Evaluating language models for the retrieval and categorization of lexical collocations

Luis Espinosa-Anke[1], **Joan Codina-Filbà**[2], **Leo Wanner**[3,2]
[1]School of Computer Science and Informatics, Cardiff University, UK
[2]TALN Research Group, Pompeu Fabra University, Barcelona, Spain
[3]Catalan Institute for Research and Advanced Studies (ICREA), Barcelona, Spain
espinosa-ankel@cardiff.ac.uk, joan.codina|leo.wanner@upf.edu

Abstract

Lexical collocations are idiosyncratic combinations of two syntactically bound lexical items (e.g., *"heavy rain"*, *"take a step"* or *"undergo surgery"*). Understanding their degree of compositionality and idiosyncrasy, as well their underlying semantics, is crucial for language learners, lexicographers and downstream NLP applications alike. In this paper we analyse a suite of language models for collocation understanding. We first construct a dataset of apparitions of lexical collocations in context, categorized into 16 representative semantic categories. Then, we perform two experiments: (1) unsupervised collocate retrieval, and (2) supervised collocation classification in context. We find that most models perform well in distinguishing light verb constructions, especially if the collocation's first argument acts as a subject, but often fail to distinguish, first, different syntactic structures within the same semantic category, and second, finer-grained categories which restrict the set of correct collocates[1].

1 Introduction

Language models (LMs) such as BERT (Devlin et al., 2018), and its variants SpanBERT (Joshi et al., 2020), ALBERT (Lan et al., 2019), RoBERTa (Liu et al., 2019), etc. have proven extremely flexible, as they behave as unsupervised multitask learners (Radford et al., 2019), and can be leveraged in a wide array of NLP tasks almost out-of-the-box (see, e.g., the GLUE and SuperGLUE results in Wang et al. (2019b) and Wang et al. (2019a), respectively). They have also been harnessed as supporting resources for knowledge-based NLP (Petroni et al., 2019), as they capture a wealth of linguistic phenomena (Rogers et al., 2020). Recently, a great deal of research analyzed the degree to which they encode, e.g., morphological (Edmiston, 2020), syntactic (Hewitt and Manning, 2019), or lexico-semantic structures (Joshi et al., 2020). However, less work explored so far how LMs interpret phraseological units at various degrees of compositionality. This is crucial for understanding the suitability of different text representations (e.g., static vs contextualized word embeddings) for encoding different types of multiword expressions (Shwartz and Dagan, 2019), which, in turn, can be useful for extracting latent world or commonsense information (Zellers et al., 2018).

One central type of phraselogical units are *lexical collocations*, defined as restricted co-occurrences of two syntactically bound lexical items (Kilgarriff, 2006), such that one of the items (the *base*) conditions the selection of the other item (the *collocate*) to express a specific meaning. For instance, the base *lecture* conditions the collocates *give* or *deliver* to express the meaning 'perform', the base *applause* conditions the selection of the collocate *thunderous* to express the meaning 'intense', and so on. Lexical collocations are of high relevance to lexicography, NLP and second language learning alike, and constitute a challenge for computational models because of their heterogeneity in terms of idiosyncrasy and degree of semantic composition (Mel'čuk, 1995).

In this paper, we analyze a suite of LMs in the context of two tasks that involve lexical collocation modeling. First, **unsupervised collocate retrieval**, where we *mask* a collocation's collocate (e.g., *"heavy"* in *"heavy rain"*), and quantify how well a LM of choice (BERT in particular) predicts, via its masked language modeling (MLM) objective, a valid collocate for that particular base ({*"heavy"*, *"torrential"*, *"violent"*, ...} for the base *"rain"* and the meaning *intense*). Second, **su-**

[1] The resources associated with this paper are available at https://github.com/luisespinosaanke/lexicalcollocations.

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 1406–1417
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

pervised in-context collocation categorization, where we fine-tune LMs on the task of predicting a semantic category of a collocation in terms of its *lexical function* (LF), given its sentential context; cf. Section 3.1. Modeling, recognizing, and classifying collocations in corpora has obvious applications for automatically creating and expanding lexicographic resources, as well as for various downstream NLP applications, among them, e.g., machine translation (Seretan, 2014), word sense disambiguation (Maru et al., 2019), or natural language generation (Wanner and Bateman, 1990). The two main contributions of this paper thus are:

1. A "collocations-in-context" dataset, with instances of collocations of 16 different semantic categories (in terms of LFs) in context, and with a fixed and lexical (i.e., no overlapping) train/dev/test split (Section 3).

2. An evaluation framework for assessing the degree of compositionality of lexical collocations, pivoting around two tasks: unsupervised *collocate* retrieval (Section 4) and in-context *collocation* categorization (Section 5).

Our results suggest that modeling collocations in context is a challenge, even for widely used LMs, and that this is particularly true for less semantic (and thus less compositional and more idiosyncratic) collocations. We also find that jointly recognizing the semantics and the syntactic structure (e.g., whether the collocate acts as subject or object in verbal constructions) of a collocation also constitutes non-trivial challenges for current architectures. Moreover, as a byproduct of our analysis, we also find an interesting behaviour in LMs when modeling antonymy in adjectives, specifically that their representations undergo substantial transformations as they flow through BERT's transformer layers, with many contextualized embeddings clustered together in the tip of a narrow cone that seems to represent adjectives in collocations denoting intensity (*"heavy"* rain) and weakness (*"minor"* issue).

2 Related Work

In this section, we discuss related works in two methodological areas that are relevant to this paper, namely conditioning MLMs (Section 2.1) and recognition of multiword epressions (MWEs) (Section 2.2).

2.1 Conditioning MLMs

A Masked Language Model (MLM) can be used as a *proxy* for gaining insights into how language is encoded by the weights of the (usually transformer-based) LM architecture. Moreover, simply asking an LM to predict words in context (without task-specific fine-tuning) has proved useful in NLP applications dealing with lexical items (affixes, words or phrases). For example, Wu et al. (2019) use BERT's MLM for augmenting their training data in sentiment analysis tasks; Qiang et al. (2019) use BERT for lexical simplification by *conditioning* the predictions over the [MASK] token by providing the original sentence as context; and Zhou et al. (2019) obtain SotA results in lexical substitution by conditioning BERT via embedding dropout on the target (unmasked) word. Inspired by the findings in these works (especially Qiang et al. (2019)), we will explore the predictions of BERT over masked lexical collocations (with and without conditioning) in Section 4, with the aim to understand whether these predictions can be used to measure the idiosyncrasy of the underlying semantics of a lexical collocation, i.e., whether the restrictions imposed by a collocation's base are due to the frozenness of the phrase itself or, on the contrary, sentential context is neccessary.

2.2 Distributional Lexical Composition

Building representations that account for non-compositional meanings within the broader spectrum of encoding semantic relations between words is a long-standing problem in computational semantics (Baroni and Zamparelli, 2010; Mitchell and Lapata, 2010; Boleda et al., 2013). Interestingly, there seems to be little agreement on how these representations should be defined, with recent attempts focusing on verbal multiword expressions (see an overview of approaches in Ramisch et al. (2018)), phrases of variable length encoded via LSTMs, based on their definitions (Hill et al., 2016), or arbitrary lexical and commonsense relations between word pairs for downstream NLP. As a testimony of the broad methods explored in the most recent literature, let us refer to, for instance, the combination of word vector averages with conditional autoencoders (Espinosa-Anke and Schockaert, 2018), expectation maximization (Camacho-Collados et al., 2019), LSTMs for predicting word pair contexts (Joshi et al., 2019), and explicit encoding of generalized lexico-syntactic patterns (Washio and Kato,

LF	BERT input	Masked sentence	Pred. collocates	Orig. collocate
Oper1	Masked sent	iran feared that the u.s and israel may [MASK] an air raid on its controversial nuclear facilities	perform, conduct, mount,	launch
	Orig sent [SEP]	iran feared that the u.s and israel may launch an air raid on its controversial nuclear facilities [SEP]	launch, conduct, order	launch
	Masked sent	iran feared that the u.s and israel may [MASK] an air raid on its controversial nuclear facilities		
Real1	Masked sent	if that happens, lindsey will use explosive triggers to [MASK] the landing gear	destroy, stop, remove	lower
	Orig sent [SEP]	if that happens , lindsey will use explosive triggers to lower the landing gear [SEP]	destroy, stop, trigger	lower
	Masked sent	if that happens , lindsey will use explosive triggers to [MASK] the landing gear		
Magn	Masked sent	the EU is driving a [MASK] bargain over Swiss demands for greater access to EU airspace .	plea, hard, new	hard
	Orig sent [SEP]	the EU is driving a hard bargain over Swiss demands for greater access to EU airspace . [SEP]	plea, hard, tough	hard
	Masked sent	the EU is driving a [MASK] bargain over Swiss demands for greater access to EU airspace .		

Table 1: Illustrative behaviour of BERT when prompted to predict a collocate in the position of a masked token, for three LFs: Oper1 (‘launch an air raid’), Real1 (‘lower the landing gear’) and Magn (‘hard bargain’), under two settings, when not conditioned (Masked sent) and when conditioned on the original (unmasked) sentence (Orig sent [SEP] Masked sent).

2018). Parting ways with the above works, in this paper we will follow the experimental setting described in Shwartz and Dagan (2019), based on injecting sentential contexts into multiword expressions (in our case, only lexical collocations) to leverage the contextual nature of current LMs. However, our goal is not to compare different combinations of feature-extraction and training/fine-tuning methods, but rather to understand lexical collocations' learnability, idiosyncrasy and their internal vector-space representations.

3 Data and Resources

3.1 Lexical Collocations

Let us first introduce the notion of lexical collocation and LF. The term *collocation* has been used in computational linguistics research to denote two different concepts. On the one hand, following Firth (1957), Church and Hanks (1989); Evert (2007); Pecina (2008) and others, a collocation has been assumed to be a combination of words that have the tendency to occur together in discourse. Typical examples are *doctor – hospital*, *mop – bucket*, *real – estate*, *look – for*, etc. On the other hand, for instance, Wanner et al. (2006); Gelbukh and Kolesnikova. (2012); Rodríguez Fernández et al. (2016); Garcia et al. (2017) adopt the definition that is common in lexicography and phraseology (Hausmann, 1985; Cowie, 1994; Mel'čuk, 1995), according to which, a collocation is an idiosyncratic combination of two lexical items, the base and the collocate, as defined above in Section 1. This interpretation states that collocations are phraseological units, although their degree of compositionality can vary. For instance, *win* [*a*] *war* is perceived to possess a higher degree of (free) composition than, e.g., *hold* [*a*] *meeting*, and *heavy rain* is less compositional than [*a*] *well-justified argument*. We adopt this definition of the notion of collocation, and in order to avoid any confusion, we refer to it, following Krenn (2000), as *lexical collocation*.

Lexical collocations can be typified with respect to the meaning of the collocate and the syntactic structure formed by the base and the collocate. LFs provide a fine-grained typology of this kind (Mel'čuk, 1996). An LF can be considered a function $f(L)$ that delivers for a base L a set of synonymous collocates that express the meaning of f. Where pertinent, f also codifies the subcategorization structure of the base+collocate combination. LFs are assigned Latin acronyms as names; cf., e.g., "Oper1" (‘operare’), which means ’perform’ and realizes the first argument of the base as subject: Oper1(*lecture*) = {*deliver*, *give*, *hold*}; "Magn" (‘magnum’), which stands for ‘intense’: Magn(*applause*) = {*thunderous*, *loud*, ...}.[2]

The encoding of LFs in NLP research has in recent years revolved around applying word embeddings-based techniques, e.g., in terms of linear projections (Rodríguez Fernández et al., 2016) and semantic generalizations (Espinosa-Anke et al., 2016). Recently, Shwartz and Dagan (2019) analyzed, from the perspective of "static" vs. "contextualized" representations and their applicability to studying compositional phenomena like "meaning shift", one specific type of lexical collocations, namely *light verb constructions* (LVCs), which are well illustrated by the LFs Oper1 and Oper2 (and also, partially, by Real1 and Real2). While we find that the above research directions (i.e., embeddings-based and contextualized representations for modeling MWEs) are complementary, in this work, we

[2] For simplicity we will write Oper1(*lecture*) = *deliver*, etc.

LF	semantic gloss	example
Oper1	'perform'; 1st argument → subject	Oper1(*support*) = *lend*
IncepOper1	'begin to perform'; 1st argument → subject	IncepOper1(*impression*) = *gain*
Oper2	'undergo'; 2nd argument → subject	Oper2(*support*) = *find*
Real1	'realize'; 1st argument → subject	Real1(*accusation*) = *prove*
Real2	'apply'; 2nd argument → subject	Real2(*support*) = *enjoy*
AntiReal2	'fail to apply'; 2nd argument → subject	AntiReal2(*war*) = *lose*
CausFunc0	'cause the existence'	CausFunc0(*hope*) = *raise*
Caus1Func0	'cause the existence; 1st argument'	Caus1Func0(*hope*) = *gain*
LiquFunc0	'cause termination of the existence'	LiquFunc0(*hope*) = *destroy*
IncepPredPlus	'increase'	IncepPredPlus(*temperature*) = *rise*
Magn	'intense'	Magn(*smoker*) = *heavy*
AntiMagn	'little', 'weak'	AntiMagn(*smoker*) = *occasional*
Ver	'genuine'	Ver(*demand*) = *legitimate*
AntiVer	'non-genuine'	AntiVer(*demand*) = *illegitimate*
Bon	'positive'	Bon(*performance*) = *good*
AntiBon	'negative'	AntiBon(*performance*) = *poor*

Table 2: LFs used in this paper. The 'semantic gloss' column provides both a definition and the actantial structure, which is required in cases where one LF may express the same semantics but with a different syntactic structure (e.g., Real1 vs. Real2).

specifically focus on the existing (and learnable) knowledge LMs have concerning lexical collocations, and whether they can be used to recognize and categorize LFs in free text.

For our experiments, we use, as initial lexical collocation source, a **collocations dataset**, LexFunc (Espinosa-Anke et al., 2019), which we have extended to cover a wider range of LFs (listed in Table 2). The original LexFunc dataset and this extended version are both the result of an initial collection of collocations categorized into LFs made available by Igor Mel'čuk. Each collocation has been manually lemmatized, and bases and collocates have been manually annotated with part-of-speech tags and their syntactic dependency relation.

With the lexical collocations of the extended LexFunc dataset at hand, we first compile from the English Gigaword[3] a **collocations corpus**, which contains the occurrences of these lexical collocations. In principle, the identification of a given collocation in corpora is a straightforward procedure, as we know its elements (base and collocate) and the syntactic dependency relation between them. However, automatic dependency parsing is far from perfect, which complicates the task. Therefore, and in order not to lose any relevant collocation occurrence in the GigaWord corpus, we apply a cascaded procedure for their identification on the lemmatized and POS- and head-modifier relation tagged GigaWord.[4] In the first stage, we identify sentences in which between the collocation elements in question one of the relevant syntactic dependency relations has been identified. In the second (more relaxed) stage, we match adjacent lemmatized collocation elements and their PoS tags. In the third stage, finally, we match lemmatized collocation elements and their PoS tags within a distance of up to 5 tokens. While this procedure inevitably introduces some noise (we might retrieve sentences where base and collocate co-occur, but not as a collocation), we performed a manual inspection on a random sample, and calculated precision of our collocation retrieval strategy, which resulted in >0.95. This confirms the quality of our retrieval strategy, and hence, our resource.

In terms of corpus statistics, Table 3 indicates the number of sentences for each LF distributed across training (70% of the sentences), development (15%) and test (15%) sets. The split was done maintaining this proportion across all LF. Note that these splits are constructed such that there are no overlapping collocations, in an effort to avoid the well-known phenomenon of *lexical memorization* (Levy et al., 2015), which may artificially inflate the results on the test set. The number of different collocations per split, globally and for each LF, also maintains the same proportions (70/15/15 ±1%), such that, e.g., AntiReal2 has 55 different collocations in the 942 sentences of the training set and 11 different collocations in the development and test sets, distributed across 205 and 200 sen-

[3] Gigaword contains newswire material, a register balancing the staticity of encyclopedic English (e.g., Wikipedia) and the noisiness of user-generated text, and is a rich source of lexical collocations in well-formed grammatical contexts (Rodríguez Fernández et al., 2016) `https://catalog.ldc.upenn.edu/LDC2011T07`.

[4] We preprocess GigaWord with the NLP4J parser `https://emorynlp.github.io/nlp4j/`.

Label	Train	Dev	Test	Total
Magn	28,748	6,113	6,164	41,025
Oper1	11,746	2,517	2,493	16,756
Real1	3,481	746	743	4,970
AntiMagn	2,959	638	649	4,246
IncepOper1	2,489	541	539	3,569
Oper2	2,408	515	520	3,443
AntiVer	1,874	397	405	2,676
AntiBon	1,815	385	393	2,593
CausFunc0	1,714	370	367	2,451
Real2	1,570	336	337	2,243
Bon	1,471	298	315	2,084
LiquFunc0	1,398	301	297	1,996
AntiReal2	942	203	200	1,345
Ver	926	198	196	1,320
Caus1Func0	686	151	149	986
IncepPredPlus	624	147	138	909
Total	64,851	13,856	13,905	

Table 3: Statistics (in number of sentences) of our *collocations in context* dataset (ordered by frequency).

tences respectively. In the overall corpus, there is an average of 18 samples per collocation (*work hard* being the most frequent one with 102 samples). *Hope*, *attack*, *criticism*, *fire* and *thread* are bases that each co-occur with more than 30 different collocates, across most LF. These bases are also among the ones with more samples in the corpus. On the other side, half of the bases are combined with one single collocate only. Overall, the statistical properties of our dataset arguably make it a faithful replica of the distribution of collocations in, at least, newswire corpora. At the same time, it is a challenging dataset, as the results we report in this paper suggest.

4 Experiment 1: Collocate Retrieval

4.1 Setup

In the first experiment, we aim to analyze how well an MLM retrieves valid collocates for a given base when being provided with the original (sentence-level) context. We use BERT (`bert-base`) (Devlin et al., 2018), as it is the *de-facto* model on top of most specialized and distilled/quantized language models. Its behaviour should thus be a good proxy for the general distributional behaviour of lexical collocations. This experiment serves, first, as an opportunity to understand *how much* semantics that is underlying LFs can be encoded via a MLM pretraining objective, and second, as a testbed for exploring conditioning strategies often used in tasks involving data augmentation and lexical substitution and simplification (cf. Section 2.1). Since this is an "in-context collocate retrieval" task, we consider it a ranking problem. Intuitively, if BERT is able to retrieve a base's valid collocates (e.g., {*heavy*, *torrential*, *violent*, ...} for *rain* as base for Magn) in the position of a masked token, this could mean that: (1) the sentence is giving enough context for the model to "know" the lexical restrictions involved in that collocation, and/or (2) the LF is sufficiently frozen, and therefore the base alone may restrict which collocates are acceptable. For the first point, and continuing with the *heavy rain* example, consider the following sentence.

(1) *Hurricane Katrina brought* [MASK] *rain to Louisiana.*

Intuitively, we would expect the sentential context to be informative enough for the model to select *heavy* or any other collocate denoting the notion of intensity, and restricted by the presence of the base *rain*. In fact, here, BERT predicts *heavy* with 79.5% probability. However, in example (2)

(2) *Policeman earns applause for staying on duty in* [MASK] *rain.*

there is much lesser evidence for the *rain* to be 'intense', and in fact BERT predicts here *'the'* with 85.1% probability. This disparity lets us investigate ways to prompt BERT to select *heavy* or any other valid collocate for example (2). Thus, in addition to simply passing one masked sentence, we explore an approach based on passing the masked sentence concatenated with the original sentence, which is a natural way to encode not only the context surrounding the word, but also the meaning of the target word itself. This strategy was successfully used for the task of unsupervised lexical simplification (Qiang et al., 2019). For the second point above, the works of Espinosa-Anke et al. (2019); Shwartz and Dagan (2019) already point to the fact that light verb constructions (LVCs) are easy to recognize in text. Further evidence is provided in Table 1, which shows BERT's top predictions for three sentences containing Oper1, Real1 and Magn collocations. It is immediately obvious that the nature of the LF itself, as well as the amount of the information provided by the sentential context, are crucial. Note that, in the case of Oper1, by providing the original sentence as context, BERT's grasp of the LF improves, as it tends to predict the correct collocate, whereas for Real1 or Magn, this improvement results in assigning higher probability to tokens which are more similar to the LF's abstract meaning (e.g., *trigger the landing gear* for Real1 or *tough bargain* for Magn).

With the above considerations in mind, we run BERT's MLM head over our GigaWord-based test set, and compute *Mean Reciprocal Rank* (MRR) and *Mean Average Precision* (MAP) over each LF. Recall that we consider as valid *hits* all the collocates for a given base and its corresponding LF. In practice, this means that for bases that have just one valid collocate, both metrics yield the same score. We lemmatize BERT's predictions using SpaCy's lemmatizer.[5]

4.2 Results and Discussion

Our results (Table 4) show, first, that conditioning BERT's MLM by passing the original sentence as additional context for the [MASK] token is useful for predicting an embedding whose semantics is more related to the original collocate. The improvements are particularly relevant for LVCs (Oper1, and, to a certain extent, Real1 and Real2), suggesting that these LFs, while perhaps easy to distinguish from others (cf. Section 5.1), they do benefit from additional contexts to be well represented. Interestingly, the Magn LF has small gains in both MRR and MAP, clearly showing that additional context helps little, and thus highlighting a strong semantic dependency between sentence meaning and the collocation's base.

A potential limitation of this setup, however, is that we cannot possibly include all possible collocates for all the bases in our resource. An estimate of the quality of BERT's predictions can be obtained by measuring the semantic similarity (for instance, by cosine distance) between the original masked collocate and the predicted collocates. In the example we already referred to above, *heavy rain*, the similarity between 'the' and 'heavy' is low, whereas, if the model predicts *hard* or even any other adjective, it should be considered less wrong. We obtain a broad picture of the quality of BERT's predictions by plotting a histogram (Figure 1) of the similarities obtained by comparing the original collocate's and BERT's predicted GloVe embeddings (Pennington et al., 2014) under both settings (MASKED and CONDITIONED) for the same three LFs as in Table 1), namely Magn, Oper1 and Real1. The conditioning strategy is helpful; it contributes not only to retrieving the original collocate (which would be trivial if we do not mask it), but also candidates with clearly similar meanings. We see, for instance, more cases for Oper1 and Real1, where the correct verb is predicted, whereas for Magn we see a more sustained improvement across all similarities, but not necessarily for retrieving the original collocate.

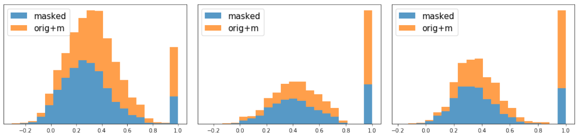

Figure 1: Histograms showing the distribution of similarities between gold and predicted (lemmatized) collocates for the three LFs Magn, Oper1 and Real1 (from left to right).

	MASKED		CONDITIONED	
	MRR	MAP	MRR	MAP
AntiBon	13.78	13.57	61.64	57.37
AntiMagn	30.16	27.11	82.18	66.89
AntiReal2	39.14	36.32	70.58	62.13
AntiVer	11.38	10.68	40.5	37.12
Bon	25.13	24.87	62.49	59.34
Caus1Func0	52.77	45.93	95.94	80.36
CausFunc0	61.14	53.53	86.2	73.52
IncepOper1	54.38	48.36	88.52	74.61
IncepPredPlus	7.27	6.375	12.27	10.67
LiquFunc0	45.51	43.12	71.14	63.60
Magn	33.43	31.67	74.72	66.92
Oper1	73.12	62.58	95.22	81.15
Oper2	63.84	53.86	93.63	76.86
Real1	59.87	53.31	90.96	75.02
Real2	55.36	46.87	76.60	64.64
Ver	33.06	29.37	72.02	63.18

Table 4: Results of the collocate retrieval experiment when passing to the MLM the masked sentence (MASKED) or with the original sentence as context (CONDITIONED).

5 Experiment 2: Collocation categorization

In the second experiment, we test the performance of a number of well-known LMs for the task of LF categorization using the train/test splits we sampled and annotated from GigaWord (Section 3). This experiment serves two purposes. First, we expect to learn about the *predictability* of LFs in context,

[5] https://spacy.io/api/lemmatizer.

which is a long-standing problem in computational lexicography and the cornerstone of automatic construction of collocation resources. Second, previous work has shown that some LFs are quite easy to distinguish, without (Espinosa-Anke et al., 2019) and with sentential context (Shwartz and Dagan, 2019). However, it is still unclear whether, by focusing exclusively on the phenomenon of collocations, and excluding, e.g., idiomatic expressions or non-compositional phrasal verbs (which are not only semantically but, more importantly, syntactically different from collocations), an LM can indeed be used to construct a resource for second language learners, or whether (and to what extent) an LM can be trained to select appropriate collocates. Our setting is essentially a sentence-pair classification problem, where the second sentence is the lexical collocation itself. Specifically, a training instance is a tuple <sentence, collocation, label>, as in Example (3) (where we use **;** as a wildcard concatenation token, as these are different across LMs):

(3) *The US military **launched an airstrike** on what it described as a safehouse in the Iraqi town of Fallujah* **;** *launched an airstrike* → ***Oper1***

We use as labels the LFs listed in Table 2, with their respective training/test splits, and train all LMs with the same hyperparameters.[6] The considered LMs are BERT (base and large, uncased) (Devlin et al., 2018), RoBERTa (base and large) (Liu et al., 2019), DistilBERT (Sanh et al., 2019), ALBERT (Sanh et al., 2020) and XLNet (base and large) (Yang et al., 2019). We use the implementation in the Transformers Python library (Wolf et al., 2020).[7]

5.1 Results and discussion

The results of this experiment (cf. Table 5) clearly highlight what has already pointed out in Shwartz and Dagan (2019): the prototypical LVCs (as modeled by Oper1) can be identified with a rather high quality. Interesting enough, this is not true for LVCs captured by Oper2, whose only difference to Oper1 is the subcategorization frame: while in Oper1, it is the 1st argument of the base that is realized as the grammatical subject, in Oper2, it is the 2nd argument. Frequency cannot explain this discrepancy since, e.g., IncepOper1, which appears in our corpus in nearly the same number of sentences as Oper2, is categorized with a significantly higher quality.

Results are also lower for some other verbal LFs with more semantic load, among them, e.g., Real1/2 and Caus1Func0, suggesting that the semantics expressed in the notions of 'realize' and 'cause', especially when the 2nd argument of the collocation functions as a subject, are more challenging. Again, these results cannot be fully explained by the amount of training data and neither by the semantic load. Thus, the categorization of IncepPredPlus achieves the highest score (the best model on IncepPredPlus obtains an average F1 of 95.21 with little variability across runs), and it is clearly an LF with a semantic load, namely 'increase'. Interestingly, Ver ('genuine') and Bon ('positive') are the worst categorized LFs in our sample, while their antonyms AntiVer and AntiBon are categorized considerably better.

As for the considered LMs, the best overall performing model is the RoBERTa family, with an overall F1 score of **71.19%** for RoBERTa-base and 70.6% for RoBERTa-large, and both models accounting for the best results on 7 of the 16 target LFs. The second best results are achieved by XLNet (base and large), with XLNet-large being the best model on both Magn and AntiMagn, two LFs which have been traditionally challenging to tell apart due to the fact that the representations of antonyms are clustered together in distributional spaces. We also note that, interestingly, DistilBERT is the best at categorizing Oper1 and Real2, which may suggest that small models may be sufficient to obtain good performnace on categorizing LVCs.

In order to gain further insights on *why* a LM may err in the task of in-context collocation categorization, we display a confusion matrix obtained from random runs for the two LMs with the highest avg score for Oper1 (Distilbert) and Oper2 (XLNet-base) (Figure 2) – the two LVC LFs that differ only in terms of their subcategorization patterns (cf. above). We may hypothesize that the categorization of a collocation based mainly on its actantial structure is challenging, and indeed, we observe that for these two models,[8] syntax-based categorization over the same semantics proves hard. Specifically, XLNet-base has as the greatest source for confusion regarding Oper1, precisely, Oper2;

[6] All models are trained for 1 epoch, with a learning rate of $4e-05$, the Adam with weight decay optimizer and a warmup ratio of 0.06.

[7] `https://huggingface.co/transformers/`.

[8] As a matter of fact, it can be assumed that this applies to all the LMs evaluated in this paper.

and this also occurs with Real1 vs. Real2 (which also differ only with respect to their subcategorization pattern). The results for DistilBERT show a greater spread among the misclassifications of Oper1, namely across Oper2, Caus1Func0 and IncepOper1, and for Oper2 across Oper1 and CausFunc0. Caus1Func0 and IncepOper1 have the same subcategorization pattern as Oper1 (but different semantics). In the case of CausFunc0 ('cause the existence', e.g., CausFunc0(*hope*) = *raise*), the subcategorization pattern is very similar to Caus1Func0, only that the grammatical subject of the corresponding syntactic construction is not an argument of the base. As we can observe, CausFunc0 is easily miscategorized as a full LVC. Finally, let us highlight the fact that while Magn is generally well categorized, the few misclassifications come, as would be expected, from collocations which convey a similar notion of amplification (e.g., Bon), but interestingly, also collocations that convey opposite semantics, such as AntiMagn or AntiBon.

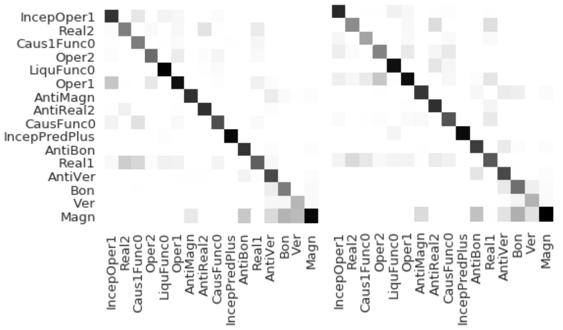

Figure 2: Confusion matrix for the two best performing models on average on Oper2 (Xlnet-base, left) and Oper1 (DistilBERT, right).

6 Subspace Analysis

In this section, we further explore the semantics of some selected LFs. We generate visualizations of PCA-projected BERT vectors for all collocation mentions of Magn, AntiMagn, Oper1 and Oper2. These four LFs are sufficiently frequent, and they encode different morphosyntactic structures.[9] We can see that antonymy (Ono et al., 2015; Schwartz et al., 2015; Nguyen et al., 2016) is relatively well captured in contextualized models, although the subspaces are clearly different between the embedding and the last transformer layer. More specifically, as the representations of collocates for Magn and AntiMagn undergo the self-attention-based transformations through BERT's layers, many of these contextualized embeddings tend to group in a narrow cone, with many antonymic collocates indistinguishably overlapping with each other. Similarly, we also observe a tendency of representation overlap in the Oper1 vs Oper2 case, with the embeddings in the last transformer layer showing a cluttered distribution, suggesting that there is little inherent knowledge in BERT to categorize a collocation into the syntactic typification of a LF.

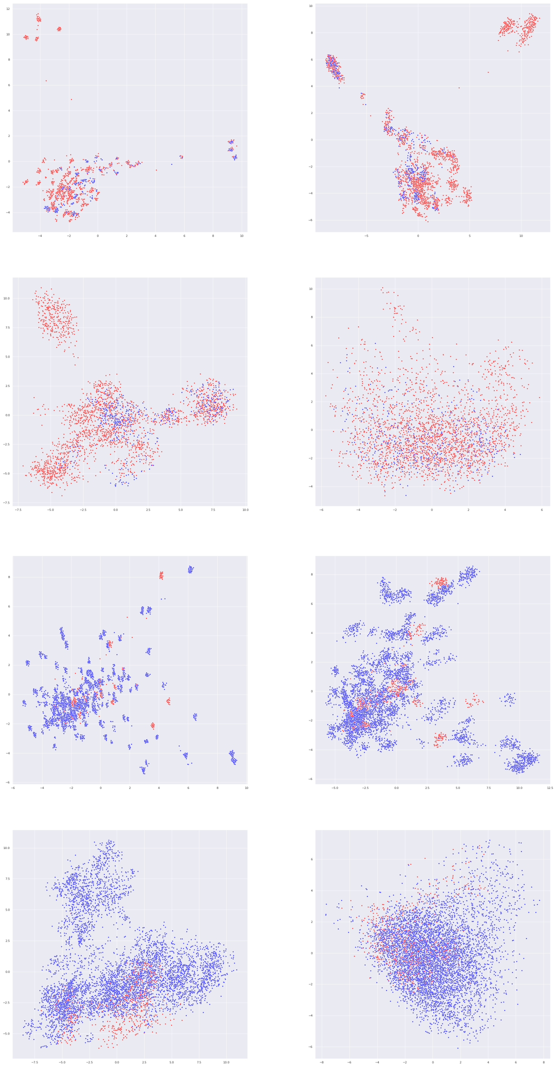

Figure 3: Oper1 (red) and Oper2 (blue) collocate embeddings for BERT's embedding layer (top row, left), and for the 1st (top row, right), and 5th and 12th transformer layers (second row, left and right, respectively). The bottom quadrant corresponds to Magn (blue) vs AntiMagn (red), with the same arrangements (embedding, 1st, 5th and 12th layer).

[9] <Magn,AntiMagn> are most frequently expressed by *adj+noun* combinations, whereas <Oper1,Oper2> are always realized by a *verb+noun* pattern.

	BERT-base			BERT-large			Albert			DistilBERT		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
AntiBon	67.92±5.99	72.01±4.03	69.85±4.67	71.73±11.50	72.43±3.03	71.81±6.29	77.00±3.90	75.31±1.66	76.11±2.21	69.89±5.13	70.90±4.55	70.35±4.42
AntiMagn	85.59±3.30	78.22±4.83	81.60±1.14	85.05±2.93	74.93±3.54	79.66±3.26	84.55±1.90	74.83±0.69	79.38±0.58	85.29±3.56	75.55±2.09	80.10±2.38
AntiReal2	81.67±1.48	78.50±2.29	80.02±0.56	84.74±3.38	78.50±3.50	81.40±0.77	81.39±12.70	82.50±4.33	81.58±6.90	71.23±2.33	79.16±1.89	74.98±2.02
AntiVer	61.32±4.03	72.51±4.82	66.28±1.91	61.21±2.44	71.76±6.26	66.01±3.67	62.35±3.00	79.25±1.28	69.78±2.37	63.33±2.20	73.33±10.40	67.63±3.18
Bon	50.95±4.34	43.06±3.96	46.54±2.64	53.52±9.98	56.71±4.09	54.91±6.98	55.23±7.97	49.52±8.27	52.19±8.06	54.52±13.50	51.64±10.80	**53.02±12.10**
Caus1Func0	67.18±5.53	46.97±15.10	**53.77±10.10**	51.49±7.12	41.83±9.77	45.47±5.95	59.19±3.67	43.62±7.56	50.13±6.25	64.21±10.60	42.05±8.06	50.51±8.25
CausFunc0	72.37±10.20	59.30±7.90	64.41±2.97	71.10±7.40	58.21±6.02	63.75±4.08	78.63±7.92	60.39±5.61	68.07±3.99	74.12±2.79	65.57±1.59	69.56±1.52
IncepOper1	79.30±3.50	76.43±1.95	77.82±2.28	78.44±7.49	75.13±5.46	76.40±1.19	81.60±4.03	76.87±1.31	79.14±2.53	79.61±1.22	77.55±4.01	78.53±2.39
IncepPredPlus	97.21±3.03	91.06±0.41	94.02±1.63	93.85±6.16	90.57±1.91	92.08±2.32	97.73±2.21	85.99±6.69	91.43±4.68	98.46±2.01	90.57±0.72	94.34±0.68
LiquFunc0	88.02±6.34	91.02±2.36	89.43±3.96	88.35±3.32	87.54±2.75	87.94±2.94	88.70±2.86	91.80±5.57	90.10±1.22	85.72±2.55	87.76±1.40	86.70±0.93
Magn	92.47±0.75	93.91±0.77	93.18±0.32	92.87±0.80	93.67±2.23	93.26±1.50	92.74±0.52	93.78±0.93	93.26±0.62	92.80±1.16	94.58±0.32	93.68±0.74
Oper1	78.89±1.08	91.69±0.92	84.80±0.27	78.01±1.94	90.26±1.54	83.69±1.76	78.69±0.36	92.77±2.82	85.14±1.18	79.58±1.50	91.89±1.94	**85.28±1.27**
Oper2	70.16±2.87	48.07±6.46	56.79±3.97	66.19±1.82	52.17±5.14	58.24±3.04	71.03±9.65	51.98±2.25	59.85±3.83	67.82±4.49	48.01±2.22	56.17±2.34
Real1	70.03±2.29	59.71±4.38	64.44±3.48	69.91±2.31	59.21±2.24	64.12±2.24	70.52±4.10	62.18±1.63	66.02±1.60	70.81±8.46	62.44±4.33	66.25±5.57
Real2	58.36±2.86	53.41±7.12	55.48±3.12	64.99±5.87	54.79±5.44	59.13±1.98	63.52±2.53	47.47±8.75	54.14±6.64	69.69±0.79	54.30±12.80	**60.51±8.24**
Ver	40.15±7.96	24.31±3.82	30.23±4.92	37.59±12.30	22.27±2.90	27.38±4.40	31.54±16.10	22.61±11.60	25.95±13.40	43.15±25.20	20.91±9.93	28.07±14.50
Average	72.60±4.10	67.51±4.45	69.29±3.00	71.82±5.42	67.50±4.11	69.08±3.22	73.40±5.21	68.18±4.43	70.14±4.13	73.14±5.47	67.89±4.82	69.73±4.41

	XLNet-base			XLNet-large			RoBERTa-base			RoBERTa-large		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
AntiBon	67.74±9.42	74.72±2.49	70.88±6.14	70.30±6.48	76.75±3.99	73.25±4.11	71.57±11.90	75.82±9.03	73.51±10.20	77.44±10.10	76.42±10.10	**76.82±10.00**
AntiMagn	85.26±5.43	76.52±2.05	80.62±3.36	86.18±5.10	77.86±4.67	**81.78±4.62**	90.02±4.44	77.19±3.90	83.10±3.90	84.52±9.63	75.86±9.63	79.84±5.16
AntiReal2	81.96±3.72	77.83±2.92	79.83±3.16	85.06±10.40	78.83±1.60	**81.65±5.21**	81.72±12.30	80.00±2.78	80.61±6.89	80.64±3.95	82.66±3.95	81.64±4.08
AntiVer	65.29±2.14	67.57±1.99	66.37±0.36	67.60±3.54	70.20±6.34	68.86±4.91	69.37±1.08	75.39±9.19	**72.11±4.89**	69.93±9.95	72.92±9.95	71.18±5.85
Bon	51.70±6.46	47.40±2.16	49.41±4.10	52.08±15.90	47.51±3.42	49.24±8.88	53.95±7.90	46.87±2.06	50.01±4.13	50.84±7.58	54.70±7.58	52.70±7.85
Caus1Func0	64.71±9.52	44.29±15.40	50.52±9.85	64.46±24.90	40.93±11.70	47.48±5.79	68.68±12.20	47.87±22.40	52.47±15.40	59.92±11.70	42.05±11.70	49.41±10.20
CausFunc0	75.65±9.76	59.58±7.49	66.35±6.80	79.15±14.70	60.39±2.38	68.17±5.83	78.06±6.82	67.21±2.45	**72.15±3.85**	75.56±8.38	63.30±8.38	68.81±6.47
IncepOper1	73.37±3.87	78.16±1.08	75.65±1.94	79.35±4.77	75.20±2.35	77.13±1.86	79.38±4.43	79.03±1.76	79.16±2.54	83.29±8.40	76.49±8.40	**79.61±4.24**
IncepPredPlus	99.21±0.01	91.54±2.32	**95.21±1.26**	97.89±2.00	89.37±2.21	93.43±1.88	96.31±4.49	90.33±0.83	93.18±1.71	92.79±12.40	91.06±12.40	91.58±6.04
LiquFunc0	87.38±2.63	92.48±4.18	89.83±2.63	87.95±4.60	91.91±6.18	89.68±1.09	88.75±0.35	93.04±4.01	90.82±1.89	90.15±2.41	91.58±2.41	**90.86±2.81**
Magn	91.93±0.85	93.91±0.48	92.91±0.54	93.41±1.66	94.24±1.66	**93.82±1.62**	92.57±1.31	94.95±1.60	93.74±1.43	92.90±1.29	94.73±1.29	93.81±1.68
Oper1	77.80±0.97	91.09±0.20	83.92±0.48	78.90±0.94	91.80±0.78	84.86±0.22	78.33±2.36	91.18±1.78	84.25±1.46	78.41±1.85	92.40±1.85	84.81±1.79
Oper2	72.13±8.13	52.69±3.66	**60.85±5.12**	74.05±4.18	50.70±7.88	59.86±4.83	72.05±1.20	50.70±8.57	59.18±5.79	70.19±5.83	52.56±5.83	59.97±1.06
Real1	72.00±1.43	59.84±0.99	65.36±0.96	66.09±3.48	64.60±4.19	65.33±3.82	71.11±4.52	61.05±4.51	65.66±4.10	72.59±4.37	62.49±4.37	**66.90±5.70**
Real2	66.37±1.17	51.23±4.45	57.77±3.13	64.01±12.30	54.20±7.40	58.67±9.52	66.12±3.52	52.91±8.66	58.67±6.65	64.58±7.64	51.03±7.64	56.92±9.77
Ver	32.58±12.90	15.98±10.50	21.37±12.90	32.94±19.20	26.70±22.00	29.21±21.10	48.10±29.10	22.78±5.97	**30.38±11.30**	41.46±4.40	18.02±4.40	24.77±3.23
Average	72.82±4.90	67.18±3.90	69.18±3.92	73.71±8.38	68.20±5.55	70.15±5.33	75.38±6.75	69.15±5.59	**71.19**±5.38	74.08±6.87	68.64±6.87	70.60±5.37

Table 5: Average Precision, Recall and F1 results for the collocation classification experiment, computed by averaging the results of three independent runs. We also report standard deviation figures. Results are provided per LF as well as the average over each metric (Average).

7 Conclusions

We have analyzed LMs in tasks revolving around modeling, recognizing and categorizing lexical collocations. We conclude that some prominet types of LVCs require little context to be well encoded, as opposed to other LFs involving, e.g., nouns and adjectives, and that predictability of LFs is challenging, not a function of training data, and that syntax plays a major role.

8 Future Work

In the future, we will make this work multilingual using linguistic equivalences as anchors, in the spirit of cross-lingual embedding research, in order to align collocations of the same LF across languages (e.g., in English and Norwegian we *take a nap*, in German, we 'make' it, in Portuguese we 'pull' it, in Spanish, we 'throw' it, etc.). We would also like to explore the idea of "semantic masking" for collocate discovery, where we would train models for dynamically *masking* (or removing) idiosyncratic information such that only the semantics of the collocate remain, thus largely corresponding to a latent abstraction over the LF. This approach has been applied recently in the lexical substitution task, with the limitation, however, that the dropout rate was tuned in a validation set, whereas a promising avenue to explore would be to automatically learn the embedding dropout in a fully supervised setting. Finally, motivated by the observed large gap in performance between the categorization of, e.g., Oper1 and Oper2, Bon and AntiBon, Ver and AntiVer, we plan to investigate in more depth the codification of collocational information in pretrained LMs.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. This work was partially supported by the European Commission via its H2020 Program under the contract number 870930. The experiments were executed using the computational facilities of the Advanced Research Computing @Cardiff (ARCCA) at Cardiff University.

References

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1183–1193.

Gemma Boleda, Marco Baroni, Louise McNally, et al. 2013. Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013): long papers; 2013 Mar 20-22; Postdam, Germany. Stroudsburg (USA): Association for Computational Linguistics (ACL); 2013. p. 35-46.* ACL (Association for Computational Linguistics).

Jose Camacho-Collados, Luis Espinosa-Anke, Shoaib Jameel, and Steven Schockaert. 2019. A latent variable model for learning distributional relation vectors. In *International Joint Conferences on Artificial Intelligence*.

Kenneth W. Church and Patrick Hanks. 1989. Word Association Norms, Mutual Information, and Lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 76–83, Vancouver, Canada.

Anthony P. Cowie. 1994. Phraseology. In R.E. Asher and J.M.Y. Simpson, editors, *The Encyclopedia of Language and Linguistics, Vol. 6*, pages 3168–3171. Pergamon, Oxford.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Daniel Edmiston. 2020. A systematic analysis of morphological content in bert models for multiple languages. *arXiv preprint arXiv:2004.03032v1*.

Luis Espinosa-Anke, Jose Camacho-Collados, Sara Rodríguez Fernández, Horacio Saggion, and Leo Wanner. 2016. Extending wordnet with fine-grained collocational information via supervised distributional learning. In *Proceedings of COLING 2016: Technical Papers. The 26th International Conference on Computational Linguistics; 2016 Dec. 11-16; Osaka (Japan): COLING; 2016. p. 900-10.* COLING.

Luis Espinosa-Anke and Steven Schockaert. 2018. Seven: Augmenting word embeddings with unsupervised relation vectors. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2653–2665.

Luis Espinosa-Anke, Steven Schockaert, and Leo Wanner. 2019. Collocation classification with unsupervised relation vectors. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5765–5772.

Stefan Evert. 2007. Corpora and Collocations. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics. An International Handbook*. Mouton de Gruyter, Berlin.

John R. Firth. 1957. Modes of Meaning. In J.R. Firth, editor, *Papers in Linguistics, 1934-1951*, pages 190–215. Oxford University Press, Oxford.

Marcos Garcia, Marcos García Salido, and Margarita Alonso Ramos. 2017. Using bilingual word-embeddings for multilingual collocation extraction. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 21–30.

A. Gelbukh and O. Kolesnikova. 2012. *Semantic Analysis of Verbal Collocations with Lexical Functions*. Springer, Heidelberg.

Franz Josef Hausmann. 1985. Kollokationen im Deutschen Woerterbuch: ein Beitrag zur Theorie des lexicographischen Biespiels. *Lexikographie und Grammatik*.

John Hewitt and Christopher D Manning. 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)*, pages 4129–4138.

Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30.

Mandar Joshi, Eunsol Choi, Omer Levy, Daniel S Weld, and Luke Zettlemoyer. 2019. pair2vec: Compositional word-pair embeddings for cross-sentence inference. In *NAACL-HLT (1)*.

Mandar Joshi, Kenton Lee, Yi Luan, and Kristina Toutanova. 2020. Contextualized representations using textual encyclopedic knowledge. *arXiv preprint arXiv:2004.12006*.

Adam Kilgarriff. 2006. Collocationality (And How to Measure it). In *Proceedings of the 12th Euralex International Congress on Lexicography (EURALEX)*, pages 997–1004, Turin, Italy. Springer-Verlag.

Brigitte Krenn. 2000. *The Usual Suspects: Data-Oriented Models for the Identification and Representation of Lexical Collocations*, volume 7 of *Saarbrücken Dissertations in Computational Linguistics and Language Technology*. DFKI and Universität des Saarlandes.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations.

Omer Levy, Steffen Remus, Chris Biemann, Ido Dagan, and Israel Ramat-Gan. 2015. Do supervised distributional methods really learn lexical inference relations? In *NAACL 2015*, Denver, Colorado, USA. Association for Computational Linguistics.

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*.

Marco Maru, Federico Scozzafava, Federico Martelli, and Roberto Navigli. 2019. Syntagnet: challenging supervised word sense disambiguation with lexical-semantic combinations. 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 3525–3531.

I.A. Mel'čuk. 1996. Lexical functions: A tool for the description of lexical relations in the lexicon. In Leo Wanner, editor, *Lexical Functions in Lexicography and Natural Language Processing*, pages 37–102. Benjamins Academic Publishers, Amsterdam.

Igor Mel'čuk. 1995. Phrasemes in Language and Phraseology in Linguistics. In M. Everaert, E.-J. van der Linden, A. Schenk, and R. Schreuder, editors, *Idioms: Structural and Psychological Perspectives*, pages 167–232. Lawrence Erlbaum Associates, Hillsdale.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 2016. Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 454–459.

Masataka Ono, Makoto Miwa, and Yutaka Sasaki. 2015. Word embedding-based antonym detection using thesauri and distributional information. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 984–989.

Pavel Pecina. 2008. A Machine Learning Approach to Multiword Expression Extraction. In *Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions)*, pages 54–57, Marrakech, Morocco.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12.

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.

Jipeng Qiang, Yun Li, Yi Zhu, and Yunhao Yuan. 2019. A simple bert-based approach for lexical simplification. *arXiv preprint arXiv:1907.06226*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Carlos Ramisch, Silvio Cordeiro, Agata Savary, Veronika Vincze, Verginica Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, et al. 2018. Edition 1.1 of the parseme shared task on automatic identification of verbal multiword expressions.

Sara Rodríguez Fernández, Luis Espinosa-Anke, Roberto Carlini, and Leo Wanner. 2016. Semantics-driven recognition of collocations using word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics; 2016 Aug. 7-12; Berlin (Germany).[place unknown]: ACL; 2016. Vol. 2, Short Papers; p. 499-505.* ACL (Association for Computational Linguistics).

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *arXiv preprint arXiv:2002.12327*.

V. Sanh, L. Debut, J. Chaumond, and Th. Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*, Vancouver, BC, Canada.

V. Sanh, L. Debut, J. Chaumond, and Th. Wolf. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of ICLR*.

Roy Schwartz, Roi Reichart, and Ari Rappoport. 2015. Symmetric pattern based word embeddings for improved word similarity prediction. In *Proceedings of the nineteenth conference on computational natural language learning*, pages 258–267.

V. Seretan. 2014. On collocations and their interaction with parsing and translation. *Informatics*, 1(1):11–31.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Proceedings of NeurIPS*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of ICLR*.

Leo Wanner and John A. Bateman. 1990. A collocational based approach to salience sensitive lexical selection. In *Proceedings of the 5th International Workshop on Natural Language Generation*, Dawson, PA.

Leo Wanner, Bernd Bohnet, and Mark Giereth. 2006. Making sense of collocations. *Computer Speech and Language*, 20(4):609–624.

Koki Washio and Tsuneaki Kato. 2018. Neural latent relational analysis to capture lexical semantic relations in a vector space. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 594–600.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *International Conference on Computational Science*, pages 84–95. Springer.

Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q.V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, BC, Canada.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. Bert-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373.