Using Distributional Principles for the Semantic Study of Contextual Language Models

Olivier Ferret

Using Distributional Principles for the Semantic Study of Contextual Language Models

Olivier Ferret
Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
olivier.ferret@cea.fr

Abstract

Many studies were recently done for investigating the properties of contextual language models but surprisingly, only a few of them consider the properties of these models in terms of semantic similarity. In this article, we first focus on these properties for English by exploiting the distributional principle of substitution as a probing mechanism in the controlled context of SemCor and WordNet paradigmatic relations. Then, we propose to adapt the same method to a more open setting for characterizing the differences between static and contextual language models.

1 Introduction

The introduction of contextual word embeddings such as ELMo (Peters et al., 2018) or BERT (Devlin et al., 2019) is considered as a major breakthrough in Natural Language Processing, especially for classification or sequence labeling tasks that can be addressed by supervised machine learning approaches. However, this kind of embeddings also represents a significant change from the viewpoint of distributional semantics. While previous approaches, including static word embeddings such as Skip-gram or CBOW models (Mikolov et al., 2013), built the distributional representation of a word by cumulating the contexts in which it was found, the neural language models such as ELMo or BERT produce a representation for each occurrence of the word.

While this specificity is not a difficulty – and even more so an advantage – in the context of a supervised classification or sequence labeling task, it is a problem in contexts requiring a distributional representation of words at the type level and not only at the token level. This is the case for the evaluation of the semantic similarity of words, which is a classical tool for investigating the semantic properties of distributional representations through datasets, such as Simlex-99 (Hill et al., 2015), built for evaluating the correlation of the similarity computed from these representations with human judgments.

As illustrated by Rogers et al. (2020), the properties of contextual word embeddings have been the focus of many recent studies, especially in the case of BERT. Surprisingly, only a few of them consider the semantic properties of these models. Moreover, most of them are performed at the word type level, which involves building representations of words from the representation of their occurrences, which is neither a straightforward nor a neutral operation.

In this article, we propose to exploit the distributional principles to investigate the semantic properties of contextual word embeddings in terms of paradigmatic semantic relations without the requirement to build representations at the word type level. More precisely, we present the following main contributions:

- we show that the distributional hypothesis can be applied at the word token level with contextual word embeddings and that ELMo and BERT exhibit specific semantic properties falling more into the scope of semantic similarity than semantic relatedness;
- we show how the transposition of this method to the word type level can be used to study the differences between contextual and static word embeddings.

2 Related work

2.1 Distributional semantic similarity

The work on distributional semantics is closely linked to the notion of semantic similarity since one major criterion for judging the quality of a distributional representation is its ability to account for semantic similarity, either by its correlation with direct human judgment (Rubenstein and Goodenough, 1965) or more indirectly through the extraction of semantic similarity relations (Grefenstette, 1996; Landauer and Dumais, 1997). In our work, we focus more particularly on the second option, which is historically linked to the notion of distributional thesaurus: the semantic similarity relations of a word are extracted by retrieving its nearest neighbors according to the similarity of their distributional representations (Grefenstette, 1994; Lin, 1998; Curran and Moens, 2002). This perspective was dominant for count-based approaches (Weeds et al., 2004; Ferret, 2010; Padró et al., 2014) but is also represented among studies on static word embeddings (Schnabel et al., 2015; Antoniak and Mimno, 2018; Ferret, 2018). In this article, we extend the use of this paradigm to the exploration of the semantic properties of contextual embeddings.

2.2 Semantic Study of Contextual Word Embeddings

In (Rogers et al., 2020), the semantic studies of BERT-like models mainly point out work focusing on semantic roles in sentences, either in the context of semantic role labeling (Tenney et al., 2019) or tasks close to lexical substitution (Ettinger, 2020; Garí Soler et al., 2019). However, the work of Vulić et al. (2020) clearly presents the largest set of experiments concerning the semantic properties of BERT and RoBERTa models. It includes experiments concerning their correlation with human judgment in terms of semantic similarity or their ability to classify word pairs according to different types of relations (Chersoni et al., 2016; Xiang et al., 2020). Other studies also considered contextual embeddings from a semantic viewpoint but in more specific contexts: the impact of their training objectives (Mickus et al., 2020), their level of contextualization (Ethayarajh, 2019), their possible biases (Bommasani et al., 2020), their ability to represent word senses (Coenen et al., 2019), to build representations for rare words (Schick and Schütze, 2020), to account for selectional preferences (Metheniti et al., 2020) or to interpret logical metonymy (Rambelli et al., 2020). Finally, (Chronis and Erk, 2020) is linked to the representation of word senses through the notion of prototype but mainly applies it for characterizing semantic similarity versus semantic relatedness and abstractness versus concreteness. While it has links with some of these studies, our work proposes a specific method, based on the distributional hypothesis, that aims at characterizing contextual word embeddings by focusing on paradigmatic relations observed at the word token level without the necessity to build representations at the word type level.

3 Principles and Methodology

With contextual embedding models, the similarity between embeddings can only be computed for words in context, i.e. occurrences of words. More precisely, the embedding built for an occurrence of a word with this kind of model integrates two dimensions: one dimension, resulting from the training on a language modeling task, accounts for the aggregation of all the contexts of the occurrences of the word, as for all distributional models; the second dimension takes into account the local context of the considered occurrence, i.e. its surrounding words.

In our case, we focus on the first dimension for evaluating the similarity between words, which requires being able to control the second one. We propose to perform this control in a very simple manner: for evaluating the similarity of two words, we put the two words in the same context, i.e. at the same position in the same sentence, and compute a word embedding for each of them by applying a contextual language model. In practice, this control is implemented by a substitution: for two words w_1 and w_2, a sentence S_i containing an occurrence of w_1 is selected and a contextualized representation of w_1 is computed; then, w_1 is replaced by w_2 in sentence S_i and a contextualized representation of w_2 is computed similarly as for w_1. By computing the similarity between the representations of w_1 and w_2, we can evaluate to what extent w_1 can be replaced by w_2 and, by following Harris (1954)'s principle of *substitutability*, also obtain an evaluation of their

semantic similarity. This evaluation is limited to the selected sentence but the application of this strategy to a set of sentences gives a more global picture of the semantic relationship between the two words. It should be noted that w_1 and w_2 do not have symmetric roles: we evaluate the similarity of w_2 with respect to $w1$ since the representation of w_2 is built in the context of a sentence in which w_1 initially occurs. In the rest of the paper, the terms *key*, *target*, and *test sentence* will refer to w_1, w_2, and S_i respectively.

We first exploit the presented principle of substitution for investigating the semantic properties of ELMo and BERT. More precisely, the idea is, for each word of a set of keys, to gather a set of targets such that each (key, target) pair is linked by a semantic relation of a known type. Given a test sentence containing an occurrence of the key, its targets can be ranked according to their similarity value with the key following the principle of substitution outlined previously. The resulting ranking indirectly ranks the types of semantic relations associated with the targets and gives some insights into the semantic properties of ELMo or BERT.

Our work focuses more particularly on paradigmatic relations, more precisely synonymy [SYN], hypernymy [HYPE], hyponymy [HYPO], and cohyponymy [COHYP] and adopts WordNet 3.0 as a reference resource for these types of relations. However, WordNet (Miller, 1990) is based on the notion of synset while sentences contain words and not word senses. To bypass this difficulty, we use as test sentences sentences from SemCor (Miller et al., 1993), a subset of the Brown Corpus whose open class words were tagged with WordNet synsets. Hence, the sense of an occurrence of a key is known and the relation with the target depends on this sense, which makes the use of substitution particularly accurate. For instance, the second sense of the key *disaster* (*an event resulting in great loss and misfortune*) has the following test sentence in SemCor:

[1] Since the 1946 **disaster** there have been 15 tsunami in the Pacific, but only one was of any consequence.

This sense of *disaster* has words such as *cataclysm* or *catastrophe* as synonyms, *misfortune* as hypernym, *tsunami* or *meltdown* as hyponyms, and *adversity* or *misadventure* as cohyponyms. For evaluating to which extent ELMo for instance accounts more for synonymy than for hypernymy, the test sentence [1] is turned into sentences [2] and [3] by substituting a synonym or a hypernym for the key.

[2] Since the 1946 **catastrophe** there have been 15 tsunami in the Pacific, but only one was of any consequence.

[3] Since the 1946 **misfortune** there have been 15 tsunami in the Pacific, but only one was of any consequence.

The three sentences are encoded using ELMo and the representation of both the key in sentence [1] and the two targets in sentences [2] and [3] are extracted from ELMo's internal layers. Since ELMo has three layers – one input non-contextual layer, layer 0, and two contextual layers, layer 1 and layer 2 – we actually have three representations for each word. In this example and for layer 1, the similarity, classically evaluated by the *cosine* measure, between the representations of the key and the target synonym *catastrophe* is equal to 0.89 while the similarity of the key and the target hypernym *misfortune* is only equal to 0.55, giving a clear advantage to synonymy over hypernymy in that case. The process is the same for BERT, except that we have 12 layers (layer 1 to layer 12) in that case. A global picture of the semantic properties of ELMo or BERT is obtained by considering a significant number of keys and targets in the context of a large number of test sentences from SemCor. Moreover, we enrich this picture by considering the relations between the ELMo and BERT embeddings and the more classical static word embeddings by adding as targets DIST_NGH the most similar neighbors to the keys of our study retrieved with the *cosine* measure by a Skip-gram model trained on a large corpus.

This investigation of the semantic properties of ELMo and BERT can be viewed as a kind of semantic probing task and related to the work of Schick and Schütze (2020) and their *WordNet Language Model Probing*. While their overall objective is different from ours, their probing task is also significantly different since their notion of pattern is more adapted to syntagmatic relations than to paradigmatic relations.

4 Experiments: Study of Contextual Word Embeddings

4.1 Experimental Setup

The implementation of the principles of the previous section is associated with some choices about test

			ELMo			BERT					
	avg. #tgt	**random**	**L0**	**L1**	**L2**	**L1**	**L3**	**L5**	**L8**	**L10**	**L12**
SYN	2.1	5.3	30.9	29.3	26.9	33.5	34.5	36.1	**36.7**	36.0	35.1
HYPE	1.9	6.7	4.3	6.1	6.2	7.8	9.3	9.9	10.9	11.1	**11.2**
HYPO	5.9	14.5	11.4	13.8	**14.8**	10.5	11.1	11.7	12.3	11.9	12.3
COHYP	8.3	29.5	6.7	7.9	**9.6**	6.4	6.7	7.5	7.7	7.8	7.9
DIST_NGH	10	44.0	**46.6**	42.9	42.5	41.7	38.4	34.7	32.3	33.1	33.5

Table 1: P@1 ($\times 100$) for the ranking, based on SemCor's sentences, of the reference targets for a set of keys[1].

sentences and semantic relations between keys and targets. First, both keys and targets are restricted to nouns. The number of targets for each key is limited to 40 to have comparable results for all keys. Moreover, we only retain keys having targets for all the five types of relations we consider, once again for the homogeneity of results. We also limit the number of targets for each type of relation to 10, with an additional limit of 30 for all WordNet relations. In practice, this limit only concerns hyponymy and cohyponymy relations, whose number tends to be high. The first column of Table 1 gives the average number of targets of each key according to their type.

For WordNet relations, we adopt the following definitions for each type of target:

- synonyms [SYN]: all the words of $Synset_{key}$, the synset of the sense associated with the occurrence of the considered key in a test sentence;
- hypernyms [HYPE]: all the words of the synsets $Synset_{hype}$ having a direct hypernymy relation with $Synset_{key}$;
- hyponyms [HYPO]: all the words of the synsets having a direct hyponymy relation with $Synset_{key}$;
- cohyponyms [COHYP]: all the words of the synsets, except $Synset_{key}$, having a direct hyponymy relation with the $Synset_{hype}$ synsets.

As mentioned previously, the DIST_NGH targets are obtained by a Skip-gram model, trained on a 1 billion word subset of the Annotated English Gigaword corpus (Napoles et al., 2012) with the best hyperparameter values from (Baroni et al., 2014). More precisely, we use this model to retrieve the 10 first neighbors, among a vocabulary of 20,813 nouns, of each key that are not present in the set of WordNet targets. In that case, this selection is not done according to the sense of the key in a test sentence since we do not have access to word senses.

Concerning test sentences, an upper limit on the number of sentences for each key is also fixed: no more than 20 sentences for each sense of a key. Finally, our evaluation is based on 41,079 SemCor's sentences, which represents around 4.5 sentences by sense on average and 7.9 sentences for each of the 5,241 keys. These keys cover a large spectrum of frequencies since if we refer to the subpart we use of the Gigaword corpus, the most frequent key, *year*, has 2,991,899 occurrences while the least frequent keys such as *inadvertence* have only 22 occurrences.

4.2 Evaluation

The results of the ranking of the targets selected for all our keys are presented in Table 1 for each layer of ELMo and BERT and each type of target. Moreover, the second column provides the P@1 values of a random ranker (average values over 100 runs). P@1 in this context corresponds to the proportion of sentences that rank first a target linked to a key with a specific type of relation (one by row).

This table first shows that the different layers of ELMo and BERT do not have exactly the same semantic properties, which is a confirmation of previous findings as those of Ethayarajh (2019) or Wu et al. (2020). It also shows some similarities and differences between ELMo and BERT. The most obvious difference is that BERT has much higher results than ELMo for synonyms and hypernyms but lower results for hyponyms and cohyponyms, meaning that BERT favors semantic similarity over semantic relatedness (Budanitsky and Hirst, 2006) as we can

[1]The percentage of cases in which the first target of a key corresponds to a paradigmatic relation is equal to $100 - P@1(\text{DIST\_NGH}) \times 100$. For instance, 56% for the random ranker.

consider that hyponyms and cohyponyms, despite their paradigmatic nature, are less strong than synonyms and hypernyms in terms of semantic similarity. The trend is even more obvious for the DIST_NGH relations coming from the static embeddings, which are more likely to be syntagmatic relations.

ELMo and BERT also exhibit some differences in their layers. While the capacity of ELMo to rank synonyms first steadily decreases as we consider higher layers, it first increases until layer 8 for BERT, then tends to decrease. This observation has also to be put in relation to the results of Ethayarajh (2019), who observed that word representations are more context-specific as the level of their layers increases. From the viewpoint of ELMo, it means that having more contextual representations is likely to favor semantic relatedness over semantic similarity. This is not unexpected since from the semantic viewpoint, contextual relations are syntagmatic rather than paradigmatic relations. This result is less obvious for BERT since its first layers have the best precision for DIST_NGH relations but a change more compatible with the results of ELMo occurs after layer 8 concerning the ranking of synonyms and DIST_NGH relations.

However, ELMo and BERT also have strong convergences. First, the most significant effect of the ranking by ELMo and BERT is obtained for synonyms. P@1 is up to nearly six times higher for the best ELMO's layer and nearly seven times for the best BERT's layer compared to the random ranking, which indicates that the application of the distributional hypothesis as described in Section 3 is an interesting method for identifying the synonyms of a word among a list of its distributional neighbors. It also illustrates the fact that even if some differences can be noted between layers in terms of semantic orientation, they all have a strong bias towards synonymy. We can also observe that both ELMo and BERT have a strong inverse correlation between P@1 for synonyms and P@1 for DIST_NGH relations: when a layer tends to favor strict semantic similarity, it logically obtains worse results for semantic relatedness at the same time. However, this correlation does not lead to results for DIST_NGH relations much lower than the results of a random ranker for ELMo, which suggests that ELMo is not radically different from static embeddings from the viewpoint of the semantic similarity it conveys. This trend is less clear-cut for BERT since its results for synonyms are more comparable to those for DIST_NGH relations but BERT nevertheless gives significant importance to DIST_NGH relations. A closer look at the first ranked DIST_NGH word also shows that it frequently has a strong semantic relationship with the key. In some cases, it is one of its synonyms but for a different sense, that is close to the sense of the current occurrence. For instance, in the sentence:

> Land reform programs need to be supplemented with programs for promoting rural credits and [...] in **agriculture**.

the word ranked first by the first layer of ELMo, *farming*, is a synonym of the second sense of *agriculture* – *the practice of cultivating the land or raising stock* – whereas this occurrence is tagged in SemCor with its first sense – *a large-scale farming enterprise*. In some other cases, the top word is actually a synonym of the key but is not considered as such in WordNet. For instance, the top word *capability* for the key *ability*. Finally, it is also frequent to find as the top word an antonym of the key, as *inaction* for the key *action*. While this is not an intended outcome, antonyms are known to be distributionally very similar to synonyms and their presence confirms the observed trend to favor semantic similarity.

5 Contextual Word Embeddings vs Static Embeddings

5.1 Principles

Besides the investigation of the semantic properties of contextual word embeddings, the method we have presented can also be adapted to study the differences between contextual and static word embeddings concerning these properties. More precisely, the idea is to define targets by replacing WordNet semantic relations with relations characterizing static embeddings. Due to the distributional nature of these embeddings, we choose to associate as targets with each key, corresponding to a word W_i, its most similar distributional neighbors according to the considered static embeddings. As previously, targets are reranked according to a set of test sentences and we use paradigmatic relations in WordNet for determining a posteriori which types of relations contextual word embeddings favor compared to static embeddings.

However, since static embeddings are defined at

the word type level, their comparison with contextual word embeddings requires aggregating the contextual representations of keys and targets built from test sentences for producing type level representations. We classically distinguish between early and late fusion. The early fusion consists in this context in aggregating the contextual representations of a key or a target extracted from the encoding by ELMo or BERT of the test sentences where it occurs. For this aggregation, we consider three operators (Bommasani et al., 2020): *average*, *element-wise maximum*, and *minimum*. The late fusion approach (Curran, 2002) produces a reranking of the neighbors of a word for each test sentence and merges the resulting rankings according to methods that are typically used in Information Retrieval for merging ranked lists of retrieved documents. More precisely, we experiment with two kinds of methods: the *Borda*, *Condorcet* (Nuray and Can, 2006) and *Reciprocal Rank* (RRF) (Cormack et al., 2009) fusions based on ranks and the *CombSum* fusion based on similarity values, normalized with the Zero-one method (Wu et al., 2006).

The definition of static embeddings at the word level also influence the way test sentences are selected: a static embedding is not associated with a particular sense of a word but, according to the idea developed by McCarthy et al. (2004) that most words have one predominant sense in a specific corpus, it is supposed to be related to the predominant sense of the word it is associated with in the corpus used for its building. Hence, we adopt a strategy for selecting test sentences accordingly. Its first step consists in selecting randomly for each word W_i a large enough set of sentences containing the word, at most N_{sent}, from the corpus used for retrieving the targets. The resulting set of contexts for each word can be considered statistically representative of its various senses.

The second step aims at selecting a smaller number of context sentences for performing the reranking at a reasonable cost while taking into account the predominant sense hypothesis. Following this hypothesis, we assume that most of the test sentences first selected for a word refer to its predominant sense. As a consequence, averaging the representations of the word resulting from the encoding of the set $\mathcal{C} = \{S_j\}$ of its test sentences by ELMo or BERT should lead to a representation of the word, denoted $v_{(W_i,\mathcal{C})}$, very close to its predominant sense. Considering that we select a fixed number $N_\mathcal{C}$ of test sentences, we propose the following options:

- *random*: it is our base option in which $N_\mathcal{C}$ sentences are randomly selected among the N_{sent} initially selected for the word;
- *closest_avg*: we select test sentences S_j such that the representation of the word $v_{(W_i,S_j)}$ is closest to $v_{(W_i,\mathcal{C})}$, with the idea to favor the homogeneity among the test sentences towards the predominant sense of the word;
- *farthest_avg*: this is the opposite of *closest_avg*. We select test sentences such that $v_{(W_i,S_j)}$ is farthest to $v_{(W_i,\mathcal{C})}$ to increase the presence of minor senses of the word;
- *uniform*: the idea is to account for the diversity of word's senses by selecting $N_\mathcal{C}$ that are uniformly distributed in terms of the similarity of $v_{(W_i,S_j)}$ to $v_{(W_i,\mathcal{C})}$.

5.2 Experimental Setup

The main difference with the setup of our first experiment concerns test sentences: they are selected from the corpus used for training the static embeddings and keys are not semantically disambiguated in them. More precisely, we select 10 test sentences for each key with a size between 10 and 90 words for having a significant and focused context. For the static embeddings, we rely on the same Skip-gram model as the one used for extracting the DIST_NGH relations. In a first experiment (see Section 5.3), we apply the reranking process to the same 5,241 keys as in Section 4 for having a direct comparison with Table 1. Test sentences are selected randomly and we apply an early fusion by averaging the representations extracted from test sentences. In a second experiment (see Section 5.4), we consider a larger number of 10,302 keys for testing the various options presented in the previous section. In both cases, the targets correspond to the first 10 distributional neighbors of the keys, retrieved by the *cosine* measure.

As in (Piasecki et al., 2018), our gold standard neighbors are obtained by extracting from WordNet the words linked to a key through the same types of relations as in Section 4 but the number of relations is larger since we consider all the synsets in which the key is present (see the first column of Table 2). We report the precision at the first rank (P@1) of the retrieved neighbors for each type of relation in

	avg.		ELMo			BERT					
	#ref. rel.	Skip-gram	L0	L1	L2	L1	L3	L5	L8	L10	L12
SYN	3.5	18.6	22.6	**23.0**	21.4	20.1	20.6	20.9	21.4	21.3	21.9
HYPE	5.0	6.9	5.8	6.6	6.3	8.1	8.3	8.6	8.9	**9.0**	8.5
HYPO	10.8	11.8	13.3	**14.2**	13.7	10.5	10.8	11.3	11.5	10.9	10.2
COHYP	56.3	27.9	33.5	**34.8**	33.0	30.0	30.7	31.2	31.3	31.2	31.6

Table 2: P@1 ($\times 100$) for the ranking of Skip-gram's distributional neighbors by contextual models.

the first experiment and add precisions at ranks 2 and 5 (P@2 and P@5) for the second experiment but without detailing these measures according to the different types of relations.

5.3 Skip-gram versus Contextual Models

Table 2 compares the distributional neighbors retrieved by the Skip-gram model with their ranking by ELMo and BERT according to the methodology we propose. The first observation is that while these models significantly[2] favor synonyms and cohyponyms compared to Skip-gram, the situation is more complex for hypernyms and hyponyms: ELMo degrades Skip-gram's results for hypernyms but improves them for hyponyms while BERT does exactly the opposite. We can also note that the Skip-gram model largely favors cohyponymy over the other lexical relations, a trend that tends to be increased by ELMo and BERT, to a greater extent by ELMo. Interestingly, in Table 1, the ranking of cohyponyms by ELMo and BERT is much worse than the one obtained by a random ranker. This difference illustrates the contextual nature of these models but also its limits. In the experiment of Section 4, keys are semantically disambiguated in test sentences and targets are related to their sense. In such a situation, a contextual model can favor the targets most semantically linked to their key, such as synonyms, because they produce representations that are specific to the considered context, which is related to the sense of the key. On the contrary, in this second experiment, the targets cover all the senses of the key and in such a configuration, corresponding to the word type level, the representations built by contextual models favor semantic relatedness over semantic similarity and are closer to static embeddings, even if they globally tend to outperform them. This effect has a much greater impact for synonyms in the case of BERT than for ELMo, which probably results from a greater sensitivity of BERT to the context than ELMo. However, it does not modify the pattern of results among BERT's layers, with best results around layer 8, while ELMo's best results are fairly surprisingly obtained by a more contextualized layer than previously. We can note that the observation about BERT is close to the findings of Chronis and Erk (2020), who found that BERT's layer 7 "is optimal for estimating similarity" in their comparison with relatedness.

Table 2 also shows that BERT is globally closer to the Skip-gram model than ELMo from the viewpoint of paradigmatic relations, except for hypernyms. This is surprising if we consider the results of Table 1, where BERT globally outperforms ELMo for paradigmatic relations, especially for synonyms. We analyze this observation as a consequence of the use of wordpieces by BERT for dealing with out-of-vocabulary words. In the experiments of Section 4, 73% of words are part of BERT's vocabulary while this ratio decreases to 49.4% in the experiments of Table 2. Following Bommasani et al. (2020) and others, we build the representation of a word split into wordpieces by averaging the representations of its wordpieces. While this is considered as the best strategy in such a situation, our results show that it has a significant negative impact when the number of words split into wordpieces reaches a certain level and introduces a form of bias into results as all words do not have the same status.

5.4 Impact of Reranking Options

We have seen in Section 5.1 that the method we propose depends on the definition of several hyperparameters and options. Table 3 shows the impact of two parameters of the method that influence both its per-

[2]Differences are judged significant according to a paired Wilcoxon test if $p \leq 0.01$.

	P@1	P@2	P@5
n=10, s=10	**41.8**	34.5	24.0
s=10, n=5	41.3†	33.6	21.6
s=10, n=15	41.6†	**34.8**†	24.7
n=15, s=5	41.2†	34.6†	**24.8**
n=15, s=15	41.6†	**34.8**†	**24.8**

Table 3: Impact of the number of neighbors (n) and test sentences (s) for ELMo$_{L1}$ (reference configuration of Table 2: n=10, s=10; †: non-significant difference compared to this reference).

	P@1	P@2	P@5
average	**41.6**	**34.8**	**24.7**
max	40.6	33.9	24.3
min	40.5	33.9	24.3
Borda	41.1	34.4	24.6
Condorcet	40.9	34.4	24.6
RRF	41.1†	34.4	24.6
CombSum	41.0	34.3	24.6†

Table 4: Impact of the fusion method for token level representations with ELMo$_{L1}$ (reference method in Table 3: *average*).

formance and speed: the number of test sentences (s) and the number of reranked neighbors (n). The evaluation of this impact is done according to ELMo's layer 1, which globally obtains the best results in reranking targets from Skip-gram, as indicated by Table 2. Moreover, it is also performed by merging the four types of semantic relations we consider in Table 2. We can globally observe that the proposed method is not very sensitive to changes in the value of these two parameters as their impact on results is generally not significant. These experiments also show that reranking only a small number of neighbors under-exploits the capabilities of the method, especially when the number of test sentences is small. However, it is interesting to note that the proposed method can be effective with only a restricted number of test sentences. Finally, we adopt s=10 and n=15 as the more interesting compromise between the quality of results and the speed of the method for our following evaluations.

In Section 5.1, we have seen that considering keys and targets at word level requires aggregating token level representations, with several possible options.

	P@1	P@2	P@5
random	41.6	34.8	24.7
closest_avg	41.4†	34.5	**24.8**†
farthest_avg	39.9	33.6	24.1
uniform	**41.8**†	**34.9**†	**24.8**†

Table 5: Impact of the selection method of test sentences for ELMo$_{L1}$ (reference method in Table 3: *random*).

	P@1	P@2	P@5
initial$_{high}$	41.6	35.3	26.4
initial$_{low}$	30.0	24.1	16.8
reranking$_{high}$	50.8	43.1	31.1
reranking$_{low}$	32.8	26.6	18.4

Table 6: Comparison of the *initial* ranking of the targets given by Skip-gram and their *reranking* by ELMo$_{L1}$ (corresponding to *uniform* in Table 5) according to the frequency of keys (*high* or *low*).

Taking as reference the *average* option adopted in the evaluation of Section 5.3, Table 4 shows that while the best results are obtained by the *average* option, an early fusion approach, the late fusion approaches globally outperform the other early fusion options. However, the results of the different settings are fairly homogeneous. This is particularly true for all the results of the late fusion approaches, which suggests that for a key or a target, the rerankings produced by the test sentences are very similar.

The last evaluation of the options of the reranking concerns the way test sentences are selected from their initial set. As for the aggregation of token representations, the differences between the tested options reported in Table 5 are not very high. Only the *farthest_avg* option is significantly worse than *random*, our reference, which is not very surprising since favoring the most atypical test sentences does not seem a priori a good option. Finally, the *uniform* method appears as the best choice both because it obtains the best results, even if they are not statistically different from those of *random*, and more importantly, it is deterministic, which is not the case of *random*.

5.5 Complementary Analyses and Examples

While Table 2 shows the differences between static and contextual embeddings according to the type of semantic relation, it is also interesting to perform

Keys	Initial order of targets	Targets after reranking
response	reaction$_{syn}$, wake$_{coh}$, action, criticism, rebuke, condemnation, denunciation$_{coh}$, explanation$_{coh}$, comment$_{coh}$	reaction$_{syn}$, reply$_{syn}$, answer$_{syn}$, counteraction, rebuke, explanation$_{coh}$, unresponsiveness, condemnation, action
bureau	department, telephone, reporting, newsroom, enumerator, agency$_{syn}$, office$_{syn}$, correspondent, authority	agency$_{syn}$, department, office$_{syn}$, authority$_{syn}$, newsroom, correspondent, enumerator, chief, deputy
supposition	contrary, conjecture$_{syn}$, theory$_{hype}$, fact, notion, assumption$_{syn}$, characterisation, assertion, reductionism	assumption$_{syn}$, assertion, conjecture$_{syn}$, hypothesis$_{syn}$, notion, belief, surmise$_{syn}$, theory$_{hype}$, characterisation
lookout	alert, loiterer, vigilance, binoculars, spotter$_{syn}$, prowl, watch$_{syn}$, warning, rubbernecker, sentry$_{syn}$	prowl, sentry$_{syn}$, spotter$_{syn}$, sentinel$_{syn}$, watch$_{syn}$, sightseer, alert, picnicker, vigilance

Table 7: Examples of reranking by ELMo$_{L1}$ of the first targets of some keys (*syn*: synonym, *coh*: cohyponym, *hype*: hypernym).

such a comparison according to the frequency of keys. Table 6 illustrates it for ELMo by the split of results into two equally balanced frequency slices – *high* and *low*. While the reranking method leads to an improvement for the two slices, this improvement is clearly more significant for high-frequency words. Since we use the same number of test sentences for all words, we assume that this finding is not linked to the number of occurrences of the words but their number of senses. As a contextual model, ELMo produces more specific representations than the Skip-gram model we use for retrieving the targets and these focused representations are more effective for retrieving relevant neighbors in terms of paradigmatic relations, especially for polysemous words such as high-frequency words.

Finally, Table 7 presents more qualitatively for some keys the differences between the first targets of the Skip-gram model and their reranking by ELMo. While it shows the global tendency of ELMo to favor synonymy more than Skip-gram does, it also illustrates, in accordance with the results of Table 1, that this feature is particularly effective when a significant proportion of the reranked targets are linked to their key with paradigmatic relations.

6 Conclusion and Perspectives

In this article, we have investigated how the distributional principle of substitution can be used as a form of probe for testing the kind of semantic similarity conveyed by ELMo and BERT. Experiments with WordNet paradigmatic relations and SemCor at the level of word senses have shown that the contextual nature of these models clearly favors synonymy in terms of lexical relations. Moreover, we have adapted the same method for studying the differences between contextual and static embeddings, with the conclusion that the bias of contextual embeddings towards semantic similarity at the token level is reduced at the word type level but is still present.

One extension of this work is to address the problem raised by the representation of words split into wordpieces by BERT. More precisely, we plan to test two types of solutions: one is a specific mechanism for building word representations from their wordpieces as the One-Token Approximation of Schick and Schütze (2020); the other relies on character-based models such as the recent CharacterBERT (El Boukkouri et al., 2020) or CharBERT (Ma et al., 2020) models.

Acknowledgments

This work has been partially funded by French National Research Agency (ANR) under project ADDICTE (ANR-17-CE23-0001).

References

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 238–247, Baltimore, Maryland.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 4758–4781, Online.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1):13–47.

Emmanuele Chersoni, Giulia Rambelli, and Enrico Santus. 2016. CogALex-V shared task: ROOT18. In *5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)*, pages 98–103, Osaka, Japan, December.

Gabriella Chronis and Katrin Erk. 2020. When is a bishop not like a rook? When it's like a rabbi! Multi-prototype BERT embeddings for estimating semantic relationships. In *24th Conference on Computational Natural Language Learning (CoNLL 2020)*, pages 227–244, Online.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. 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*, volume 32, pages 8594–8603. Curran Associates, Inc.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *32^{nd} International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'09)*, pages 758–759.

James R. Curran and Marc Moens. 2002. Improvements in automatic thesaurus extraction. In *Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, pages 59–66, Philadelphia, USA.

James Curran. 2002. Ensemble methods for automatic thesaurus extraction. In *2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 222–229.

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

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Jun'ichi Tsujii. 2020. CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. In *28^{th} International Conference on Computational Linguistics (COLING 2020)*, pages 6903–6915, Barcelona, Spain (Online, dec. International Committee on Computational Linguistics.

Kawin Ethayarajh. 2019. How Contextual Are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*, pages 55–65, Hong Kong, China.

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.

Olivier Ferret. 2010. Testing semantic similarity measures for extracting synonyms from a corpus. In *7th International Conference on Language Resources and Evaluation (LREC'10)*, pages 3338–3343, Valletta, Malta.

Olivier Ferret. 2018. Using pseudo-senses for improving the extraction of synonyms from word embeddings. In *56^{th} Annual Meeting of the Association for Computational Linguistics (ACL 2018), short paper session*, pages 351–357, Melbourne, Australia. Association for Computational Linguistics.

Aina Garí Soler, Marianna Apidianaki, and Alexandre Allauzen. 2019. Word Usage Similarity Estimation with Sentence Representations and Automatic Substitutes. In *Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 9–21, Minneapolis, Minnesota.

Gregory Grefenstette. 1994. *Explorations in automatic thesaurus discovery*. Kluwer Academic Publishers.

Gregory Grefenstette, 1996. *Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic and Window Based Approaches*, page 205–216. MIT Press, Cambridge, MA, USA.

Zellig S. Harris. 1954. Distributional Structure. *Word*, 10(2-3):146–162.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211–240.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *17^{th} International Conference on Computational Linguistics and 36^{th} Annual Meeting of the Association for Computational Linguistics (ACL-COLING'98)*, pages 768–774, Montréal, Canada.

Wentao Ma, Yiming Cui, Chenglei Si, Ting Liu, Shijin Wang, and Guoping Hu. 2020. CharBERT: Character-aware pre-trained language model. In *28th International Conference on Computational Linguistics (COLING 2020)*, pages 39–50, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding Predominant Word Senses in Untagged Text. In *42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 279–286, Barcelona, Spain.

Eleni Metheniti, Tim Van de Cruys, and Nabil Hathout. 2020. How relevant are selectional preferences for transformer-based language models? In *28th International Conference on Computational Linguistics (COLING 2020)*, pages 1266–1278, Barcelona, Spain (Online).

Timothee Mickus, Mathieu Constant, Denis Paperno, and Kees Van Deemter. 2020. What do you mean, BERT? Assessing BERT as a Distributional Semantics Model. *Society for Computation in Linguistics*, 3, January.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR 2013, workshop track*.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A Semantic Concordance. In *Human Language Technology*, Plainsboro, USA.

George A. Miller. 1990. WordNet: An On-Line Lexical Database. *International Journal of Lexicography*, 3(4).

Courtney Napoles, Matthew R. Gormley, and Benjamin Van Durme. 2012. Annotated Gigaword. In *NAACL Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*, pages 95–100, Montréal, Canada.

Rabia Nuray and Fazli Can. 2006. Automatic ranking of information retrieval systems using data fusion. *Information Processing and Management*, 42(3):595–614.

Muntsa Padró, Marco Idiart, Aline Villavicencio, and Carlos Ramisch. 2014. Nothing like good old frequency: Studying context filters for distributional thesauri. In *2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 419–424, Doha, Qatar.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)*, pages 2227–2237, New Orleans, Louisiana, USA.

Maciej Piasecki, Gabriela Czachor, Arkadiusz Janz, Dominik Kaszewski, and Paweł Kedzia. 2018. Wordnet-based evaluation of large distributional models for polish. In *9th Global WordNet Conference (GWC 2018)*.

Giulia Rambelli, Emmanuele Chersoni, Alessandro Lenci, Philippe Blache, and Chu-Ren Huang. 2020. Comparing probabilistic, distributional and transformer-based models on logical metonymy interpretation. In *1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 224–234, Suzhou, China.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Timo Schick and Hinrich Schütze. 2020. Rare Words: A Major Problem for Contextualized Embeddings and How to Fix It by Attentive Mimicking. In *Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, pages 8766–8774, New York, USA.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pages 298–307, Lisbon, Portugal.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing Pretrained Language Models for Lexical Semantics. In *2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 7222–7240, Online.

Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *20th International Conference on Computational Linguistics (COLING 2004)*, pages 1015–1021, Geneva, Switzerland.

Shengli Wu, Fabio Crestani, and Yaxin Bi. 2006. Evaluating score normalization methods in data fusion. In *Third Asia Conference on Information Retrieval Technology (AIRS'06)*, pages 642–648. Springer-Verlag.

John Wu, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2020. Similarity Analysis of Contextual Word Representation Models. In *58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 4638–4655, Online.

Rong Xiang, Emmanuele Chersoni, Luca Iacoponi, and Enrico Santus. 2020. The CogALex shared task on monolingual and multilingual identification of semantic relations. In *Workshop on the Cognitive Aspects of the Lexicon*, pages 46–53, Online. Association for Computational Linguistics.