

Exploring the Representation of Word Meanings in Context: A Case Study on Homonymy and Synonymy

Marcos Garcia

CiTIUS – Research Center in Intelligent Technologies

Universidade de Santiago de Compostela, Galiza

marcos.garcia.gonzalez@usc.gal

Abstract

This paper presents a multilingual study of word meaning representations in context. We assess the ability of both static and contextualized models to adequately represent different lexical-semantic relations, such as homonymy and synonymy. To do so, we created a new multilingual dataset that allows us to perform a controlled evaluation of several factors such as the impact of the surrounding context or the overlap between words, conveying the same or different senses. A systematic assessment on four scenarios shows that the best monolingual models based on Transformers can adequately disambiguate homonyms in context. However, as they rely heavily on context, these models fail at representing words with different senses when occurring in similar sentences. Experiments are performed in Galician, Portuguese, English, and Spanish, and both the dataset (with more than 3,000 evaluation items) and new models are freely released with this study.

1 Introduction

Contrary to static vector models, which represent the different senses of a word in a single vector (Erk, 2012; Mikolov et al., 2013), contextualized models generate representations at token-level (Peters et al., 2018; Devlin et al., 2019), thus being an interesting approach to model word meaning in context. In this regard, several studies have shown that clusters produced by some contextualized word embeddings (CWEs) are related to different senses of the same word (Reif et al., 2019; Wiedemann et al., 2019), or that similar senses can be aligned in cross-lingual experiments (Schuster et al., 2019).

However, more systematic evaluations of polysemy (i.e., word forms that have different related meanings depending on the context (Apresjan, 1974)), have shown that even though CWEs present some correlations with human judgments

(Nair et al., 2020), they fail to predict the similarity of the various senses of a polysemous word (Haber and Poesio, 2020).

As classical datasets to evaluate the capabilities of vector representations consist of single words without context (Finkelstein et al., 2001) or heavily constrained expressions (Kintsch, 2001; Mitchell and Lapata, 2008), new resources with annotations of words in free contexts have been created, including both graded similarities (Huang et al., 2012; Armendariz et al., 2020) or binary classification of word senses (Pilehvar and Camacho-Collados, 2019; Raganato et al., 2020). However, as these datasets largely include instances of polysemy, they are difficult to solve even for humans (in fact, the highest reported human upper bound is about 80%) as the nuances between different senses depend on non-linguistic factors such as the annotator procedure or the target task (Tuggy, 1993; Kilgarriff, 1997; Hanks, 2000; Erk, 2010).

In this paper, we rely on a more objective and simple task to assess how contextualized approaches (both neural network models and contextualized methods of distributional semantics) represent word meanings in context. In particular, we observe whether vector models can identify unrelated meanings represented by the same word form (homonymy) and the same sense conveyed by different words (synonymy). In contrast to polysemy, there is a strong consensus concerning the representation of homonymous senses in the lexicon, and it has been shown that homonyms are cognitively processed differently than polysemous words (Klepousniotou et al., 2012; MacGregor et al., 2015). In this regard, exploratory experiments in English suggest that some CWEs correctly model homonymy, approximating the contextualized vectors of a homonym to those of its paraphrases (Lake and Murphy, 2020), and showing stronger correlation with human judgments to those

of polysemous words (Nair et al., 2020). However, as homonyms convey unrelated meanings depending on the context, it is not clear whether the good performance of CWEs actually derives from the contextualization process or simply from the use of explicit lexical cues present in the sentences.

Taking the above into account, we have created a new multilingual dataset (in Galician, Portuguese, English, and Spanish) with more than 3,000 evaluation items. It allows for carrying out more than 10 experiments and controlling factors such as the surrounding context, the word overlap, and the sense conveyed by different word forms. We use this resource to perform a systematic evaluation of contextualized word meaning representations. We compare different strategies using both static embeddings and current models based on deep artificial neural networks. The results suggest that the best monolingual models based on Transformers (Vaswani et al., 2017) can identify homonyms having different meanings adequately. However, as they strongly rely on the surrounding context, words with different meanings are represented very closely when they occur in similar sentences. Apart from the empirical conclusions and the dataset, this paper also contributes with new BERT and *fastText* models for Galician.¹

Section 2 presents previous studies about word meaning representation. Then, Section 3 introduces the new dataset used in this paper. In Section 4 we describe the models and methods to obtain the vector representations. Finally, the experiments and results are discussed in Section 5, while Section 6 draws some conclusions of our study.

2 Related Work

A variety of approaches has been implemented to compute word meaning in context by means of standard methods of distributional semantics (Schütze, 1998; Kintsch, 2001; McDonald and Brew, 2004; Erk and Padó, 2008). As compositional distributional models construct sentence representations from their constituents vectors, they take into account contextualization effects on meaning (Mitchell and Lapata, 2008; Baroni and Zamparelli, 2010; Baroni, 2013). However, these approaches often have scalability problems as their representations grow exponentially with the size of the sentences. Therefore, the datasets used to

evaluate them are composed of highly restricted phrases (Grefenstette and Sadrzadeh, 2011).

The rise of artificial neural networks on natural language processing popularized the use of vector representations, and the remarkable performance of neural language models (Melamud et al., 2016; Peters et al., 2018) led to a productive line of research exploring to what extent these models represent linguistic knowledge (Rogers et al., 2020). However, few of these works have focused on lexical semantics, and most of the relevant results in this field come from evaluations in downstream tasks. In this regard, Wiedemann et al. (2019) found that clusters of BERT embeddings (Devlin et al., 2019) seem to be related to word senses, while Schuster et al. (2019) observed that clusters of polysemous words correspond to different senses in a cross-lingual alignment of vector representations.

Probing LSTMs on lexical substitution tasks, Aina et al. (2019) showed that these architectures rely on the lexical information from the input embeddings, and that the hidden states are biased towards contextual information. On an exploration of the geometric representations of BERT, Reif et al. (2019) found that different senses of a word tend to appear separated in the vector space, while several clusters seem to correspond to similar senses. Recently, Vulić et al. (2020) evaluated the performance of BERT models on several lexical-semantic tasks in various languages, including semantic similarity or word analogy. The results show that using special tokens ([CLS] or [SEP]) hurts the quality of the representations, and that these tend to improve across layers until saturation. As this study uses datasets of single words (without context), type-level representations are obtained by averaging the contextualized vectors over various sentences.

There are several resources to evaluate word meaning in free contexts, such as the Stanford Contextual Word Similarity (Huang et al., 2012) and CoSimLex (Armendariz et al., 2020), both representing word similarity on a graded scale, or the Word-in-Context datasets (WiC), focused on binary classifications (i.e., each evaluation item contains two sentences with the same word form, having the same or different senses) (Pilehvar and Camacho-Collados, 2019; Raganato et al., 2020). These datasets include not only instances of homonymy but mostly of polysemous words. In this regard, studies on polysemy using Transformers have obtained diverse results: Haber and Poesio (2020)

¹Dataset, models, and code are available at https://github.com/marcospln/homonymy_acl21/.

found that BERT embeddings correlate better with human ratings of co-predication than with similarity between word senses, thus suggesting that these representations encode more contextual information than word sense knowledge. Nevertheless, the results of Nair et al. (2020) indicate that BERT representations are correlated with human scores of polysemy. An exploratory experiment of the latter study also shows that BERT discriminates between polysemy and homonymy, which is also suggested by other pilot evaluations reported by Lake and Murphy (2020) and Yu and Ettinger (2020).

Our study follows this research line pursuing objective and unambiguous lexical criteria such as the representation of homonyms and synonyms. In this context, there is a broad consensus in the psycholinguistics literature regarding the representation of homonyms as different entries in the lexicon (in contrast to polysemy, for which there is a long discussion on whether senses of polysemous words are stored as a single core representation or as independent entries (Hogeweg and Vicente, 2020)). In fact, several studies have shown that homonyms are cognitively processed differently from polysemous words (Klepousniotou et al., 2012; Rabagliati and Snedeker, 2013). In contrast to the different senses of polysemous words, which are simultaneously activated, the meanings of homonyms are in conflict during processing, with the not relevant ones being deactivated by the context (MacGregor et al., 2015). To analyze how vector models represent homonymy and synonymy in context, we have built a new multilingual resource with a strong inter-annotator agreement, presented below.

3 A New Multilingual Resource of Homonymy and Synonymy in Context

This section briefly describes some aspects of lexical semantics relevant to our study, and then presents the new dataset used in the paper.

Homonymy and homography: Homonymy is a well-known type of lexical ambiguity that can be described as the relation between distinct and unrelated meanings represented by the same word form, such as *match*, meaning for instance ‘sports game’ or ‘stick for lighting fire’. In contrast to polysemy (where one lexeme conveys different related senses depending on the context, e.g., *newspaper* as an organization or as a set of printed pages), it is often assumed that homonyms are different lexemes that have the same lexical form (Cruse, 1986), and

therefore they are stored as independent entries in the lexicon (Pustejovsky, 1998).

There are two main criteria for homonymy identification: Diachronically, homonyms are lexical items that have different etymologies but are accidentally represented by the same word form, while a synchronic perspective strengthens unrelatedness in meaning. Even if both approaches tend to identify similar sets of homonyms, there may be ambiguous cases that are diachronically but not synchronically related (e.g., two meanings of *banco* – ‘bench’ and ‘financial institution’ – in Portuguese or Spanish could be considered polysemous as they derive from the same origin,² but as this is a purely historical association, most speakers are not aware of the common origin of both senses). In this study, we follow the synchronic perspective, and consider homonymous meanings those that are clearly unrelated (e.g., they unambiguously refer to completely different concepts) regardless of their origin.

It is worth mentioning that as we are dealing with written text we are actually analyzing homographs (different lexemes with the same spelling) instead of homonyms. Thus, we discard instances of phonologically identical words which are written differently, such as the Spanish *hola* ‘hello’ and *ola* ‘wave’, both representing the phonological form /ola/. Similarly, we include words with the same spelling representing different phonological forms, e.g., the Galician-Portuguese *sede*, which corresponds to both /sede/ ‘thirst’, and /sɛde/ ‘headquarters’.

In this paper, *homonymous senses* are those unrelated meanings conveyed by the same (homonym) word form. For instance, *coach* may have two homonymous senses (‘bus’ and ‘trainer’), which can be conveyed by other words (synonyms) in different contexts (e.g., by *bus* or *trainer*).

Structure of the dataset: We have created a new resource to investigate how vector models represent word meanings in context. In particular, we want to observe whether they capture (i) different senses conveyed by the same word form (homonymy), and (ii) equivalent senses expressed by different words (synonymy). The resource contains controlled sentences so that it allows us to observe how the context and word overlap affect word representations.

To allow for different comparisons with the same

²In fact, several dictionaries organize them in a single entry: <https://dicionario.priberam.org/banco>, <https://dle.rae.es/banco>.

Sense	Sentences 1-3	Sentence 4	Sentence 5
(1)	We're going to the airport by coach . We're going to the airport by bus . We're going to the airport by <i>bicycle</i> .	[...] the coach was badly delayed by roadworks.	They had to travel everywhere by bus .
(2)	That man was appointed as the new coach . That man was appointed as the new trainer . That man was appointed as the new <i>president</i> .	She has recently joined the amateur team as coach .	They need a new trainer for the young athletes.

Table 1: Example sentences for two senses of *coach* in English ('bus' and 'trainer'). Sentences 1 to 3 include, in the same context, the target word, a synonym, and a word with a different sense (in italic), respectively. Sentences 4 and 5 contain the target word and a synonym in different contexts, respectively.

and different contexts, we have included five sentences for each meaning (see Table 1 for examples): three sentences containing the target word, a synonym, and a word with a different sense, all of them in the same context (sentences 1 to 3), and two additional sentences with the target word and a synonym, representing the same sense (sentences 4 and 5, respectively). Thus, for each sense we have four sentences (1, 2, 4, 5) with a word conveying the same sense (both in the same and in different contexts) and another sentence (3) with a different word in the same context as sentences 1 and 2.

From this structure, we can create datasets of sentence triples, where the target words of two of them convey the same sense, and the third one has a different meaning. Thus, we can generate up to 48 triples for each pair of senses (24 in each direction: sense 1 vs. sense 2, and vice-versa). These datasets allow us to evaluate several semantic relations at the lexical level, including homonymy, synonymy, and various combinations of homonymous senses. Interestingly, we can control for the impact of the context (e.g., are contextualized models able to distinguish between different senses occurring in the same context, or do they incorporate excessive contextual information into the word vectors?), the word overlap (e.g., can a model identify different senses of the same word form depending on the context, or it strongly depends on lexical cues?), or the POS-tag (e.g., are homonyms with different POS-tags easily disambiguated?).

Construction of the dataset: We compiled data for four languages: Galician, Portuguese, Spanish, and English.³ We tried to select sentences compatible with the different varieties of the same language

³Galician is generally considered a variety of a single (Galician-)Portuguese language. However, they are divided in this resource, as Galician has recently been standardized using a Spanish-based orthography that formally separates it from Portuguese (Samartim, 2012).

(e.g., with the same meaning in UK and US English, or in Castilian and Mexican Spanish). However, we gave priority to the European varieties when necessary (e.g., regarding spelling variants).

The dataset was built using the following procedure: First, language experts (one per language) compiled lists of homonyms using dedicated resources for language learning, together with WordNet and other lexicographic data (Miller, 1995; Montraveta and Vázquez, 2010; Guinovart, 2011; Rademaker et al., 2014). Only clear and unambiguous homonyms were retained (i.e., those in the extreme of the *homonymy-polysemy-vagueness* scale (Tuggy, 1993)). These homonyms were then enriched with frequency data from large corpora: Wikipedia and SLI GalWeb (Agerri et al., 2018) for Galician, and a combination of Wikipedia and Europarl for English, Spanish and Portuguese (Koehn, 2005). From these lists, each linguist selected the most frequent homonyms, annotating them as ambiguous at type or token level (*absolute homonymy* and *partial homonymy* in Lyons' terms (Lyons, 1995)). As a substantial part were noun-verb pairs, only a few of these were included. For each homonym, the language experts selected from corpora two sentences (1 and 4) in which the target words were not ambiguous.⁴ They then selected a synonym that could be used in sentence 1 without compromising grammaticality (thus generating sentence 2), and compiled an additional sentence for it (5), trying to avoid further lexical ambiguities in this process.⁵ For each homonym, the linguists selected a word with a different meaning (for sen-

⁴Sentences were selected, adapted, and simplified using GDEX-inspired constraints (Kilgarriff et al., 2008) (i.e., avoiding high punctuation ratios, unnecessary subordinate clauses, etc.), which resulted in the creation of new sentences.

⁵In most cases, this synonym is the same as that of sentence 2, but this is not always the case. Besides, in some cases we could not find words conveying the same sense, for which we do not have sentences 2 and 5.

Language	Hom.	Senses	Sent.	Triples	Pairs	WiC	κ
Galician	22	47 (4)	227	1365	823	197	0.94
English	14	30 (5)	138	709	463	129	0.96
Portuguese	11	22 (1)	94	358	273	81	0.96
Spanish	10	23 (3)	105	645	391	101	0.95
Total	57	122	564	3077	1950	508	0.94

Table 2: Characteristics of the dataset. First three columns display the number of homonyms (*Hom*), senses, and sentences (*Sent*), respectively. Senses in parentheses are the number of homonymous pairs with different POS-tags). Center columns show the size of the evaluation data in three formats: triples, pairs, and WiC-like pairs, followed by the Cohen’s κ agreements and their micro-average. The total number of homonyms and senses is the sum of the language-specific ones, regardless of the fact that some senses occur in more than one language.

tence 3), trying to maximize the following criteria: (i) to refer unambiguously to a different concept, and to preserve (ii) semantic felicity and (iii) grammaticality. The size of the final datasets varies depending on the initial lists and on the ease of finding synonyms in context.

Results: Apart from the sentence triples explained above, the dataset structure allows us to create evaluation sets with different formats, such as sentence pairs to perform binary classifications as in the WiC datasets. Table 2 shows the number of homonyms, senses, and sentences of the multilingual resource, together with the size of the evaluation datasets in different formats.

As the original resource was created by one annotator per language, we ensured its quality as follows: We randomly extracted sets of 50 sentence pairs and gave them to other annotators (5 for Galician, and 1 for each of the other three varieties, all of them native speakers of the target language). We then computed the Cohen’s κ inter-annotator agreement (Cohen, 1960) between the original resource and the outcome of this second annotation (see the right column of Table 2). We obtained a micro-average $\kappa = 0.94$ across languages, a result which supports the task’s objectivity. Nevertheless, it is worth noting that few sentences have been carefully modified after this analysis, as it has shown that several misclassifications were due to the use of an ambiguous synonym. Thus, it is likely that the final resource has higher agreement values.

4 Models and Methods

This section introduces the models and procedures to obtain vector representations followed by the evaluation method.

4.1 Models

We have used static embeddings and CWEs based on Transformers, comparing different ways of obtaining the vector representations in both cases:

Static embeddings: We have used skip-gram *fastText* models of 300 dimensions (Bojanowski et al., 2017).⁶ For English and Spanish, we have used the official vectors trained on Wikipedia. For Portuguese, we have used the model provided by Hartmann et al. (2017), and for Galician we have trained a new model (see Appendix C for details).⁷

Contextualized embeddings: We have evaluated multilingual and monolingual models:⁸

Multilingual models: We have used the official multilingual BERT (mBERT cased, 12 layers) (Devlin et al., 2019), XLM-RoBERTa (Base, 12 layers) (Conneau et al., 2020), and DistilBERT (DistilmBERT, 6 layers) (Sanh et al., 2019).

Monolingual models: For English, we have used the official BERT-Base model (uncased). For Portuguese and Spanish, BERTimbau (Souza et al., 2020) and BETO (Cañete et al., 2020) (both cased). For Galician, we trained two BERT models (with 6 and 12 layers; see Appendix C).

4.2 Obtaining the vectors

Static models: These are the methods used to obtain the representations from the static models:

Word vector (WV): Embedding of the target word (homonymous senses with the same word form will have the same representation).

⁶In preliminary experiments we also used *word2vec* and GloVe models, obtaining slightly lower results than *fastText*.

⁷These Portuguese and Galician models obtained better results (0.06 on average) than the official ones.

⁸To make a fair comparison we prioritized *base* models (12 layers), but we also report results for *large* (24 layers) and *6* layers models when available.

Language	Exp1		Exp2		Exp3		Exp4		Total		Full	
Galician	122	105	183	149	278	229	135	135	718	618	1365	1157
English	77	52	89	58	144	91	68	68	378	269	709	494
Portuguese	45	41	37	37	80	74	41	41	203	193	358	342
Spanish	65	49	87	71	146	110	59	59	357	289	645	517

Table 3: Number of instances of each experiment and language. Numbers on the right of each column are those triples where the three target words belong to the same morphosyntactic category (left values are the total number of triples). *Total* are the sums of the four experiments, while *Full* refers to all the instances of the dataset.

Sentence vector (*Sent*): Average embedding of the whole sentence.

Syntax (*Syn*): Up to four different representations obtained by adding the vector of the target word to those of their syntactic heads and dependents. This method is based on the assumption that the syntactic context of a word characterizes its meaning, providing relevant information for its contextualized representation (e.g., in ‘He swims to the bank’, *bank* may be disambiguated by combining its vector with the one of *swim*).⁹ Appendix D describes how heads and dependents are selected.

Contextualized models: For these models, we have evaluated the following approaches:

Sentence vector (*Sent*): Vector of the sentence built by averaging all words (except for the special tokens [CLS] and [SEP]), each of them represented by the standard approach of concatenating the last 4 layers (Devlin et al., 2019).

Word vector (*WV*): Embedding of the target word, combining the vectors of the last 4 layers. We have evaluated two operations: vector concatenation (*Cat*), and addition (*Sum*).

Word vector across layers (*Lay*): Vector of the target word on each layer. This method allows us to explore the contextualization effects on each layer.

Vectors of words split into several sub-words are obtained by averaging the embeddings of their components. Similarly, MWEs vectors are the average of the individual vectors of their components, both for static and for contextualized embeddings.

4.3 Measuring sense similarities

Given a sentence triple where two of the target words (*a* and *b*) have the same sense and the third (*c*) a different one, we evaluate a model as follows (in a similar way as other studies (Kintsch, 2001; Lake and Murphy, 2020)): First, we obtain

⁹We have also evaluated a contextualization method using selectional preferences inspired by Erk and Padó (2008), but the results were almost identical to those of the *WV* approach.

three cosine similarities between the vector representations: $sim1 = \cos(a, b)$; $sim2 = \cos(a, c)$; $sim3 = \cos(b, c)$. Then, an instance is labeled as *correct* if those words conveying the same sense (*a* and *b*) are closer together than the third one (*c*). In other words, $sim1 > sim2$ and $sim1 > sim3$: Otherwise, the instance is considered as *incorrect*.

5 Evaluation

This section presents the experiments performed using the new dataset and discusses their results.

5.1 Experiments

Among all the potential analyses of our data, we have selected four evaluations to assess the behavior of a model by controlling factors such as the context and the word overlap:

Homonymy (Exp1): The same word form in three different contexts, two of them with the same sense (e.g., *coach* in sentences [1:1, 1:4, 2:1]¹⁰ in Table 1). This test evaluates if a model correctly captures the sense of a unique word form in context. **Hypothesis:** Static embeddings will fail as they produce the same vector in the three cases, while models that adequately incorporate contextual cues should correctly identify the outlier sense.

Synonyms of homonymous senses (Exp2): A word is compared with its synonym and with the synonym of its homonym, all three in different contexts (e.g., *coach=bus≠trainer* in [1:1, 1:5, 2:2]). This test assesses if there is a bias towards one of the homonymous senses, e.g., the most frequent one (MacGregor et al., 2015). **Hypothesis:** Models with this type of bias may fail, so as in Exp1, they should also appropriately incorporate contextual information to represent these examples.

Synonymy vs homonymy (Exp3): We compare a word to its synonym and to a homonym, all in

¹⁰First and second digits refer to the sense and sentence ids.

different contexts (e.g., *coach=bus≠coach* in [1:1, 1:5, 2:1]). Here we evaluate whether a model adequately represents both (i) synonymy in context –two word forms with the same sense in different contexts– and (ii) homonymy –one of the former word forms having a different meaning. **Hypothesis:** Models relying primarily on lexical knowledge are likely to represent homonyms closer than synonyms (giving rise to an incorrect output), but those integrating contextual information will be able to model the three representations correctly.

Synonymy (Exp4): Two synonyms in context vs. a different word (and sense), in the same context as, at least, one of the synonyms (e.g., [2:1, 2:2, 2:3]). It assesses to what extent the context affects word representations of different word forms. **Hypothesis:** Static embeddings may pass this test as they tend to represent type-level synonyms closely in the vector space. Highly contextualized models might be puzzled as different meanings (from different words) occur in the same context, so that the models should have an adequate trade-off between lexical and contextual knowledge.

Table 3 displays the number of sentence triples for each experiment as well as the total number of triples of the dataset. To focus on the semantic knowledge encoded in the vectors –rather than on the morphosyntactic information–, we have evaluated only those triples in which the target words of the three sentences have the same POS-tag (numbers on the right).¹¹ Besides, we have also carried out an evaluation on the full dataset.

5.2 Results and discussion

Table 4 contains a summary of the results of each experiment in the four languages. For reasons of clarity, we include only *fastText* embeddings and the best contextualized model (BERT). Results for all models and languages can be seen in Appendix A. BERT models have the best performance overall, both on the full dataset and on the selected experiments, except for Exp4 (in which the outlier shares the context at least with one of the synonyms) where the static models outperform the contextualized representations.

In Exp1 and Exp2, where the context plays a crucial role, *fastText* models correctly labeled between 50%/60% of the examples (depending on

¹¹On average, BERT-base models achieved 0.24 higher results (*Add*) when tested on all the instances (including different POS-tags) of the four experiments.

the language and vector type, with better results for *Sent* and *Syn*). For BERT, the best accuracy surpasses 0.98 (Exp1 in English), with an average across languages of 0.78, and where word vectors outperform sentence representations. These high results and the fact that *WVs* work better in general than *Sent* may be indicators that Transformers are properly incorporating contextual knowledge.

Solving Exp3 requires both dealing with contextual effects and homonymy (as two words have the same form but different meaning) so that static embeddings hardly achieve 0.5 accuracy (*Sent*, with lower results for both *WV* and *Syn*). BERT’s performance is also lower than in Exp1 and Exp2, with an average of 0.67 and *Sent* beating *WVs* in most cases, indicating that the word vectors are not adequately representing the target senses.

Finally, *fastText* obtains better results than BERT on Exp4 (where the outlier shares the context with at least one of the other target words), reaching 0.81 in Spanish with an average across languages of 0.64 (always with *WVs*). BERT’s best performance is 0.41 (in two languages) with an average of 0.42, suggesting that very similar contexts may confound the model.

To shed light on the contextualization process of Transformers, we have analyzed their performance across layers. Figure 1 shows the accuracy curves (vs. the macro-average *Sent* and *WV* vectors of the contextualized and static embeddings) for five Transformers models on Galician, the language with the largest dataset (see Appendix A for equivalent figures for the other languages).

In Exp1 to Exp3 the best accuracies are obtained at upper layers, showing that word vectors appropriately incorporate contextual information. This is true especially for the monolingual BERT versions, as the multilingual models’ representations show higher variations. Except for Galician, Exp1 has better results than Exp2, as the former primarily deals with context while the latter combines contextualization with lexical effects. In Exp3 the curves take longer to rise as initial layers rely more on lexical than on contextual information. Furthermore, except for English (which reaches 0.8), the performance is low even in the best hidden layers (≈ 0.4). In Exp4 (with context overlap between words with different meanings), contextualized models cannot correctly represent the word senses, being surpassed in most cases by the static embeddings.

Finally, we have observed how Transformers rep-

Model	Vec.	Exp1	Exp2	Exp3	Exp4	Macro	Micro	Full
Galician								
BERT-base	Sent	0.695	0.758	0.751	0.178	0.596	0.618	0.727
	Cat	0.705	0.799	0.293	0.422	0.555	0.513	0.699
<i>fastText</i>	Sent	0.562	0.685	0.476	0.141	0.466	0.468	0.618
	WV	0.21	0.564	0	0.526	0.325	0.286	0.461
	Syn (3)	0.533	0.658	0.197	0.185	0.393	0.362	0.567
English								
BERT-base	Sent	0.788	0.655	0.736	0.221	0.6	0.599	0.7
	Add	0.981	0.81	0.758	0.441	0.748	0.732	0.839
<i>fastText</i>	Sent	0.596	0.5	0.505	0.147	0.437	0.431	0.543
	WV	0.308	0.552	0.033	0.574	0.366	0.335	0.48
	Syn (3)	0.442	0.69	0.231	0.176	0.385	0.357	0.546
Portuguese								
BERT-base	Sent	0.683	0.432	0.635	0.22	0.493	0.518	0.564
	Add	0.854	0.541	0.378	0.366	0.535	0.508	0.67
<i>fastText</i>	Sent	0.61	0.622	0.527	0.171	0.482	0.487	0.55
	WV	0.024	0.541	0	0.634	0.3	0.244	0.453
	Syn (3)	0.659	0.459	0.176	0.195	0.372	0.337	0.508
Spanish								
BERT-base	Sent	0.755	0.592	0.536	0.186	0.517	0.516	0.595
	Add	0.857	0.704	0.409	0.441	0.603	0.564	0.74
<i>fastText</i>	Sent	0.449	0.338	0.445	0.085	0.329	0.346	0.429
	WV	0.122	0.62	0.018	0.814	0.393	0.346	0.479
	Syn (3)	0.367	0.577	0.173	0.237	0.339	0.318	0.553

Table 4: Summary of the BERT and *fastText* results. *Macro* and *Micro* refer to the macro-average and micro-average results across the four experiments, respectively. *Full* are the micro-average values on the whole dataset.

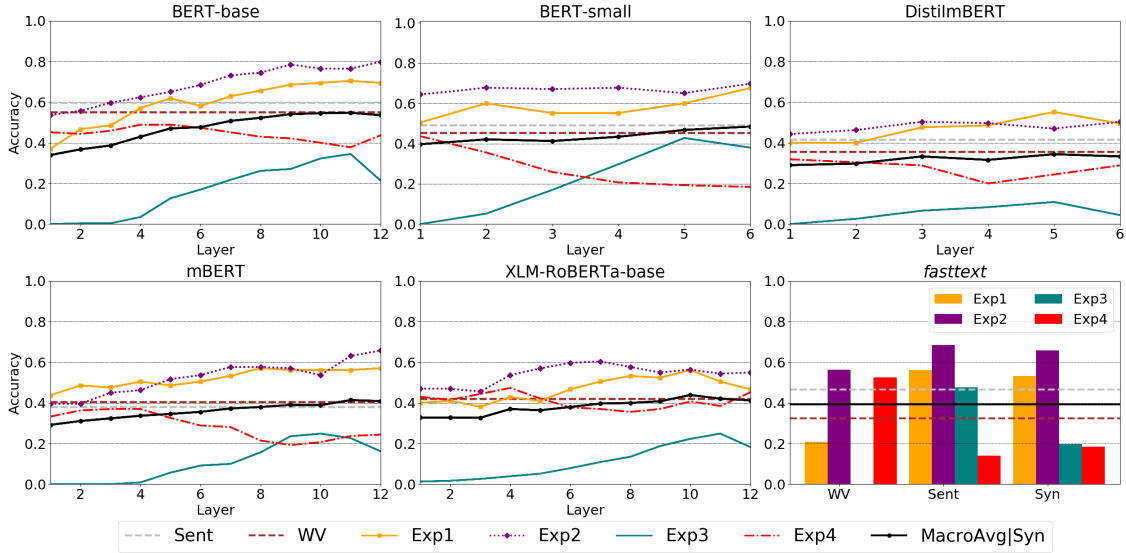
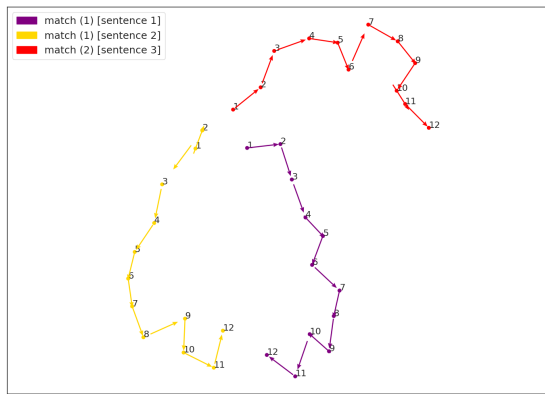


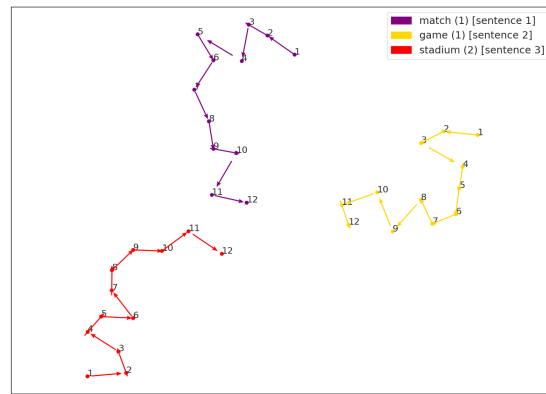
Figure 1: Results across layers and models for Galician. *Sent* and *WV* (dashed) are macro-average values. *MacroAvg|Syn* is the macro-average per layer (Transformers) and the macro-average of the *Syn* strategy (*fastText*).

representations vary across the vector space. Figure 2 shows the UMAP visualizations (McInnes et al., 2018) of the contextualization processes of Exp1

and Exp4 examples in English. In 2a, the similar vectors of *match* in layer 1 are being contextualized across layers, producing a suitable representation



(a) Exp1:
Sentence 2: “Chelsea have a *match* with United next week.”.
Sentence 3: “You should always strike a *match* away from you.”



(b) Exp4:
Sentence 2: “A *game* consists of two halves lasting 45 minutes, meaning it is 90 minutes long.”.
Sentence 3: “He was watching a football *stadium*.”

Figure 2: UMAP visualizations of word contextualization across layers (1 to 12) in Exp1 and Exp4 in English (BERT-base). In both cases, sentence 1 is “He was watching a football *match*.”, and the target word in sentence 3 is the outlier.

since layer 7. However, 2b shows how the model is not able to adequately represent *match* close to its synonym *game*, as the vectors seem to incorporate excessive information (or at least limited lexical knowledge) from the context. Additional visualizations in Galician can be found in Appendix B.

In sum, the experiments performed in this study allow us to observe how different models generate contextual representations. In general, our results confirm previous findings which state that Transformers models increasingly incorporate contextual information across layers. However, we have also found that this process may deteriorate the representation of the individual words, as it may be incorporating excessive contextual information, as suggested by Haber and Poesio (2020).

6 Conclusions and Further Work

This paper has presented a systematic study of word meaning representation in context. Besides static word embeddings, we have assessed the ability of state-of-the-art monolingual and multilingual models based on the Transformers architecture to identify unambiguous cases of homonymy and synonymy. To do so, we have presented a new dataset in four linguistic varieties that allows for controlled evaluations of vector representations.

The results of our study show that, in most cases, the best contextualized models adequately identify homonyms conveying different senses in various contexts. However, as they strongly rely on the surrounding contexts, they misrepresent words having

different senses in similar sentences.

In further work, we plan to extend our dataset with multiword expressions of different degrees of idiomaticity and to include less transparent –but still unambiguous– contexts of homonymy. Finally, we also plan to systematically explore how multilingual models represent homonymy and synonymy in cross-lingual scenarios.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments, and NVIDIA Corporation for the donation of a Titan Xp GPU. This research is funded by a *Ramón y Cajal* grant (RYC2019-028473-I) and by the Galician Government (ERDF 2014-2020: Call ED431G 2019/04).

References

- Rodrigo Agerri, Xavier Gómez Guinovart, German Rigau, and Miguel Anxo Solla Portela. 2018. *Developing new linguistic resources and tools for the Galician language*. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Laura Aina, Kristina Gulordava, and Gemma Boleda. 2019. *Putting words in context: LSTM language models and lexical ambiguity*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3342–3348, Florence, Italy. Association for Computational Linguistics.
- Ju D Apresjan. 1974. Regular polysemy. *Linguistics*, 12(142):5–32.

- Carlos Santos Armendariz, Matthew Purver, Matej Ulčar, Senja Pollak, Nikola Ljubešić, and Mark Granroth-Wilding. 2020. [CoSimLex: A resource for evaluating graded word similarity in context](#). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5878–5886, Marseille, France. European Language Resources Association.
- Marco Baroni. 2013. Composition in distributional semantics. *Language and Linguistics Compass*, 7(10):511–522.
- Marco Baroni and Roberto Zamparelli. 2010. [Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space](#). In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193, Cambridge, MA. Association for Computational Linguistics.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. [Enriching word vectors with subword information](#). *Transactions of the Association for Computational Linguistics*, 5:135–146.
- José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jui-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC at ICLR 2020*.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- David Alan Cruse. 1986. *Lexical semantics*. Cambridge University Press.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Katrin Erk. 2010. [What is word meaning, really? \(and how can distributional models help us describe it?\)](#). In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pages 17–26, Uppsala, Sweden. Association for Computational Linguistics.
- Katrin Erk. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.
- Katrin Erk and Sebastian Padó. 2008. [A structured vector space model for word meaning in context](#). In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 897–906, Honolulu, Hawaii. Association for Computational Linguistics.
- Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppín. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414.
- Marcos Garcia and Pablo Gamallo. 2010. Análise Morfossintáctica para Português Europeu e Galego: Problemas, Soluções e Avaliação. *Linguamática*, 2(2):59–67.
- Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011. [Experimental support for a categorical compositional distributional model of meaning](#). In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Xavier Gómez Guinovart. 2011. [Galnet: WordNet 3.0 do galego](#). *Linguamática*, 3(1):61–67.
- Janosch Haber and Massimo Poesio. 2020. [Assessing polyseme sense similarity through co-predication acceptability and contextualised embedding distance](#). In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 114–124, Barcelona, Spain (Online). Association for Computational Linguistics.
- Patrick Hanks. 2000. [Do Word Meanings Exist?](#) *Computers and the Humanities*, 34:205–215.
- Nathan Hartmann, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jéssica Silva, and Sandra Aluísio. 2017. [Portuguese word embeddings: Evaluating on word analogies and natural language tasks](#). In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 122–131, Uberlândia, Brazil. Sociedade Brasileira de Computação.
- Lotte Hogeweg and Agustin Vicente. 2020. On the nature of the lexicon: The status of rich lexical meanings. *Journal of Linguistics*, 56(4):865–891.
- Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. 2012. [Improving word representations via global context and multiple word prototypes](#). In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882, Jeju Island, Korea. Association for Computational Linguistics.

- Adam Kilgarriff. 1997. I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.
- Adam Kilgarriff, Milos Husák, Katy McAdam, Michael Rundell, and Pavel Rychlý. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX international congress*, pages 425–432. Documenta Universitaria Barcelona, Spain.
- Walter Kintsch. 2001. Predication. *Cognitive science*, 25(2):173–202.
- Ekaterini Klepousniotou, G Bruce Pike, Karsten Steinhauer, and Vincent Gracco. 2012. Not all ambiguous words are created equal: An EEG investigation of homonymy and polysemy. *Brain and language*, 123(1):11–21.
- Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, volume 5, pages 79–86. AAMT.
- Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. *Computational Linguistics and Intellectual Technologies*, 18:333–339.
- Brenden M. Lake and Gregory L. Murphy. 2020. [Word meaning in minds and machines](#). ArXiv preprint: 2008.01766.
- John Lyons. 1995. *Linguistic semantics: An introduction*. Cambridge University Press.
- Lucy J MacGregor, Jennifer Bouwsema, and Ekaterini Klepousniotou. 2015. Sustained meaning activation for polysemous but not homonymous words: Evidence from EEG. *Neuropsychologia*, 68:126–138.
- Scott McDonald and Chris Brew. 2004. [A distributional model of semantic context effects in lexical processing](#). In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 17–24, Barcelona, Spain.
- Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. [UMAP: Uniform Manifold Approximation and Projection](#). *Journal of Open Source Software*, 3(29):861.
- Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. [context2vec: Learning generic context embedding with bidirectional LSTM](#). In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin, Germany. Association for Computational Linguistics.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Workshop Proceedings of the International Conference on Learning Representations (ICLR) 2013*. ArXiv preprint arXiv:1301.3781.
- George A Miller. 1995. [WordNet: a lexical database for English](#). *Communications of the ACM*, 38(11):39–41.
- Jeff Mitchell and Mirella Lapata. 2008. [Vector-based models of semantic composition](#). In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio. Association for Computational Linguistics.
- Ana María Fernández Montraveta and Gloria Vázquez. 2010. La construcción del WordNet 3.0 en español. *La lexicografía en su dimensión teórica*, pages 201–220.
- Sathvik Nair, Mahesh Srinivasan, and Stephan Meylan. 2020. [Contextualized word embeddings encode aspects of human-like word sense knowledge](#). In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 129–141, Online. Association for Computational Linguistics.
- Lluís Padró and Evgeny Stanilovsky. 2012. [FreeLing 3.0: Towards wider multilinguality](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2473–2479, Istanbul, Turkey. European Language Resources Association (ELRA).
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. [Deep contextualized word representations](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.
- Mohammad Taher Pilehvar and Jose 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.
- James Pustejovsky. 1998. *The Generative Lexicon*. The MIT Press.
- Hugh Rabagliati and Jesse Snedeker. 2013. The truth about chickens and bats: Ambiguity avoidance distinguishes types of polysemy. *Psychological science*, 24(7):1354–1360.
- Alexandre Rademaker, Valeria de Paiva, Gerard de Melo, Livy Real, and Maira Gatti. 2014. [OpenWordNet-PT: A project report](#). In *Proceedings of the Seventh Global Wordnet Conference*, pages 383–390, Tartu, Estonia. University of Tartu Press.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text

- transformer. *Journal of Machine Learning Research*, 21(140):1–67.
- Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. [XL-WiC: A multilingual benchmark for evaluating semantic contextualization](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.
- Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. [Visualizing and Measuring the Geometry of BERT](#). In *Advances in Neural Information Processing Systems*, volume 32, pages 8594–8603. Curran Associates, Inc.
- 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.
- Roberto Samartim. 2012. Língua somos: A construção da ideia de língua e da identidade coletiva na galiza (pré-) constitucional. In *Novas achegas ao estudo da cultura galega II: enfoques socio-históricos e lingüístico-literarios*, pages 27–36.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. [DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter](#). In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing at NeurIPS 2019*, Vancouver, Canada.
- Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. [Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing](#). 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 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.
- Hinrich Schütze. 1998. [Automatic word sense discrimination](#). *Computational Linguistics*, 24(1):97–123.
- Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. [BERTimbau: pretrained BERT models for Brazilian Portuguese](#). In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.
- Milan Straka, Jana Straková, and Jan Hajic. 2019. [UD-Pipe at SIGMORPHON 2019: Contextualized embeddings, regularization with morphological categories, corpora merging](#). In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 95–103, Florence, Italy. Association for Computational Linguistics.
- David Tuggy. 1993. Ambiguity, polysemy, and vagueness. *Cognitive linguistics*, 4(3):273–290.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. ArXiv preprint arXiv:1706.03762.
- Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna 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. Association for Computational Linguistics.
- Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. [CCNet: Extracting high quality monolingual datasets from web crawl data](#). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.
- Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers*, pages 161–170, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Lang Yu and Allyson 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. Association for Computational Linguistics.

Appendices

A Complete results

Figure 3 and Table 5 include the results for all languages and models. We also include *large* variants (BERT and XLM-RoBERTa) when available. For static embeddings, we report results for the best *Syn* setting, which combines up to three syntactically related words with the target word (see Appendix D).

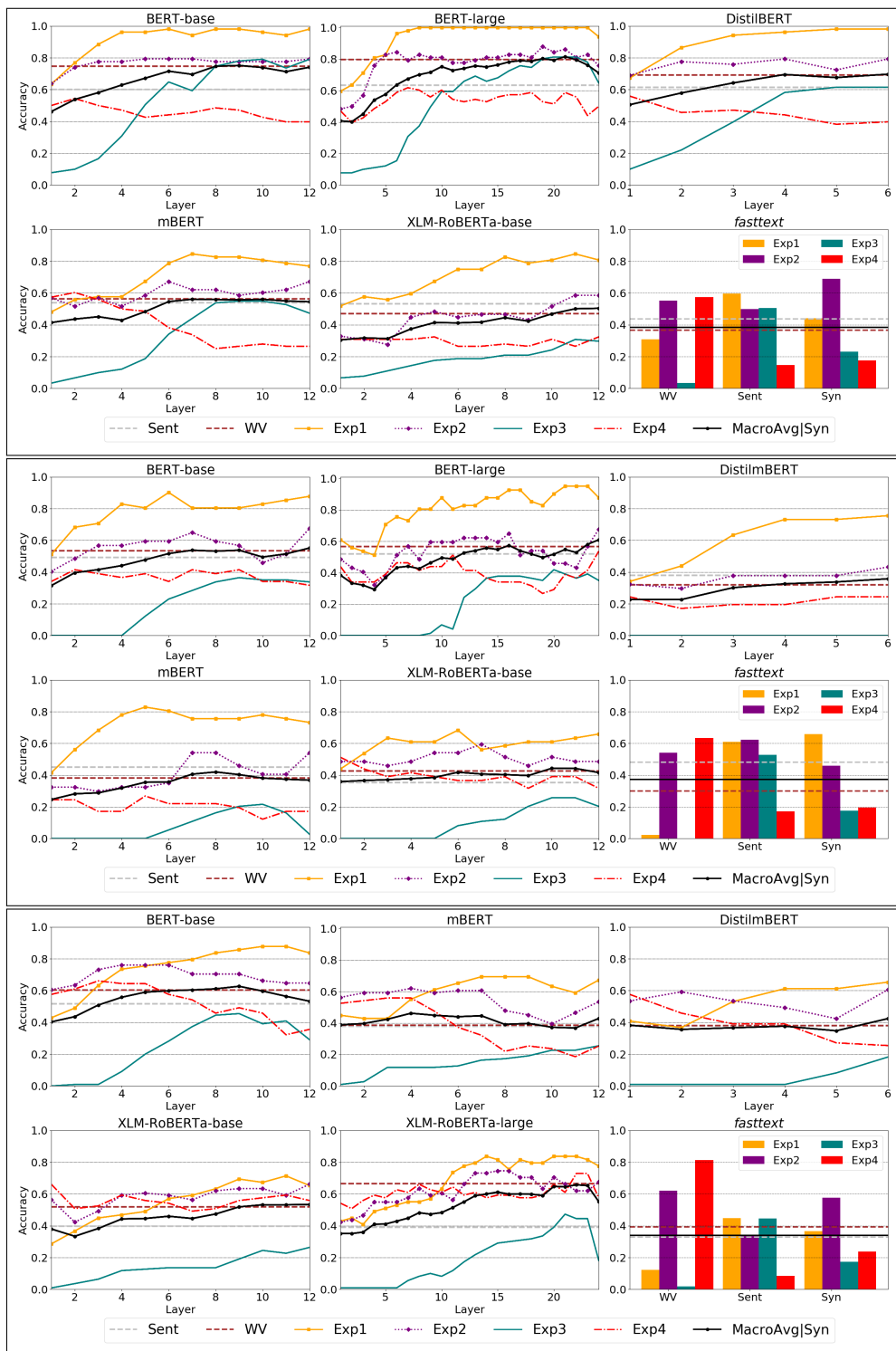
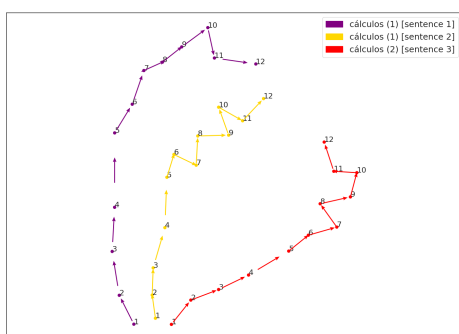


Figure 3: Results across layers and models for English (top), Portuguese (middle), and Spanish (bottom). *Sent* and *WV* (dashed) are macro-average values. *MacroAvg|Syn* is the macro-average per layer (Transformers) and the macro-average of the *Syn* strategy (*fastText*).

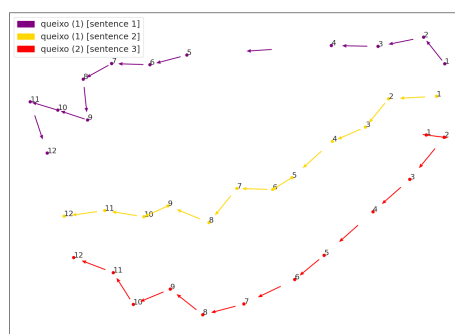
Model	Vec.	Galician			Portuguese			Spanish			English					
		E1	E2	E3	E4	Ma	Mi	F	E1	E2	E3	E4	Ma	Mi	F	
BERT	Sent	0.7	0.76	0.75	0.18	0.6	0.62	0.73	0.56	0.76	0.59	0.54	0.19	0.52	0.52	0.6
	Cat	0.71	0.8	0.29	0.42	0.56	0.51	0.7	0.66	0.86	0.7	0.41	0.44	0.6	0.56	0.74
	Add	0.7	0.8	0.28	0.42	0.55	0.51	0.7	0.67	0.86	0.7	0.41	0.44	0.6	0.56	0.74
BERT ₂	Sent	0.61	0.6	0.59	0.16	0.49	0.5	0.64	0.6	-	-	-	-	-	-	-
	Cat	0.62	0.71	0.3	0.2	0.46	0.43	0.65	0.68	-	-	-	-	-	-	-
	Add	0.61	0.71	0.29	0.2	0.45	0.43	0.65	0.68	-	-	-	-	-	-	-
mBERT	Sent	0.48	0.4	0.49	0.16	0.38	0.39	0.53	0.54	0.51	0.41	0.41	0.19	0.38	0.38	0.5
	Cat	0.57	0.61	0.23	0.22	0.41	0.38	0.62	0.54	0.61	0.45	0.23	0.24	0.38	0.35	0.63
	Add	0.57	0.62	0.21	0.22	0.4	0.37	0.61	0.55	0.63	0.44	0.22	0.25	0.39	0.35	0.63
XLM-b	Sent	0.52	0.51	0.49	0.16	0.42	0.43	0.54	0.45	0.51	0.44	0.46	0.19	0.4	0.41	0.51
	Cat	0.56	0.54	0.22	0.38	0.42	0.39	0.56	0.61	0.67	0.62	0.23	0.54	0.52	0.46	0.69
	Add	0.55	0.54	0.2	0.39	0.42	0.38	0.56	0.61	0.67	0.62	0.23	0.56	0.52	0.47	0.69
XLM-l	Sent	0.42	0.34	0.42	0.16	0.33	0.34	0.44	0.44	0.49	0.48	0.39	0.2	0.39	0.39	0.47
	Cat	0.48	0.5	0.22	0.42	0.4	0.37	0.49	0.58	0.84	0.63	0.46	0.71	0.66	0.62	0.76
	Add	0.46	0.51	0.2	0.43	0.4	0.37	0.5	0.58	0.84	0.66	0.46	0.71	0.67	0.62	0.77
DmBERT	Sent	0.51	0.49	0.5	0.16	0.42	0.43	0.57	0.5	0.51	0.44	0.45	0.12	0.38	0.39	0.51
	Cat	0.52	0.52	0.07	0.24	0.34	0.29	0.51	0.47	0.61	0.49	0.01	0.34	0.36	0.3	0.53
	Add	0.54	0.56	0.07	0.26	0.36	0.31	0.51	0.47	0.61	0.52	0.01	0.37	0.38	0.31	0.54
<i>fastText</i>	Sent	0.56	0.69	0.48	0.14	0.47	0.47	0.62	0.55	0.45	0.34	0.45	0.09	0.33	0.35	0.43
	WV	0.21	0.56	0	0.53	0.33	0.29	0.46	0.45	0.12	0.62	0.02	0.81	0.39	0.35	0.48
	Syn (3)	0.53	0.66	0.2	0.19	0.39	0.36	0.57	0.51	0.37	0.58	0.17	0.24	0.34	0.32	0.55

Table 5: Complete results for the four languages. BERT are BERT-Base models, and BERT₂ refers to a second BERT model for each language (small for Galician, and large for Portuguese and English). XLM-b and XLM-l are XLM-RoBERTa base and large models, respectively. DmBERT is the multilingual version of DistilBERT, and *fastText* the *fastText* embeddings. *Ma* and *Mi* refer to the macro-average and micro-average results across the four experiments, respectively. *F* are the micro-average values on the whole dataset.

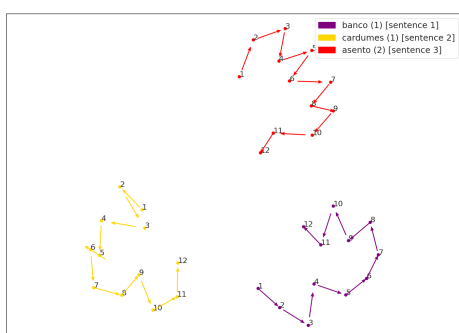
B Contextualization process



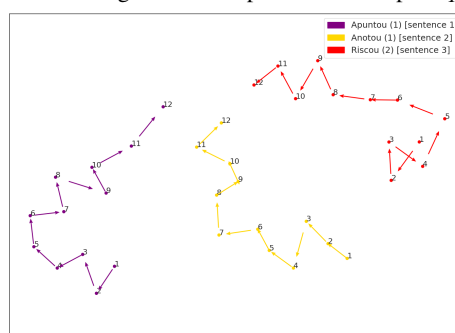
(a) Sent. 1: “Ten que haber algún erro nos *cálculos* porque o resultado non é correcto.”
Sent. 2: “Segundo os meus *cálculos* acabaremos en tres días.”
Sent. 3: “Tivo varios *cálculos* biliare.”



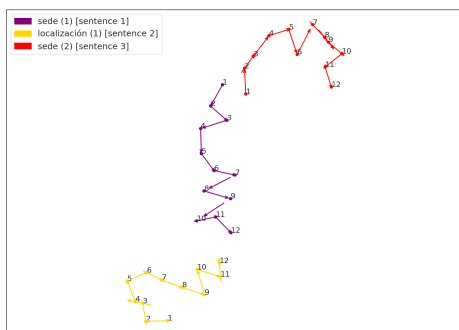
(b) Sent. 1: “De sobremesa tomou *queixo* con marmelo.”
Sentence 2: “Fomos a unhas xornadas gastronómicas do *queixo*.”
Sentence 3: “Achegouse a ela e pasoulle a man polo *queixo*.”



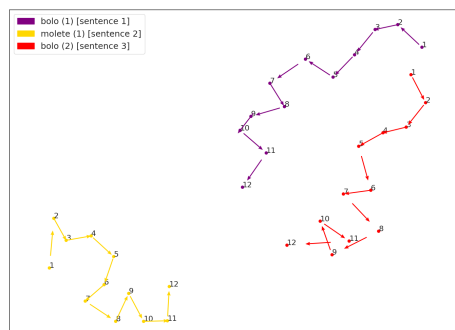
(c) Sentence 1: “Eran tantos que parecían un *banco* de xurelos.”
Sent.2: “Desde a rocha víanse pequenos *cardumes* de robaliza.”
Sentence 3: “Este *asento* de pedra é algo incómodo.”



(d) Sent.1: “*Apuntou* todos os números de teléfono na axenda.”
Sentence 2: “*Anotou* todos os números de teléfono na axenda.”
Sentence 3: “*Riscou* todos os números de teléfono na axenda.”



(e) Sent. 1: “Vai ter lugar a elección da próxima *sede* dos Xogos Olímpicos.”
Sent. 2: “A *localización* do evento será decidida esta semana.”
Sent. 3: “Vou á fonte por auga, que teño *sede*.”



(f) Sentence 1: “Encántalle comer o *bolo* de pan antes da sopa.”
Sentence 2: “O *molete* tiña a codia un pouco dura.”
Sentence 3: “Para atraeren as robalizas iscaban *bolo* vivo.”

Figure 4: Examples in Galician using BERT-base (English translations of the sentences in Appendix E).

First row shows examples of Ex1. In Figure 4a *cálculos* is correctly contextualized since layer 3. In Figure 4b, the outlier sense of *queixo* is not correctly contextualized in any layer.

Second row shows examples of Exp2 (4c) and Exp4 (4d). In Figure 4c, the synonyms *banco* and *cardume* are closer to the outlier *asento* in layer 1 (and from 4 to 7), but the contextualization process is not able to correctly represent the senses in the vector space. In Figure 4d, the result is correct from layer 7 to 11, but in general the representations of words in similar sentences point towards a similar region.

Third row includes examples of Exp3. In Figure 4e, the occurrences of the homonym *sede* are correctly contextualized as the one in the first sentence approaches its synonym *localización* in upper layers. The equivalent example of Figure 4f is not adequately solved by the model, as both senses of *bolo* are notoriously distant from *molete*, synonym of the first homonymous sense.

C Galician models

Training corpus: We combined the SLI GalWeb (Agerri et al., 2018), CC-100 (Wenzek et al., 2020), the Galician Wikipedia (April 2020 dump), and other news corpora crawled from the web. Following Raffel et al. (2020), sentences with a high ratio of punctuation and symbols, and duplicates were removed. The final corpus has 555M words (633M tokens tokenized with FreeLing (Padró and Stanilovsky, 2012; Garcia and Gamallo, 2010)). The corpus was divided into 90%/10% splits for train and development.

fastText model: We trained a *fastText* skip-gram model for 15 iterations with 300 dimensions, window size of 5, negative sampling of 25, and a minimum word frequency of 5. We used the same 90% split used to train the BERT models, but with automatic tokenization (\approx 600M tokens).

BERT models: We used the 90% train split of the corpus (with the original tokenization) to train two BERT models, with 6 and 12 layers:

BERT-small (6 layers): This model has been trained from scratch using a vocabulary of 52,000 (sub-)words and a batch size of 208. It has been training during 1M steps (\approx 20 epochs) in 14 days.

BERT-base (12 layers): Following Kuratov and Arkhipov (2019), we initialized the model from the official pre-trained mBERT, therefore having the same vocabulary size (119,547). We trained it on the Galician corpus during 600k steps (\approx 13 epochs in 28 days) with a batch size of 198.

Both models were trained with the *Transformers* library (Wolf et al., 2020) on a single NVIDIA Titan XP GPU (12GB), a block size of 128, a learning rate of 0.0001, a masked language modeling (MLM) probability of 0.15, and a weight decay of 0.01. They have been trained only with the MLM objective.

D Syntax (*Syn* method)

To get the heads and dependents of each target word we have used the following hierarchies: For nouns: *HeadVerb* (the head verb, if any) > *DepVerb* (dependents of the head verb with one of the following relations: *obj*, *nmod*, *obl*) > *DepAdj* (a dependent adjective) > *DepNoun* (a dependent noun). For verbs: *Head* (only if it is a verb or a noun) > *Obj* (its direct object, if any) > *Arg* (a dependent with one of these relations: *nsubj*, *nmod*, *obl*). Using

these hierarchies we have evaluated representations built by adding from 1 to 4 vectors to the one of each target word. As shown in Table 5, combining 3 syntactically related words to the target one obtains the best results.

For the experiments, we have parsed the datasets using the 2.5 *Universal Dependencies* models provided by UDPipe (Straka et al., 2019).

E English translations (Figure 4)

Figure 4a, sentence 1: “There must be some error in the *calculations* because the result is incorrect”. Sentence 2: “According to my *calculations* we will finish in three days”. Sentence 3: “[He/she] had several *gallstones*”.

Figure 4b, sentence 1: “For dessert [he/she] ate *cheese* with quince”. Sentence 2: “We went to a *cheese* gastronomy days”. Sentence 3: “[He/She] approached her and ran his hand over her *chin*”.

Figure 4c, sentence 1: “They were so many that they looked like a *school* of mackerel”. Sentence 2: “From the rock small *shoals* of sea bass could be seen”. Sentence 3: “This stone *seat* is somewhat uncomfortable”.

Figure 4d, sentences 1 and 2: “[He/She] *wrote down* all the phone numbers on the phone book.” Sentence 3: “[He/She] *crossed out* all the phone numbers on the phone book”.

Figure 4e, sentence 1: “The choice of the next *venue* for the Olympics will take place”. Sentence 2: “The *location* of the event will be decided this week”. Sentence 3: “I’ll get water from the spring, I am *thirsty*”.

Figure 4f, sentence 1: “[He/She] loves to eat the *bread cake* before soup”. Sentence 2: “The *bread* had a slightly hard crust”. Sentence 3: “They used live *sand lance* to attract sea bass”.