Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 153-164 recibido 30-03-2022 revisado 19-05-2022 aceptado 23-05-2022

Evaluating Contextualized Vectors from both Large Language Models and Compositional Strategies

Evaluando vectores contextualizados generados a partir de grandes modelos de lenguaje y de estrategias composicionales

Pablo Gamallo, Marcos Garcia, Iria de-Dios-Flores
Centro de Investigación en Tecnoloxías Intelixentes (CITIUS)
Universidade de Santiago de Compostela, Galiza
{pablo.gamallo, marcos.garcia.gonzalez, iria.dedios}@usc.gal

Abstract: In this article, we compare contextualized vectors derived from large language models with those generated by means of dependency-based compositional techniques. For this purpose, we make use of a word-in-context similarity task. As all experiments are conducted for the Galician language, we created a new Galician evaluation dataset for this specific semantic task. The results show that compositional vectors derived from syntactic approaches based on selectional preferences are competitive with the contextual embeddings derived from neural-based large language models.
Keywords: Large Language Models, Contextualized Vectors, Comositionality, Semantic Similarity, Selection Preferences, Syntactic Dependencies.

Resumen: En este artículo, comparamos los vectores contextualizados derivados de grandes modelos de lenguaje con los generados mediante técnicas de composición basadas en dependencias sintácticas. Para ello, nos servimos de una tarea de similitud de palabras en contextos controldados. Como se trata de una experimentación orientada a la lengua gallega, creamos un nuevo conjunto de datos de evaluación en gallego para esta tarea semántica específica. Los resultados muestran que los vectores composicionales derivados de enfoques sintácticos basados en restricciones de selección son competitivos con los *embeddings* contextuales derivados de los modelos de lenguaje de gran tamaño basados en arquitecturas neuronales.
Palabras clave: Grandes Modelos de Lenguaje, Vectores Contextualizados, Composicionalidad, Similitud Semántica, Restricciones de Seleccion, Dependencias Sintácticas.

1 Introduction

Large Language Models (LLMs) are a disruptive breakthrough in Artificial Intelligence that have received an increasing amount of attention in many Natural Language Processing (NLP) tasks. As in the case of classical models, it is possible to use two different approaches to evaluate LLMs: intrinsic and extrinsic evaluations. Intrinsic evaluation consists of using a metric to evaluate the language model itself, without considering any task in which it may be involved. Extrinsic evaluation consists of evaluating the models by employing them in a downstream NLP task. This strategy allows us to compare how their final representation affects the accomplishment of the target task.

Perplexity is one of the most popular metrics for intrinsically evaluating language models. It measures how good a language model is at predicting real sentences. Although perplexity measurements allow researchers to assess the quality of a model in a fast and inexpensive way, it is not considered a fair metric to compare models because the final value is highly dependent on the models' size and vocabulary. In addition, while this metric can be easily applied to classical language models, it is not well-defined for auto-encoding LLM (Salazar et al., 2020), such as masked language models such as BERT (Devlin et al., 2019).

Most commonly, LLMs are evaluated on several NLP tasks by making use of extensive and comprehensive benchmarks (e.g., GLUE (Wang et al., 2019)). However, extrinsic evaluation also has some drawbacks. First, it is a costly and computationally slow process,

ISSN 1135-5948. DOI 10.26342/2022-69-13

since it requires supervised fine-tuning (i.e., training a new model with annotated examples to adapt it to the task). Second, hyperparameters for fine-tuning are likely to have an important influence on the results of the evaluations (Shibayama et al., 2020). And third, the most comprehensive datasets to evaluate LLMs are only available for either English or a dozen of mid-resource languages (Lin et al., 2021), but not for low-resource languages such as Galician.

As an alternative to fine-tuning, which adjusts the vector weights with new annotated data, it is possible to optimize a pre-trained language model for many different tasks by making use of prompt tuning, which is a sort of zero-shot learning approach based on the optimization of the model by embedding the description of the task in the input. So LLMs can also be externally evaluated through the evaluation of their prompted-based tasks.

Another way to evaluate LLMs is to do so on the basis of some of their components, for instance word embeddings. LLMs transform input sentences into contextual vectors of each token constituent. These sensitive context word embeddings are seen as components that are dynamically derived from the LLM. Contextual embeddings can be evaluated in a manner analogous to the way non-contextual and static word embeddings are evaluated. While the latter are evaluated intrinsically on subtasks searching for word similarity and analogy completion out of context, the former can be evaluated by means of tasks that measure both in-context word similarity and sentence similarity. For LLMs, these tasks are simpler and faster than extrinsic evaluations, since they do not require supervision or fine-tuning, and allow us to directly check the quality of the model that generated the contextual embeddings. It should be noted that, even though this type of evaluation is known as intrinsic evaluation of embeddings, it is not actually intrinsic for the LLM from which the embeddings are derived. To avoid terminological confusion, we will call it *vector-based evaluation* of LLMs.

Importantly, contextual embeddings can be generated not only from LLMs, but also by compositional techniques that combine static embeddings, as described in numerous works on compositional distributional semantics (Baroni, 2013; Weir et al., 2016; Gamallo et al., 2019; Wijnholds, Sadrzadeh, and Clark, 2020). In some of these approaches, static embeddings representing the meaning of words in a sentence are combined by syntactic dependencies in an entirely compositional manner, resulting in contextualized vectors of each constituent word (Gamallo et al., 2019; Weir et al., 2016).

The aim of this work is to compare contextual embeddings generated from LLMs with those generated by using syntax-based compositional techniques. All embeddings will be evaluated in a word-in-context similarity task. To do so a very specific dataset is needed because the compositional techniques, due to their linguistic complexity, can only be evaluated on controlled and simple syntactic constructions (e.g. adjective-noun, noun-verb, noun-verb-noun, etc). For this purpose, we created a syntactically controlled dataset in Galician language. In sum, the main contributions of the paper are the following:

- Creation of a new Galician dataset to perform word-in-context similarity tasks.
- Vector-based evaluation (via contextualized word embeddings) of four different LLMs, namely three BERT monolingual models for Galician, and the official multilingual one (mBERT).
- Evaluation of dependency-based compositional vectors generated from Galician Wikipedia.
- Comparison of the performance of all these dynamic and contextually sensitive embeddings against the same dataset.

The rest of the article is organized as follows. The next section introduces some related work (2). Then, the different types of language models, both LLMs and dependency-based, are defined in Section 3. The results and the dataset used in the evaluation are described and analyzed in Section 4. Finally, the conclusions are presented in Section 5.

2 Related work

The first well-known datasets to evaluate contextualized vectors in controlled syntactic constructions are those described in Mitchell and Lapata (2008; 2010). The authors did

not actually use the term *contextualized vectors* for what they called the representation of the meaning of sentences in vector space by means of vector composition. In their work, the meaning of phrases or sentences is represented as the combination of constituent word vectors together with arithmetic operations such as addition and component-wise multiplication. The main drawback of this approach is that it is not fully compositional because word order and syntactic functions are not taken into account. The dataset created by Mitchell and Lapata (2008) in order to evaluate vector composition contains pairs of intransitive English sentences (subject-verb constructions) differing only in the verb. In Mitchell and Lapata (2010) the dataset contains pairs of verb-object constructions differing also in the verb.

Later, Grefenstette et al. (2011a) and Kartsaklis and Sadrzadeh (2013) built very similar evaluation datasets, always for English. These datasets also consist of pairs of sentences, all of which are subject-verb-object transitive constructions that differ only in the verb. Yet, unlike the previous work by Mitchell and Lapata, the semantic approaches that were evaluated on these datasets were full compositional models based on functional words represented as high-dimensional tensors (Baroni, Bernardi, and Zamparelli, 2014). The main concern with these approaches is that they require several high-order tensor representations of verbs with several arguments, something which is computationally inefficient.

To facilitate linguistic preprocessing, all sentences in those datasets are presented as sequences of lemmas, for instance *ball_ricochet* instead of *the ball ricocheted*. Thus, they are not true sentence but n-grams of lemmas representing controlled syntactic constructions.

The increasing development of context-sensitive word embeddings derived from neural language models has marginalized syntax-based and compositional semantic models. One of the main reasons for the low interest in these models is the difficulty to adapt them to open phrases and sentences with any type of syntactic construction. Purely compositional models, due to their linguistic complexity, are so far only successfully applied to datasets with controlled syntactic expressions. In contrast, as context-sensitive embeddings derived from LLMs are built in open syntactic environments, most datasets available and used in shared tasks are composed of open text without syntactic constraints (Pilehvar and Camacho-Collados, 2019; Armendariz et al., 2020). However, it is worth mentioning that the syntactically controlled datasets cited above have been used to compare the two contextual approaches, that is, both compositional embeddings built by means of dependencies and contextual embeddings derived from LLMs, e.g., Wijnholds et al. (2020) and Gamallo et al. (2021) for English, and Gamallo et al. (2021) for Portuguese and Spanish.

There are recent studies which also take advantage of syntactically controlled datasets (such as BiRD (Asaadi, Mohammad, and Kiritchenko, 2019)) to probe the compositional abilities of LLMs. In this respect, Yu and Ettinger (2020) found that Transformer-based models mostly rely on word content, and therefore miss additional information provided by compositional operations.

Finally, recent approaches (Nguyen et al., 2020; Bai et al., 2021) use syntactic information to improve self-attention mechanism, resulting in interesting attempts to include compositional semantic strategies to build the contextualized meaning of words from LLMs.

3 Contextualized Word Vectors from Galician Language Models

Contextualized word vectors can be derived from different types of language models following distributional-based strategies. In our work we explore contextualized word vectors from two types of models for the Galician language, described below: (a) BERT-based LLMs, and (b) transparent models with syntactic dependencies.

3.1 Contextualized Word Vectors from BERT-based Models

Besides the official multilingual model (mBERT, with 12 hidden layers) provided by Devlin et al. (2019), we evaluate the following monolingual models: Bertinho-base, with 12 layers (Vilares, Garcia, and Gómez-Rodríguez, 2021), and two models of Bert-Galician ('base' and 'small') released by Garcia (2021), with 12 and 6 layers, respectively. Concerning the size of the training corpus of each model, mBERT and Bertinho-base were

trained on the Wikipedia, which contains about 42M tokens, while the two versions of Bert-Galician were trained on a larger corpus with about 500M tokens.

To obtain the contextualized vector of a word in the input sentence, we use the standard approach of adding the last four layers, as they have been found to provide more context-specific representations (Ethayarajh, 2019; Vulić et al., 2020). When the tokenizer divides a word into several sub-words (or affixes), only the first subword is considered since it represents the lexical stem of the full token.

We also generate sentence embeddings from LLMs by making use a pooling strategy. This is the same strategy used by Sentence-BERT for English (Reimers and Gurevych, 2019). The main difference with regard to Sentence-BERT is that the Galician pretrained models of our experiments are not fine-tuned with annotated collections of semantically similar pairs of sentences. The basic pooling strategy used to generate our sentence embeddings consists of computing the mean of all output vectors.

3.2 Contextualized Word Vectors from a Galician Dependency-Based Model

3.2.1 Selectional Preferences

Dependency-based distributional models, also known as structured vector spaces, allow us to directly deal with issues related to semantic compositionality and selectional preferences between syntactically related words. To build such a syntax-based model in a transparent way, we opt for a count-based strategy with explicit and sparse dimensions representing lexical-syntactic contexts of words. For instance, given the dependency $(obj, catch, ball)$, representing the heading verb *catch* occurring with dependent noun *ball* in the direct object relation (obj), we extract two lexical-syntactic contexts: either being a dependent noun occurring with *catch* in *obj* relation, or being a verbal head occurring with *ball* in *obj* relation.

The high number of dimensions (lexical-syntactic contexts) of the vector space is reduced by selecting the N most relevant contexts per word (Biemann and Riedl, 2013; Padró et al., 2014; Gamallo, 2017), where N is a global, arbitrarily defined constant whose usual values range from 100 to 1000 (Padró et al., 2014). The relevance value of a context with regard to a word is computed by means of a lexical association measure (e.g., pointwise mutual information, loglikelihood, etc.). This is an explicit, transparent, and static representation of word meaning, very similar to the predictive-based and also static (i.e. out of context) representation known as word embeddings (Mikolov, Yih, and Zweig, 2013).

In order to build contextualized word vectors from these dependency-based representations, we follow the concept of selectional preference formalized in Erk and Padó (2008), which states that the two words related by a dependency relation impose restrictions on each other. Let $A(obj, catch, ball)$ denote the lexical association A between verbal head *catch* and dependent noun *ball* via relation *obj* in a parsed corpus, then the selectional preferences, noted h and d, imposed by the two lemmas on each other in relation *obj* are computed in equations 1 and 2.

$$\vec{h}_{ball}(obj) = \sum_{h:A(obj,h,ball)>\theta} \vec{h} \qquad (1)$$

$$\vec{d}_{catch}(obj) = \sum_{d:A(obj,catch,d)>\theta} \vec{d} \qquad (2)$$

Where $h : A(obj, h, ball) > \theta$ is the set of heading verbs (e.g., *catch, throw, organize...*) that have *ball* as *obj* with a lexical association value higher than threshold θ, and $d : A(obj, catch, d) > \theta$ is the set of dependent nouns (e.g., *ball, baseball, cold, drift...*) occurring with *catch* via *obj* with an association value higher than θ. Note that the former represents the paradigmatic class of those relevant verbs having *ball* as direct object, while the latter is the paradigmatic class of relevant nouns appearing as direct objects of *catch*. In both cases, the selectional preferences imposed by the two related lemmas result in two new compositional vectors, $\vec{h}_{ball}(obj)$ and $\vec{d}_{catch}(obj)$, created by the iterative sum of the static vectors, respectively noted $\vec{h}$ and $\vec{d}$, of the paradigmatic classes (see equations 1 and 2).

Once the selectional preferences are built, they are combined by component-wise multiplication with the static vectors of both the head and dependent lemmas, giving rise to

two new contextualized vectors: the vector of the head lemma, $\vec{catch}_{(obj,h,ball)}$, contextualized with the selectional preferences of *ball* (equation 3), and the vector of the dependent lemma, $\vec{ball}_{(obj,catch,d)}$, contextualized with the selectional preferences of *catch* (equation 4).

$$\vec{catch}_{(obj,h,ball)} = \vec{catch} \odot \vec{h}_{ball}(obj) \quad (3)$$

$$\vec{ball}_{(obj,catch,d)} = \vec{ball} \odot \vec{d}_{catch}(obj) \quad (4)$$

At the end of this compositional process, the two contextualized vectors represent the in-context meaning of the two related words, which are more precise than the out-of-context meaning of the initial static vectors: *catch* means in the context of *ball* some event similar to *grab*, and not to *contract* as in *catch a disease*, while *ball* means in the context of *catch* a spherical object and not and dancing event as in *attend a ball*. In sum, these two contextualized vectors represent discriminated and disambiguated word senses.

3.2.2 Incremental Contextualization

So far we have defined the process of compositional semantics between two dependent words, but this process can be extended to the sentence level. Given the dependency parse tree of an input sentence, the contextualization of all constituent words in the sentence is the result of applying the compositional operations carried out by all dependencies identified in the parse tree in an iterative and incremental way. Thus, at the end of the process, each word of the sentence, including the root one, is assigned a contextualized vector. The order in which compositional operations are applied is not predetermined and the incremental and iterative process can go either from left-to-right or from right-to-left.

3.2.3 Galician Model

To build the language model and their corresponding vector space, the Galician Wikipedia (dump file of November 2019) was parsed with LinguaKit (Gamallo et al., 2018).[1] The Linguakit module used for this purpose is *dep*(endencies), which in turn makes use of the PoS tagger and lemmatizer modules. Since it is a small corpus containing about 42.7 million tokens, we used lemmas as the main lexical unit. Lemmas appearing less than 100 times were filtered out, and lexico-syntactic contexts with frequency less than 50 were removed. Then, for each lemma, we selected the 500 most relevant lexico-syntactic contexts by means of loglikelihood as lexical association measure. The final model resulted in a non-zero matrix of about 50k different lemmas and over 33k different contexts. In total, the vector space consists of a non-zero matrix with about 4.251 million word-context pairs. All static vectors of out-of-context Galician words are derived from this language space. See Gamallo (2017) for more details on how dependency-based vectors are built.

The software used to dynamically build contextualized word vectors from the Galician static vectors is freely available.[2] This is an improved upgrade we implemented on the basis of an older version that was fully described in Gamallo (2019). For the Galician corpus, the θ parameter was set to 0. Previous experiments did not show any improvement by assigning positive values to this parameter. It means that the second selection of relevant contexts made by this parameter is not justified with small corpus sizes such as the one used here.

Although the compositional strategy is designed to work with any type of sentence, due to the difficulty of the task, the implemented version only applies to linguistic expressions with a fixed and predefined syntactic structure (e.g., adjective-noun, subject-verb-object, and son on).

4 Evaluation

In order to compare contextual embeddings generated from BERT-based models with those generated with the compositional dependency-based strategy, we use a word-in-context similarity task in Galician. For this purpose, we created a syntactically controlled Galician dataset with subject-verb-object sentences.

4.1 Test Dataset

Following the structure of the English dataset described in Grefenstette and Sadrzadeh (2011a), a new Galician dataset with 192 sentence pairs of subject-verb-object sentences

[1] https://github.com/citiususc/Linguakit

[2] https://github.com/gamallo/DepFunc

was built.[3]

As it is not possible to make a direct translation of the original English sentences, since the selection preferences are very different from one language to another, we chose analogous examples with 68 different polysemous verbs and 149 different nouns in subject and object position. All sentences consist of just one basic nominal phrase as subject, a verb as predicate, and a basic nominal phrase as direct object.

In each pair, one transitive sentence consisting of a verb with its subject and direct object is compared to another transitive sentence combining the same subject and object with a semantically related verb that is chosen to be either appropriate or inappropriate in the same context. For instance, *a empresa compra un político* ('the company buys a politician') is semantically appropriate and very close to *a empresa suborna un político* ('the company bribes a politician') as *comprar* ('buy') is a very close synonym of *subornar* ('bribe') in this context, where the subject is a person or organization and the object is also a person or organization. However, the same pair of verbs have a very dissimilar behavior in a different context, e.g., *o director compra unha acción / ??o director suborna unha acción* ('the director buys a share' / '??the director bribes a stock'), as the verb *subornar* ('bribe') cannot be applied on objects that are not provided with the human feature. The selectional preferences imposed by that verb are not fulfilled by the direct object.

Unlike the original English dataset from which it is inspired, we created complete sentences, and not just triples of lemmas. However, all sentences were also lemmatized to enable to be evaluated with the dependency-based approach. Most verbs and their arguments were adapted (and not literally translated) to Galician from the English original dataset.

Three native speakers of Galician (and expert linguists) were asked to rate the degree of semantic correctness and similarity of each sentence pair using a 1 to 7 Likert scale. The average scores per annotator are 4.22, 3.45 and 3.94, with the following standard deviations: 1.96, 2.02 and 1.97, respectively. In order to measure the reliability of the ratings provided by the annotators, we calculated an intraclass correlation coefficient (ICC) using the *irr* package in R (Gamer et al., 2019). The agreement ICC was 0.71, indicating a high reliability among raters. Then, the average of the three scores per pair was computed.

As in many intrinsic evaluations of word embeddings, we compute Spearman correlation between human scores (the average of the three evaluators) and the predictions returned by the systems. Both human evaluators and systems should provide high scores to semantically similar sentence pairs with a high degree of semantic correctness.

4.2 Types of Sentence Similarity

Contextualized word embeddings are powerful semantic artifacts that can be used to measure the similarity between two sentences from different points of view. In the following subsections, we define different types of sentence similarity depending on which is the most representative constituent of the sentence.

4.2.1 BERT Sentence Similarity

As all constituent words are fully contextualized, we assume that any of them can represent the whole sentence semantically from a specific point of view. For example, in the sentence *the president signed the decree*, in addition to the verb, the contextualized subject refers to the president who signed the decree, while the contextualized direct object designates the decree that is signed by the president. So, in a transitive sentence, each of the three contextualized word vectors (subject, verb, or object) might be used to compute similarity at the sentence level (and not just at the word level). Moreover, it is also possible to build a new vector representing the whole sentence by combining the embeddings of its constituent words. In total, we can build the following four vectors:

BERT - verb : Contextualized vector of the verb head, resulting from adding the 4 last layers.

BERT - subj : Contextualized vector of the subject word, resulting from adding the 4 last layers.

BERT - obj : Contextualized vector of the direct object, resulting from adding the 4 last layers.

[3]The dataset is available with the software (Cf. footnote 2).

BERT - sentence : Mean of all output vectors.

Note that for English, Sentence-BERT (Reimers and Gurevych, 2019) generates fixed sized vectors of sentences in a way that is similar to our BERT-sentence strategy. There are, however, two significant differences: Sentence-BERT was derived from BERT-large (with 24 layers) and was fine-tuned with two very large dataset collections: SNLI (Bowman et al., 2015) and MultiNLI (Williams, Nangia, and Bowman, 2018) containing 1 million sentence pairs which were annotated for semantic tasks such as inference, contradiction, and entailment. So, while Sentence-BERT is a fine-tuned model trained with a supervised technique on annotated corpora, BERT-sentence is a fully unsupervised model.

4.2.2 Dependency-Based Sentence Similarity

This strategy builds compositional vectors in an incremental way. Thus, it is sensitive to the order of application of the identified syntactic dependencies. The semantic meaning of *The company buys the politician* can be interpreted either from left to right (see 5 below) or from right to left (see 6), according to the order in which the two dependencies of the sentence are applied:

$$(nsubj, buy, company), (obj, buy, politician) \quad (5)$$

$$(obj, buy, politician), (nsubj, buy, company) \quad (6)$$

Then, considering the direction of the compositional process, several compositional vectors representing the meaning of the transitive sentence are built:

left-to-right - verb : This builds the compositional vector of the verb head *buy*. It results from being contextualized first by the selectional preferences imposed by the nominal subject *company* and then by the selectional preferences of the direct object *politician*.

left-to-right - obj : This builds the compositional vector of the direct object *politician*. It results from being contextualized by the preferences imposed by *buy* previously combined with the subject *company*.

left-to-right - sentence : The addition of the two previous left-to-right values (head and dep).

right-to-left - verb : This builds the compositional vector of the verb head *buy*. It results from being contextualized first by the selectional preferences imposed by the direct object *politician* and then by the selectional preferences of the subject *company*.

right-to-left - subj : This builds the compositional vector of the subject *company*. It results from being contextualized by the preferences imposed by *buy* previously combined with the direct object *politician*.

right-to-left - sentence : The addition of the two previous right-to-left values (head and dep).

Note that, in the left-to-right direction, the object is fully contextualized by the verb and the subject. By contrast, the subject is not contextualized by the object, so that this partially contextualized sense of the subject is not used to represent the sentence. The same occurs for the direct object in the right-to-left compositional processes. The incremental direction is not relevant in BERT LLMs because the BERT-like strategy relies on bidirectional scanning by jointly conditioning on both left and right context in all layers. Hence, any constituent word is contextualized by the other words in both left-to-right and right-to-left direction.

4.3 Results

All the models and their different vector configurations were evaluated using the Galician dataset described above.

Table 1 shows the results of the four BERT models for the four types of contextualized vectors introduced in subsection 4.2.1, by using Spearman correlation between the system scores and the human evaluators. We observe that the most significant element of the meaning of the sentence is the contextualized sense of the verb -something expected because the verb is the syntactic root. The verb provides even better results than the sentence method, as a representative of the whole sentence, in three of the four models. By contrast, the subject tends to be the least significant constituent. Perhaps this might

be explained by the fact that in transitive constructions the object is mostly determined by the verb (through more restrictive selection preferences) than by the subject.

If we focus on the comparison of the four LLMs, we observe that the best one is clearly BERT-base (57 for the verb), followed by BERT-small (46). The big differences between these two LLMs and Bertinho-base (21) and mBERT (25) are quite remarkable.

Table 2 shows the results obtained with different contextualized vectors derived from the dependency-based model (see subsection 4.2.2). In the first row, the table also shows a non-compositional baseline strategy just comparing the similarity of verb vectors out of context.

As it was the case with BERT models, the verb is also the most representative constituent of the meaning of the sentence: it achieves 47 and 41 correlation in the two directions, compared to only 18 and 15 for subject and object respectively. It follows that the verbal root, once contextualized by the sense of the arguments, can be taken as the meaning of the whole sentence.

Although they are not totally comparable, we also show in the last rows of the table the best values obtained by compositional systems applied to the English dataset described in (Grefenstette and Sadrzadeh, 2011b) and from which we have created the Galician one. The values obtained on the Galician dataset by using BERT-Base-Galician outperform all the compositional methods for English, including the highest score, 54, obtained by the system described in (Wijnholds, Sadrzadeh, and Clark, 2020).

Models	ρ
BERT-base-Galician - sentence	54
BERT-base-Galician- verb	**57**
BERT-base-Galician - subj	33
BERT-base-Galician - obj	49
BERT-small-Galician- sentence	**46**
BERT-small-Galician - verb	43
BERT-small-Galician - subj	28
BERT-small-Galician - obj	36
mBERT - sentence	21
mBERT - verb	**25**
mBERT - subj	23
mBERT - obj	24
Bertinho-base - sentence	9
Bertinho-base - verb	**21**
Bertinho-base - subj	6
Bertinho-base - obj	9

Table 1: Spearman correlation between different configurations of BERT and human judgments on 192 subject-verb-object sentence pairs.

Models	ρ
baseline - verb	28
left-to-right - sentence	37
left-to-right - verb	**47**
left-to-right - obj	15
right-to-left - sentence	32
right-to-left - verb	41
right-to-left - subj	18
Hashimoto and Tsuruoka (2014)	43 (en)
Polajnar et al. (2015)	35 (en)
Wijnholds et al. (2020)	54 (en)

Table 2: Spearman correlation between different configurations of the compositional dependency-based method and human judgments on 192 subject-verb-object sentence pairs (in lemmas). The table also shows a baseline based on just comparing verbs out-of-context (first row) and some related evaluations on a quite similar dataset for English (three last rows).

4.4 Discussion

The analysis of the results presented in the previous section leads us to draw some conclusions about the strategies compared in the experiments.

BERT-base-Galician and BERT-small-Galician models clearly outperform both Bertinho-base and mBERT in the proposed semantic task. It is also important to point out that the best scores of both Bertinho-base and mBERT are still below the baseline (28).

Concerning the dependency-based strategy, its results are comparable to those of BERT-small-Galician, even if they are far from the higher correlation given by BERT-base-Galician. Let us note that the training corpus of the dependency-based method, as well as that of Bertinho-base and mBERT (Galician part) is just the Galician Wikipedia, and this is much smaller than the training corpus used for the two BERT-Galician models: ≈42 million tokens vs. ≈500 million.

As it was already reported, the contextualized sense of the verbal root is the most

representative meaning of the sentence for most strategies. It behaves better than the other constituents (subject and object), and even than computing a global meaning for the whole sentence. This is a very relevant observation as most systems computing the sentence meaning do not know which is the root word because they do not rely on a dependency tree, and so they make use of the vectors of all constituent words.

And finally, we must point out that the correlation values obtained here and in other related experiments for other languages are in low to medium ranges. This shows that this is a semantic task of great complexity that still requires improved language models, perhaps not larger or computationally deeper, but of higher quality and with deeper linguistic knowledge.

5 Conclusions

In this article, we evaluated and compared the performance of contextualized vectors built with LLMs and a fully compositional strategy based on syntactic dependencies and selectional preferences. The use of selectional preferences to build contextualized vectors is a linguistically motivated attention strategy focused on selecting only syntactically relevant contextual elements. It can be seen, therefore, as a mechanism of attention driven by syntactic information.

According to the results obtained, the compositional strategy turned out to be competitive when compared to several configurations of BERT in a specific task focused on sentence similarity in Galician. It should be noted that the computational cost of training compositional models is much lower than that of neural-based LLMs. In addition, the syntax-based vectors we have used for the compositional approach are more transparent and interpretable than those derived from the Transformer architecture. Transparent models make it easier to explain the errors and successes committed in a particular task, since it is possible to explicitly list the syntactic contexts involved in vector composition.

However, the dependency-based strategy has important weaknesses. First, as it mainly relies on syntactic parsing, it has a vulnerable exposure to parser errors. Second, this strategy cannot be easily adapted to syntactically open sentences.

In order to overcome these drawbacks, in future work we will define and implement a syntax-based model allowing us to build fully contextualized vectors for open sentences. This will enable to apply the compostional method to any sentence in as similar way as Transformers do.

Acknowledgements

This research was funded by the project "Nós: Galician in the society and economy of artificial intelligence", agreement between Xunta de Galicia and University of Santiago de Compostela, and grant ED431G2019/04 by the Galician Ministry of Education, University and Professional Training, and the European Regional Development Fund (ERDF/FEDER program), and Groups of Reference: ED431C 2020/21. In addition: Ramón y Cajal grant (RYC2019-028473-I) and Grant ED431F 2021/01 (Galician Government).

References

Armendariz, C. S., M. Purver, S. Pollak, N. Ljubešić, M. Ulčar, M. Robnik-Šikonja, I. Vulić, and M. T. Pilehvar. 2020. SemEval-2020 task 3: Graded word similarity in context (GWSC). In *Proceedings of the 14th International Workshop on Semantic Evaluation*.

Asaadi, S., S. Mohammad, and S. Kiritchenko. 2019. Big BiRD: A large, fine-grained, bigram relatedness dataset for examining semantic composition. 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 505–516, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Bai, J., Y. Wang, Y. Chen, Y. Yang, J. Bai, J. Yu, and Y. Tong. 2021. Syntax-BERT: Improving pre-trained transformers with syntax trees. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3011–3020, Online. Association for Computational Linguistics.

Baroni, M. 2013. Composition in distributional semantics. *Language and Linguistics Compass*, 7:511–522.

Baroni, M., R. Bernardi, and R. Zamparelli. 2014. Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technology (LiLT)*, 9:241–346.

Biemann, C. and M. Riedl. 2013. Text: Now in 2d! a framework for lexical expansion with contextual similarity. *Journal of Language Modelling*, 1(1):55–95.

Bowman, S. R., G. Angeli, C. Potts, and C. D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-2019, Volume 1*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Erk, K. and S. Padó. 2008. A structured vector space model for word meaning in context. In *2008 Conference on Empirical Methods in Natural Language Processing (EMNLP-2008*, pages 897–906, Honolulu, HI.

Ethayarajh, K. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *EMNLP/IJCNLP (1)*, pages 55–65. Association for Computational Linguistics.

Gamallo, P., M. Garcia, C. Piñeiro, R. Martinez-Castaño, and J. C. Pichel. 2018. LinguaKit: A Big Data-Based Multilingual Tool for Linguistic Analysis and Information Extraction. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 239–244.

Gamallo, P. 2017. Comparing explicit and predictive distributional semantic models endowed with syntactic contexts. *Language Resources and Evaluation*, 51(3):727–743.

Gamallo, P. 2019. A dependency-based approach to word contextualization using compositional distributional semantics. *Language Modelling*, 7(1):53–92.

Gamallo, P. 2021. Compositional distributional semantics with syntactic dependencies and selectional preferences. *Applied Sciences*, 11(12).

Gamallo, P., M. P. Corral, and M. Garcia. 2021. Comparing dependency-based compositional models with contextualized word embedding. In *13th International Conference on Agents and Artificial Intelligence (ICAART-2021)*.

Gamallo, P., S. Sotelo, J. R. Pichel, and M. Artetxe. 2019. Contextualized translations of phrasal verbs with distributional compositional semantics and monolingual corpora. *Computational Linguistics*, 45(3):395–421.

Gamer, M., J. Lemon, I. Fellows, and P. Singh, 2019. *irr: Various Coefficients of Interrater Reliability and Agreement*. R package version 0.84.1.

Garcia, M. 2021. Exploring the representation of word meanings in context: A case study on homonymy and synonymy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3625–3640, Online, August. Association for Computational Linguistics.

Grefenstette, E. and M. Sadrzadeh. 2011a. Experimental support for a categorical compositional distributional model of meaning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 1394–1404.

Grefenstette, E. and M. Sadrzadeh. 2011b. Experimenting with transitive verbs in a discocat. In *Workshop on Geometrical Models of Natural Language Semantics (EMNLP 2011)*.

Kartsaklis, D. and M. Sadrzadeh. 2013. Prior disambiguation of word tensors for constructing sentence vectors. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pages 1590–1601.

Lin, B. Y., S. Lee, X. Qiao, and X. Ren. 2021. Common sense beyond english: Evaluating and improving multilingual language models for commonsense reasoning. *CoRR*, abs/2106.06937.

Mikolov, T., W.-t. Yih, and G. Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia.

Mitchell, J. and M. Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT)*, pages 236–244, Columbus, Ohio.

Mitchell, J. and M. Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1439.

Nguyen, X.-P., S. Joty, S. C. H. Hoi, and R. Socher. 2020. Tree-structured attention with hierarchical accumulation.

Padró, M., M. Idiart, A. Villavicencio, and C. Ramisch. 2014. Nothing like good old frequency: Studying context filters for distributional thesauri. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 419–424.

Pilehvar, M. T. and J. Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning 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 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Reimers, N. and I. Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Salazar, J., D. Liang, T. Q. Nguyen, and K. Kirchhoff. 2020. Masked language model scoring. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Shibayama, N., R. Cao, J. Bai, W. Ma, and H. Shinnou. 2020. Evaluation of pretrained BERT model by using sentence clustering. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 279–285, Hanoi, Vietnam, October. Association for Computational Linguistics.

Vilares, D., M. Garcia, and C. Gómez-Rodríguez. 2021. Bertinho: Galician BERT Representations. *Procesamiento del Lenguaje Natural*, 66:13–26.

Vulić, I., E. M. Ponti, R. Litschko, G. Glavaš, and A. Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online, November. Association for Computational Linguistics.

Wang, A., A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR 2019*.

Weir, D. J., J. Weeds, J. Reffin, and T. Kober. 2016. Aligning packed dependency trees: A theory of composition for distributional semantics. *Computational Linguistics*, 42(4):727–761.

Wijnholds, G., M. Sadrzadeh, and S. Clark. 2020. Representation learning for type-driven composition. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 313–324, Online. Association for Computational Linguistics.

Williams, A., N. Nangia, and S. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yu, L. and A. Ettinger. 2020. Assessing phrasal representation and composition in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4896–4907, Online, November. Association for Computational Linguistics.