SEP] (to indicate the end of it). We make use of the model's built-in tokenizer, `BertTokenizer`, and we do not perform any finetuning of the encoder weights, but make use of the pretrained model as it is made available.

3.4 Correlation of SP-10K score and BERT probability

For each example sentence in our corpus, we mask the dependent word of the word pair using a [MASK] token, and we retrieve the probability that is assigned to the target word in the focal position. The probability is computed by passing the last hidden state through a *softmax* function, a feature that is also used by Wang and Cho (2019) in the context of language modeling. We are making the assumption that this result is to be treated as the conditional probability of a bi-directional language model (similar to what a traditional language model would return) even though we are aware that BERT's bi-directional nature means that this assumption is not self-evident.

Next, we compute the correlation of the masked word's probability and the plausibility score of the word pair, using the Kendall rank correlation coefficient as implemented by the `scipy` Python library. Kendall τ (tau) correlation is a non-parametric measure of the monotonicity of the relationship between two datasets. The p-value roughly indicates the probability of an uncorrelated system producing datasets that have a correlation at least as extreme as the one computed from these datasets.[2] Values close to 1 indicate strong positive correlation, while values close to -1 indicate strong disagreement. Intuitively, we are looking for a strong positive correlation, meaning that the higher the plausibility score of the word pair, the higher the probability of the dependent word in the context of the head word.

[2]In the remainder of this paper, significant results are defined as $p < 0.01$.

sentence		the	film	tells	the	story	of	that	trial	
standard	[CLS]	the	film	tells	the	[MASK]	of	that	trial	[SEP]
head	[CLS]	the	film		the	[MASK]	of	that	trial	[SEP]
context	[CLS]			tells		[MASK]				[SEP]
control	[CLS]					[MASK]				[SEP]

Figure 1: Illustration of the four attention mask settings, for a sentence with the word pair "tell story" (as a **dobj** relation). Greyed out words indicate blocked attention.

In order to determine the relative importance of selectional preference information, we compute probabilities within different attention settings by using **attention masks**; the attention mask is an array of 1s and 0s indicating which tokens we do not wish to incorporate in the way the model attends to the sequence. By using this feature, we are able to "block" certain tokens of the sentence from Bert's self-attention mechanism, and examine the impact it brings to the probability scores and the correlation. We use four different settings: the *standard* setting does not involve any masking, thus the model can attend to the whole sequence; the *head* setting blocks attention to the head word of the pair, so prediction needs to be based on other context words; the *context* setting masks all the context words, so prediction needs to be based on the head word (and BERT's special tokens), and the *control* setting masks all the words of the sequence (except for the special tokens), so that no adequate prediction should be possible (as a sanity check). A graphical illustration of the four different attention settings is given in Figure 1.

We compute correlation scores between the model's probabilities and the plausibility scores for all sentences, and provide both micro- and macro-averaged results. As mentioned before, the number of extracted example sentences per word pair differs significantly. The micro-averaged results are computed over the entire set of sentences, without taking into account this variable number of sentences. For the macro-averaged results, we first compute the average for each word pair, and then provide a global average for all the pairs (hence treating all word pairs as equally important). We will consider as a strong positive correlation a value above 0.4, and as strong negative correlation a value below -0.4.

4 Results

As seen in Table 2, we do not observe a strong positive or negative correlation for any of the five syntactic relation categories and for any of the attention mask scenarios. Out of all the attention mask scenarios, *context* mask (masking attention for the entire sequence except for the head word and BERT tokens) shows the highest positive values (fair correlation up to +/- .30), while the *head* attention mask scenario (masking attention for the head word and attending to the context for prediction) showed no correlation but was biased to negative values. Compared to the *standard* setting (all tokens are attended to), we notice the effect that the head word has in predictions; the context words affect the probability of the prediction, but the absence of the head word is detrimental. This observation supports the argument that, for the syntactic relations involving verbs as the head word, the verb's selectional preferences are relatively important and influential enough in the selection of constituents, and that BERT is aware of these preferences and constraints, and uses them to assign a proportionate probability to the dependent word.

Note that for the **amod** syntactic relations (noun and its modifier) the correlation scores are among the highest, reasonably so, as the noun should place stronger constraints on its modifier. The **nsubj** relations show slightly more positive correlations than the **dobj** ones, probably due to the constraint of animacy for a subject (word pairs included a variety of animate and inanimate subjects). The two-hop relations, **nsubj_amod** and **dobj_amod** show lower correlations, still relatively high for the *context* attention mask setting, which is in accordance with the hypothesis of Zhang et al. (2019b) that selectional preferences span further than one-hop relations.

Taking a closer look at the head words of the word pairs, we searched for strong positive or negative correlations for each head word that exists in at least two different word pairs, per syntactic relation. We

	standard	*head*	*context*	*control*
nsubj	0.03	-0.02	0.16	-0.01
dobj	0.05	-0.07	0.05	-0.05
amod	0.04	-0.06	0.24	-0.04
nsubj_amod	-0.01	-0.13	0.29	-0.00
dobj_amod	0.06	0.01	-0.03	0.02

(a) Micro-averaged results

	standard	*head*	*context*	*control*
nsubj	0.19	0.15	0.29	0.08
dobj	0.16	0.04	0.27	0.05
amod	0.15	0.03	0.35	0.03
nsubj_amod	0.01	-0.04	0.22	0.06
dobj_amod	0.14	0.10	0.20	0.07

(b) Macro-averaged results

Table 2: Kendall τ (tau) correlation coefficient of masked word probability and word pair plausibility score.

examined whether specific verbs and nouns affected positively or negatively the correlation of probability and plausibility, and whether there were common features between these head words (e.g. semantic similarity, common semantic class). We grouped the probabilities and scores of sentences per head word, and calculated the correlation coefficient for head words that were present in at least two different word pairs. Overall, for all five syntactic categories of our experiment, we do not notice distinct classes, semantic or syntactic, that the words with strong correlations could be grouped with.

For **nsubj** relations, verbs of semantic similarity (in at least one of their meanings) did not demonstrate similar patterns of probability and correlation; for example, the verbs of violence (in some contexts) "kill", "strike", "grab", "fire" show a strong positive correlation, while the verbs of the same semantic class "shoot" and "confront" have a strong negative correlation – this could be caused by the different metaphorical meanings that these two words might have, or the dependent words that they were paired with in the SP-10K dataset (favorable for "kill", detrimental for "shoot"). Concerning the type of subjects, the animate subject "man" had a high plausibility score in the SP-10K dataset and high probability scores for "kill" and "shoot", causing a strong positive correlation. The inanimate subjects had mid-range plausibility scores ("earthquake", "explosion") or low scores ("film", "tragedy") but the probability varied based on the sentence and metaphorical use; for the word pair "strike tragedy" which existed in many sentences of our dataset, the plausibility score was 5.25 and the assigned probability for "tragedy" was relatively low, even though the idiomatic phrase "tragedy struck" is fairly common.

Examining the **dobj** relations, verbs (head words) showed inconsistent correlations among the different attention mask scenarios; out of the few verbs that showed consistently positive or negative correlation, we were not able to identify semantic clusters of verbs, neither differences based on verb transitivity (monotransitive/ditransitive). The presence of a strong correlation relied more on BERT's semantic knowledge rather than world knowledge or utterance plausibility; for example, the word pair "blame customer" has (correctly) a moderate plausibility score (6.75), is found twice in the ukWaC corpus, but the assigned probability by BERT of the word "customer" is very low in the *standard* and *context* attention mask scenarios. The word pair "blame management", on the other hand, with slightly lower plausibility (6.25) is assigned a proportionally good probability. This leads us to the conclusion that, even though both syntactic pairs are grammatically correct and have commonly used words, the pretrained model has learned that "management" (someone in control and responsible of a service) is a more probable direct object for the verb "blame" than the word "customer (the receiver of a service), especially when the only given context is the verb. When attention to the head word was removed, there was no strong negative correlation between "blame" and the given plausibility score.

Concerning the **amod** word pairs, again no semantic class of nouns appear consistently in the positive or relative correlation groups. An interesting observation is that high-frequency adjectives of size and age, such as "small", "big", "old" and "new" were almost always assigned a high probability by BERT, but the variations in plausibility score (from 8.25 to 4.25) led to strong positive or negative correlations, especially since word pairs with these adjectives are quite frequent in our corpus, for example "new house", "small bird" and "new face" had many occurrences in the corpus and a strong positive correlation (high plausibility/high probability), "new material" (6.5) and "old daughter" (4.25) had lower plausibility scores and subsequently lower probabilities, in all attention mask scenarios.

In the **nsubj_amod** word pairs, again we see that high-frequency descriptive adjectives (dependent words) are still assigned higher probabilities, even though the plausibility scores are more mediocre for the word pairs of this relation, therefore high-frequency adjectives can be found in both the strong positive and negative correlation groups. We also do not notice distinctive semantic classes among the verbs (head words), and neither can we make assumptions based on the animacy of the subject, since the adjective modifiers do not follow such a constraint ("new", "local", "national", "exact", "different") and the given verbs do not have the animacy constraint either ("bring", "attract" had a strong positive correlation, "increase", "reflect" a strong negative). Some verbs that do prefer animate objects were found to have a strong positive correlation (e.g. "compare", "operate"), others to have a strong negative (e.g. "lift"). Concerning the different attention scenarios, there is a noticeable positive shift in correlations (+.30, +.20) with the *context* attention mask compared to no mask or masking the head word, which hints at the influence that the verb had in the predictions, and how the context (including the one-hop dependency to the verb subject) produced less polarizing probabilities.

Finally, for the **dobj_amod** word pairs, as in the direct object word pairs, we do not notice verb grouping based on semantics or transitivity. Many of the verbs with strong positive ("teach", "promise") or negative correlation ("claim", "confirm") are verbs with varied subcategorization frames. In this syntactic category, we observe the smallest positive shifts with the use of the *context* attention mask, and even a decrease in correlation (-.03) in the micro-averaged results. However, the results still show a weak positive correlation similar to the ones of the other syntactic relations, for the most part; this observation supports the fact that the role of the verb is quite important for predictions.

5 Discussion

5.1 The limitations of our datasets

As mentioned in Section 3.2, we would like to elaborate on the issues we faced during the creation of our prompt sentences, how we dealt with them, and whether they could seriously impede our results. First of all, we noticed some problematic word pairs in the SP-10K corpus, which were included in a group with a certain syntactic relation, which they could not possess. For example, some word pairs under the verb-direct object relation included intransitive verbs such as *laugh*, *walk*, *smile*, or verbs that could not accept the dependent word as a direct object such as *look way*, *think time*, and these pairs were still assigned plausible scores when the plausibility should have been zero (e.g. *look way*, where *way* had a score of 6.5). We assume that these problematic high scores were given by naive crowdsource annotators, because distributionally there is indeed a strong correlation between the words of a pair (which makes them highly likely to co-occur in a text), but little attention was paid to the proposed syntactic relation.[3] These word pairs with problematic head or dependent words were removed from our query for sentences, in order to make sure that they were not accidentally found in a sentence with a wrong parse tree.

On the other hand, some word pairs, especially the ones which were by design of low plausibility and had a low plausibility score, are not found in the corpus, as shown in Table 1 – almost half of the word pairs for all types of syntactic relation. However, some word pairs are very common in the corpus, and are found in disproportionately more sentences. For this reason, we provided both micro- and macro-averaged results in Section 4. In addition, there are several word pairs which are parts of idiomatic, lexicalized phrases, and are very frequent in the ukWaC corpus and almost exclusively found in the context of these idiomatic phrases, but were assigned a low score. As an example, for the **nsubj** relation, in the pairs "weather permit" (4.06) and "study find" (4.0), the subjects are inanimate (whereas the verbs generally require an animate subject) but in this specific context they are acceptable. Interestingly enough, contrary to our personal intuition, BERT seems to be in accordance with the SP-10K scores, because it does assign a moderate to low probability to the dependent words, even in the *context* attention mask setting. This could imply that BERT is able to capture, to some extent, a verb's preferences and constraints, and can make predictions based on them, when the use is not metaphorical and conflicting with usual, literal cases.

[3]The Mechanical Turk workers of the SP-10K project were presented with word pairs, without any other context, and asked to evaluate the plausibility of the second word being dependent to the first with a specific syntactic relation.

5.2 Does BERT really learn selectional preference?

In our experiment, we studied how changing the way the input sequence is attended to could shift predictions and the probability of a masked word. Our goal was to observe how much the head word affected the probability of the dependent word, and whether context played a more important role than the head word itself. The fact that the highest positive correlation values almost always came from allowing attention only to the head word (and the non-lexical BERT tokens) signifies that the head word is identified as an integral and influential part of the sequence when it comes to selecting a masked word. Interestingly enough, the **nsubj_amod** and **dobj_amod** categories showed, for the most part, similar positive correlations (especially in the *context* attention block scenario) as the one-hop syntactic relations. As Zhang et al. (2019b) have also mentioned, these two-hop relations also fall under the influence of a word's selectional preferences, and the fact that the head word in these cases is the head of the sequence (the verb, in our simple prompt sentences) could have impacted the selection of a modifier to a greater extent than the context could.

6 Conclusion

In this paper, we explored whether selectional preference information makes part of the linguistic information that is learned by a Transformer model, by examining the correlation of the plausibility of a head-dependent word pair and the assigned probability of the dependent word in a sequence including the head word. Our overall results did not show a strong correlation that would definitively prove or disprove the presence of selectional preferences, but there are indications that BERT's embeddings have captured enough syntactic-semantic information to be able to assign probability based on "the right fit" for a head word. In addition, some specific cases between syntactic relations and metaphorical uses have given us the incentive to further investigate these phenomena, and to delve deeper in the architecture of the model for answers. The code we used for our experiments will be made available at a GitHub repository[4].

As for future work, a further exploration of the attention output would be quite beneficial in understanding the attention weights during prediction, in each scenario, and how they differ per layer and attention head. Researchers who have attempted to thoroughly analyze BERT's attention behavior in syntactic probing tasks have noted that the attention maps have "a fairly thorough representation of English syntax" [sic] (Clark et al., 2019), and have noticed that specific attention heads and layers specialize on learning specific linguistic knowledge, such as syntactic dependencies (Kovaleva et al., 2019). Visualizing attention for 12 layers and 12 attention heads, over 145K sentences, is a complicated task which we intend to tackle soon, with the help of visualization libraries (Vig, 2019).

Acknowledgements

This work has been funded by CNRS (80|PRIME-2019 project MoDiCLI). Experiments presented in this paper were carried out using the OSIRIM platform[5] that is administered by IRIT and supported by CNRS, the Region Midi-Pyrénées, the French Government, and ERDF. We would like to thank our reviewers for their insightful comments and suggestions.

References

Stephen Clark and David Weir. 2001. Class-based probability estimation using a semantic hierarchy. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 95–102. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 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*, pages 276–286.

[4] https://github.com/lenakmeth/bert_selectional_preferences

[5] https://osirim.irit.fr/

Andy Coenen, Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of BERT. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8594–8603. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Katrin Erk, Sebastian Padó, and Ulrike Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4):723–763.

Katrin Erk. 2007. A simple, similarity-based model for selectional preferences. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 216–223, Prague, Czech Republic, June. Association for Computational Linguistics.

Allyson Ettinger. 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.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54.

Todd R Ferretti, Ken McRae, and Andrea Hatherell. 2001. Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, 44(4):516–547.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

Yoav Goldberg. 2019. Assessing BERT's Syntactic Abilities. *arXiv preprint arXiv:1901.05287*.

Clayton Greenberg, Vera Demberg, and Asad Sayeed. 2015. Verb polysemy and frequency effects in thematic fit modeling. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*, pages 48–57.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Jerrold J. Katz and Jerry A. Fodor. 1963. The structure of a semantic theory. *Language*, 39(2):170–210.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. 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)*, pages 4365–4374, Hong Kong, China, November. Association for Computational Linguistics.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Hang Li and Naoki Abe. 1998. Generalizing case frames using a thesaurus and the MDL principle. *Computational linguistics*, 24(2):217–244.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Diana McCarthy and John Carroll. 2003. Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29(4):639–654.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July. Association for Computational Linguistics.

Ken McRae, Michael J Spivey-Knowlton, and Michael K Tanenhaus. 1998. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38(3):283–312.

Timothee Mickus, Denis Paperno, Mathieu Constant, and Kees van Deemter. 2020. What do you mean, BERT? Assessing BERT as a Distributional Semantics Model. *Proceedings of the Society for Computation in Linguistics*, 3(1):350–361.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. Maltparser: A data-driven parser-generator for dependency parsing. In *LREC*, volume 6, pages 2216–2219.

Diarmuid Ó Séaghdha and Anna Korhonen. 2012. Modelling selectional preferences in a lexical hierarchy. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 170–179. Association for Computational Linguistics.

Diarmuid Ó Séaghdha. 2010. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 435–444. Association for Computational Linguistics.

Ulrike Padó. 2007. The integration of syntax and semantic plausibility in a wide-coverage model of human sentence processing. *Doctoral thesis*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 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)*, pages 2463–2473.

Philip Resnik. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61:127–159, November.

Alan Ritter, Mausam, and Oren Etzioni. 2010. A latent dirichlet allocation method for selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Uppsala, Sweden, July. Association for Computational Linguistics.

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 104–111. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Tim Van de Cruys. 2009. A non-negative tensor factorization model for selectional preference induction. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 83–90, Athens, Greece, March. Association for Computational Linguistics.

Tim Van de Cruys. 2014. A neural network approach to selectional preference acquisition. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 26–35, Doha, Qatar, October. Association for Computational Linguistics.

Jesse Vig. 2019. A Multiscale Visualization of Attention in the Transformer Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42.

Alex Wang and Kyunghyun Cho. 2019. BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2018. Neural network acceptability judgments. *CoRR*, abs/1805.12471.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771.

Thomas Wolf. 2019. Some additional experiments extending the tech report "Assessing BERT's syntactic abilities" by Yoav Goldberg. Technical report, Huggingface Inc.

Hongming Zhang, Jiaxin Bai, Yan Song, Kun Xu, Changlong Yu, Yangqiu Song, Wilfred Ng, and Dong Yu. 2019a. Multiplex word embeddings for selectional preference acquisition. 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)*, pages 5247–5256, Hong Kong, China, November. Association for Computational Linguistics.

Hongming Zhang, Hantian Ding, and Yangqiu Song. 2019b. SP-10K: A large-scale evaluation set for selectional preference acquisition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 722–731, Florence, Italy, July. Association for Computational Linguistics.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2019c. Semantics-aware BERT for language understanding. *arXiv preprint arXiv:1909.02209*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.