Exploration of Sentence Representations in Spanish BERT-like Models

Gonzalo Herrera, Aiala Rosá, Luis Chiruzzo
Instituto de Computación Facultad de Ingeniería
Universidad de la República, Uruguay
{gonzalo.herrera, aialar, luischir}@fing.edu.uy

Abstract

Transformer-based language models, ubiquitous in modern NLP, generate internal representations (embeddings) of words and sentences. Yet, systematic comparisons of embedding strategies from various models remain limited. In this work, we evaluate Spanish embeddings from several BERT-like models (BETO, multilingual BERT, XLM-RoBERTa, ROUBERTa) to understand their syntactic and semantic capabilities across layers. We propose sentence-level analogy tests to probe generalization. Results suggest tasks like verb negation or word reordering perform best with embeddings from earlier layers, while nuanced semantic distinctions—such as agent or patient gender—are better captured by deeper layers. Our findings provide guidelines for embedding strategies and offer a foundation for further NLP research.

Keywords: BERT, embeddings, evaluation, Spanish, Analogies, Similarity

1. Introduction

The task of finding numerical vectors to represent language information has a long history: the concept of word embeddings emerged from the idea of assigning a vector representation to each word. That vector lies in a space whose dimensionality is significantly smaller than the vocabulary size, with the extra constraint that similar words should have close vectors in this space. Collobert and Weston (2008) proposed the approach of pre-training embeddings with unlabeled text. This approach was extended to concatenate additional features to represent extra information, in particular the position of the word in the sentence and its distance to the main verb (Collobert et al., 2011). Later, two new methods were proposed to create word embeddings: word2vec (Mikolov et al., 2013a) and GloVe (Pennington et al., 2014). However, these approaches do not allow a representation to adapt based on the context the word appears in.

The introduction of the transformer architecture Vaswani et al. (2017) enabled the development of contextualized language models. Transformer-based models leverage information across the sentence to create contextual embeddings. This led to the creation of two main families of models: generative models such as GPT, which use decoder-only transformers (Radford et al., 2018), and BERT and BERT-like models, which use encoder-only transformers (Devlin et al., 2019). Encoder-only models produce contextualized vector representations (embeddings) that encode syntactic and semantic information.

Understanding the structure and properties of these embeddings is particularly important in retrieval-based applications such as semantic search and Retrieval-Augmented Generation (RAG) (Lewis et al., 2020). For RAG systems, documents are divided into chunks, which are encoded as vectors and stored in a database. Given a query, its embedding is used to retrieve the most relevant documents, which are passed to a generative model. The effectiveness of the retrieval system depends directly on the quality of the embeddings. Therefore, reliable intrinsic methods to evaluate embedding representations are essential.

Language models are commonly evaluated indirectly, through benchmarks that measure their performance on downstream tasks. While this approach aligns well with how decoder-only transformer models are commonly used, it does not fully reflect the way encoder-only models are employed for retrieval systems. This motivates the need for intrinsic evaluation methods that specifically analyze the semantic and syntactic properties of sentence embeddings. This becomes even more relevant when we consider models for languages other than English, which have been less extensively studied. Furthermore, languages such as Spanish are more morphologically rich than English, introducing additional inflectional variations that embeddings must capture to perform optimally.

In this work, we study how to intrinsically evaluate embeddings produced by BERT-like models for Spanish. We evaluate the models with two sets of tests: a semantic textual similarity test and a new sentence-level analogy task inspired by the word analogies test previously used in word embeddings (Mikolov et al., 2013b). We define a set of controlled syntactic and semantic transformations we can apply to different entities in the sentence

Proceedings of LANLP: Bridging Ibero and Latin American NLP Communities, pages 55–65

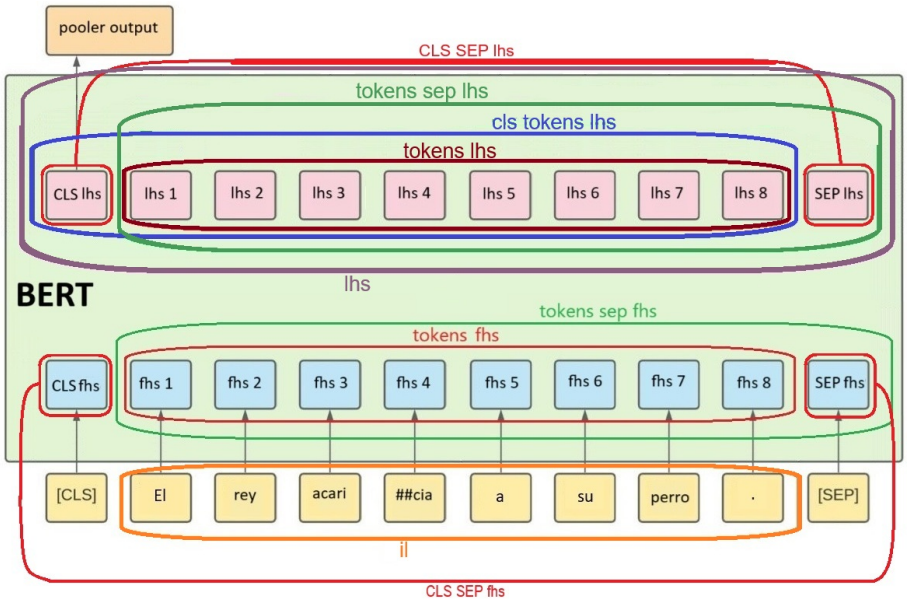


Figure 1: Structure and layers involved in BERT embeddings. The input layer is at the bottom; then the first hidden state (fhs) that consists of the input embeddings + sequence embeddings + positional embeddings; then all the other layers until the last hidden state (lhs); finally there is the pooler output that, in BERT's case is a function that takes the CLS token from the last hidden state. This diagram also shows which transformer activations we considered for defining the different word and sequence embeddings we use in this work.

and evaluate whether such transformations can be transferred to other sentences by operating with the vector representations of the sentences. We call this test the "Analogy Tests using Sentences", and will be made public so other researchers can use it as a benchmark.

This paper is organized as follows. In section 2 we analyze language model evaluation and how it has been done for BERT-like models and LLMs. In section 3 we describe the models we tested and how we extracted the sentence embeddings for the tests. In section 4 we detail all the experiments and tests we did and our proposed evaluation method. In section 5 we show the results we obtained. In section 6 we provide the final conclusions and future work.

2. Related Work

There has been a lot of work when it comes to evaluating language models. Broadly, we can classify the evaluation into two main categories: extrinsic and intrinsic evaluations.

Extrinsic evaluations consist of evaluating the language model based on its performance when used to solve different tasks, also called downstream tasks. BERT and RoBERTa (Liu et al., 2019b) models are usually evaluated with benchmarks like GLUE (Wang et al., 2018), SQuAD (Rajpurkar et al., 2016) and SWAG (Zellers et al., 2018). Other works explored the potential these models have before and after fine-tuning them for specific tasks such as semantic textual similarity and natural language inference (Choi et al., 2021).

Intrinsic evaluations on the other hand try to evaluate the internal representation the model has constructed. Word analogies (Mikolov et al., 2013b) and word similarities (van der Maaten and Hinton, 2008) are two of the main ways that were used to test traditional word embeddings.

After the release of the BERT and BERT-like models, there has been an interest in understanding the internal representation of words captured by transformers. Many works were done in order to understand the information captured by these models. One foundational work looked at how the representation of words is modified for words with multiple meanings, concluding that the embeddings of a word are clustered in separate groups based on context (Reif et al., 2019). Another similar approach is to look at how similar the representation of words remains in sentences and how they change when contextualized (Ethayarajh, 2019). Other works on polling the internal structure of different neural models exist (Liu et al., 2019a), many indicating differences in representation between syntax and semantics that can be found in different layers (Jawahar et al., 2019; Tenney et al., 2019), and the encoding of social biases (Bommasani et al., 2020). Yun et al. (2021) presents a visualization tool based on dictionary learning to better understand the inner works of transformer

Name	Layer / tokens averaged	Tokens for the example sentence
`il`	All word tokens — *input layer*	"El" + "rey" + "acari" + "##cia" + "a" + "su" + "perro" + "."
`tokens fhs`	All word tokens — *first hidden state*	fhs 1 + fhs 2 + ... + fhs 8
`tokens sep fhs`	Word tokens + [SEP]/</s> — *first hidden state*	fhs 1 ... fhs 8 + fhs SEP
`cls lhs`	[CLS]/<s> token — *last hidden state*	lhs CLS
`sep lhs`	[SEP]/</s> token — *last hidden state*	lhs SEP
`tokens lhs`	All word tokens — *last hidden state*	lhs 1 + ... + lhs 8
`cls sep lhs`	[CLS]/<s> + [SEP]/</s> — *last hidden state*	lhs CLS + lhs SEP
`cls tokens lhs`	[CLS]/<s> + word tokens — *last hidden state*	lhs CLS + lhs 1 ... lhs 8
`tokens sep lhs`	Word tokens + [SEP]/</s> — *last hidden state*	lhs 1 ... lhs 8 + lhs SEP
`lhs`	All tokens — *last hidden state*	lhs CLS + lhs 1 ... lhs 8 + lhs SEP
`pooler output`	Pooler output vector	pooler output

Table 1: Embedding representations evaluated. Column 1 lists the shorthand names used throughout the paper; Column 2 specifies which layer and token subset each name refers to; Column 3 illustrates the actual vectors selected for the sentence "El rey acaricia a su perro." (*The king pets his dog.*) as tokenized by BETO-cased.

based language models. Most recently, Bonino et al. (2025) did a comprehensive mathematical analysis of the representations of the attention layers in these models. All of these works, while truly insightful about the inner workings of BERT-like models, do not provide a way to compare them and their vector representations at sentence level. Furthermore, these works are for English, and the exploration of these same representations in Spanish with its morphological particularities has not been explored thoroughly.

3. Models

In this work we tested four variants of the BERT model and five variants of the RoBERTa model. All models were downloaded from HuggingFace.

For evaluating BERT-like models, we chose BETO (Cañete et al., 2020), both cased and uncased variants. BETO is a Spanish specialized model. The other two BERT models we tried are the multilingual BERT models, both cased and uncased.

In the case of RoBERTa, we picked FacebookAI's multilingual model, xlm-RoBERTa (Conneau et al., 2019), both the base model and the large model, two versions of xlm-RoBERTa fine-tuned for Spanish, a base model (Pandya et al., 2021) and a large model (Lange et al., 2021). Lastly, we also tried a RoBERTa model trained specifically over Spanish news text, called ROUBERTa (Filevich et al., 2024). All xlm-RoBERTa models are cased models while ROUBERTa is an uncased model.

The names we used in tables are the following: *beto-cased*, *beto-uncased*, *mbert-cased*, *mbert-uncased*, *xlm-roberta-base*, *xlm-roberta-large*, *xlm-roberta-base-sp*, *xlm-roberta-large-sp* and *rouberta*

3.1. Embedding Extraction

For each model we extract embeddings from different layers and combine different tokens from the layer. We perform different tests using the representation of the input layer of the models, the representation of the first hidden layer, the last hidden layer, and the pooler output vector. When more than one token is involved, we calculate the centroid of the tokens to get the sentence embeddings. Fig. 1 shows each extraction format from a BERT example. Table 1 presents a reference summary with representation name, how it is calculated and an example for each.

4. Description of the Experiments

We performed two different tests in order to evaluate the representations that can be extracted from the models at both the syntactic and semantic levels. The first experiment is a semantic textual similarity test, using the vector representation of the sentences and cosine similarity. The second group is an analogy test using sentences, where we created a small dataset to test the vector representation of different sentences after applying certain transformations.

4.1. Semantic Textual Similarity

The Semantic Textual Similarity (STS) test identifies whether or not two sentences share similar meanings. For this test, we used the Semeval 2015 dataset for Spanish (Agirre et al., 2015), which includes 500 sentence pairs from news articles and 251 pairs from Wikipedia. Each pair is annotated by humans in a range from 0 to 4. To create the gold standard, four annotators scored the similarity of each pair and the label is the average score.

We feed the model both sentences and calculate the cosine similarity for the vectors of both. After that, we calculate the correlation coefficient

Verb	Agents	Patients
acariciar *(to pet)*	hombre, rey, mujer, reina, *(man, king, woman, queen)*	perro, gato, perra, gata *(dog, cat, female dog, female cat)*
comer *(to eat)*	hombre, rey, mujer, reina, perro, gato, perra, gata *(man, king, woman, queen, dog, cat, female dog, female cat)*	pescado, fruta *(fish, fruit)*
llegar *(to arrive)*	hombre, rey, mujer, reina, perro, gato, perra, gata *(man, king, woman, queen, dog, cat, female dog, female cat)*	No patient
dormir *(to sleep)*	hombre, rey, mujer, reina, perro, gato, perra, gata *(man, king, woman, queen, dog, cat, female dog, female cat)*	No patient
pasear *(to walk)*	hombre, rey, mujer, reina *(man, king, woman, queen)*	perro, gato, perra, gata *(dog, cat, female dog, female cat)*
perseguir *(to follow)*	hombre, rey, mujer, reina, perro, gato, perra, gata *(man, king, woman, queen, dog, cat, female dog, female cat)*	hombre, rey, mujer, reina, perro, gato, perra, gata *(man, king, woman, queen, dog, cat, female dog, female cat)*

Table 2: List of agents and patients for each verb. For some verbs, a patient cannot be included given how the sentences were constructed. All agents and patients are also used in their plural form.

Variation	Verbs	Number of experiments
Negation	acariciar, comer, pasear, perseguir *(to pet, to eat, to walk, to follow)*	6000
Gender change of the agent	acariciar, comer, pasear, perseguir *(to pet, to eat, to walk, to follow)*	6000
Number change of the agent	acariciar, comer, pasear, perseguir *(to pet, to eat, to walk, to follow)*	6000
Gender change of the patient	acariciar, pasear, perseguir *(to pet, to walk, to follow)*	4500
Number change of the patient	acariciar, comer, pasear, perseguir *(to pet, to eat, to walk, to follow)*	6000
Temporal modifier reordering	llegar, dormir *(to arrive, to sleep)*	3000

Table 3: List of verbs for each transformation.

between the computed cosine similarity and the human-assigned gold standard scores. In addition, we plot scatter-plots to validate the correlations and get a better understanding of the results.

4.2. Analogy Tests using Sentences

This test is inspired on the word analogies test, but with sentence embeddings. For these tests, we created a controlled synthetic dataset to validate if the models are able to capture different types of morphologic, syntactic and semantic transformations.

4.2.1. Dataset Creation

The dataset was automatically generated using a python script that specified how to combine the different words to construct each sentence. To create it we selected a few verbs that could fit the different transformations we intended to test. For each verb we selected a few nouns to work as agents of the verb, a few to work as the patient of the verb if any applied, and which transformations we would test. For all the verbs with a patient we have sentences in the active voice and in the passive voice. The full list of agents and patients for each verb are shown in Table 2. All of the verbs were conjugated into the past simple, present simple and future simple to simplify the creation and evaluation. Limiting the dataset like this allows us to have a controlled dataset where we are sure every sentence is grammatically correct. The final dataset consists of 6528 sentences.

The transformations we selected are: negation, gender or number change in the agent or patient, and temporal modifier reordering. For all this, we created a collection of sentences and extended them with all the transformations that we could apply, as long as the sentence still made sense. In Table 3 we show which verbs were chosen for each transformation.

Table 4 shows the transformation for the sentence "El rey acaricia a su perro." (*The king pets his dog*) for all the transformations except temporal modifier reordering since this sentence does not have a temporal modifier. For the temporal modifier reordering we took sentences with one of three possible temporal modifiers: "en la mañana" (*in the morning*), "en la tarde" (*in the evening*) and "en la noche" (*at night*). We take the temporal modifier which may go either at the beginning or the end of the sentences, for example: "El hombre duerme en la mañana." (*The man sleeps in the morning*) would be changed into "En la mañana el hombre duerme." (*In the morning the man sleeps*).

4.2.2. Experiments

We made experiments for all the transformations we explained previously, this was done in three different sets:

– all four sentences were in the active voice

Variation	Active voice	Passive voice
Base sentence	El rey acaricia al perro. *(The king pets the dog.)*	El perro es acariciado por el rey. *(The dog is pet by the king.)*
Negation	El rey *no* acaricia al perro. *(The king does not pet the dog.)*	El perro *no* es acariciado por el rey. *(The dog is not pet by the king.)*
Gender change of the agent	*La reina* acaricia al perro. *(The queen does not pet the dog.)*	El perro es acariciado por *la reina*. *(The dog is not pet by the queen.)*
Number change of the agent	*Los reyes* acarician al perro. *(The kings do not pet the dog.)*	El perro es acariciado por *los reyes*. *(The dog is pet by the kings.)*
Gender change of the patient	El rey acaricia a la *perra*. *(The king pets the female dog.)*	*La perra* es acariciada por el rey. *(The female dog is pet by the king.)*
Number change of the patient	El rey acaricia a los *perros*. *(The king pets the dogs.)*	*Los perros* son acariciados por el rey. *(The dogs are pet by the king.)*

Table 4: Syntactic transformations used in our dataset. Starting from a base sentence in both active and passive voice, five transformations are applied: negation, gender/number change of the agent, and gender/number change of the patient. There are also composed transformations, for this example "La reina no acaricia al perro." (*The queen does not pet the dog.*) is the negation and the agent gender change, we also had other sentences such as "Las reinas acarician al perro." (*The queens pet the dog.*) and even "Las reinas no acarician a las perras." (*The queens do not pet the female dogs.*) in the dataset.

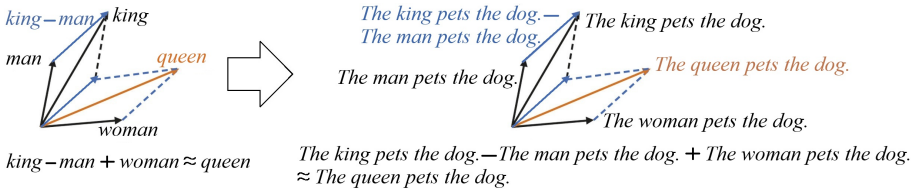


Figure 2: Vector operation. This figure illustrates the $A' - A + B \approx B'$ vector calculation for the traditional method and one example for our proposed method with full sentences. In our method we adapt all words needed for the sentences but no more than that.

- all four sentences were in the passive voice
- two sentences (A and A') were in active voice and the other two (B and B') were in passive voice.

This way we can see if we get different results when we are using the active voice versus the passive voice. We can also see if there is an impact when mixing both passive and active voice in the calculation.

For the experiments, we made sure the same verb was used for both A and B to limit variability. Even though we used the same verb, A and B might have different time conjugations. For each combination of transformation and verb we randomly drew 500 pairs of sentences for the active voice experiments, 500 pairs for the passive voice experiments, and 500 for the mixed voice experiments: 250 where A and A' were in the active voice while B and B' were in the passive voice, and 250 where A and A' were in the passive voice while B and B' were in the active voice. The final number of experiments for each transformation can be seen in Table 3.

4.2.3. Results calculation

We use the embedding of three of the sentences to evaluate whether the vector of the fourth sentence can be recovered. Given two sentences, A and B and their transformed counterparts A' and B' respectively, we compute the vector operation A' - A + B. The resulting vector is then compared against the embeddings of all the sentences in the dataset using cosine similarity. The result is considered correct every time B' is in the top-k most similar sentences. We calculated the results for top-1, top-3 and top-5. However, top-1 accuracy is zero in most of our experiments. Top-3 accuracy generally follows the same trend as top-5 accuracy, but the absolute values remain very low, which makes comparative analysis less informative.

For this reason, we focus our discussion on top-5 accuracy. This allows for a more meaningful comparison across models.

In Fig. 2, we show a transformation of the classic example king/man/woman/queen, and one with full sentences.

5. Results

Due to space limitations, we highlight the best results or those we consider most interesting to analyze.

5.1. Semantic Textual Similarity

As previously mentioned, we present the results of the STS test based on two analyses: a correla-

Model	beto cased	beto uncased	mbert cased	mbert uncased	xlm roberta base	xlm roberta base sp	xlm roberta large	xlm roberta large sp	rouberta
tokens sep fhs	0.499	0.478	0.493	0.493	0.466	**0.466**	**0.434**	**0.444**	0.424
sep lhs	**0.673**	0.527	0.407	0.603	0.028	0.114	0.295	0.135	0.335
tokens lhs	0.581	0.604	**0.616**	0.574	**0.470**	0.238	0.426	0.265	0.607
cls tokens lhs	0.585	0.604	0.615	0.572	0.465	0.235	0.427	0.263	**0.617**
cls sep lhs	0.660	0.528	0.307	**0.627**	0.049	0.109	0.086	0.195	0.429
tokens sep lhs	0.587	**0.604**	0.615	0.575	0.465	0.240	0.425	0.270	0.607

Table 5: Semantic Textual Similarity – Pearson correlation between cosine similarity and the human-ranked gold standard (higher is better). We only show the embedding extraction methods that achieve a maximum in at least one model, omitting those that are outperformed for all the models

tion analysis to get a numeric comparison of the models, and a graphical analysis to get a better understanding of the results.

5.1.1. Correlation analysis

Table 5 shows the Pearson correlation between the gold standard and the cosine similarity for each sentence pair. Most of the best results are found in the last hidden state. This is to be expected since later layers should be able to better capture contextual information as seen in previous studies (Hewitt and Manning, 2019; Jawahar et al., 2019).

BETO outperforms all the other models in the cased variation. It is interesting to see that the option with the best results is the representation of the SEP token in the last hidden state, outperforming the CLS representation for more than ten points. This suggests that relying solely on the CLS token to calculate the pooler output in the original version of BERT might not always be optimal.

Another interesting observation is that the ROUBERTa model has the best performance of all the RoBERTa models for this test, outperforming all of them with a big margin and being near most of the BERT variants.

5.1.2. Graphical analysis

When comparing figures 3a, 3b, and 3c, we can see that while BETO and ROUBERTa's similarity scores range from 0.5 to 1, Multilingual-BERT's scores range from 0.8 to 1. This indicates that although relative similarity increases alongside the gold standard, the absolute similarity values are not a reliable measure. Interestingly, ROUBERTa shows a better spatial representation than Multilingual BERT even though both have almost the same correlation between the similarity and the gold standard. This is probably due to ROUBERTa being specifically trained using news data and the dataset having 500 sentence pairs from newswire.

This reaffirms that these models already have a lot of syntactic and semantic information even without finetuning, as seen in previous studies.

5.2. Analogy Tests

We present a summarization for each transformation and group of experiments: when all the sentences are in active voice, when all the sentences are in passive voice, and when two of them are in passive voice and the other two are in active voice (which we call mixed-voices). We also present the results of specific transformations in more detail.

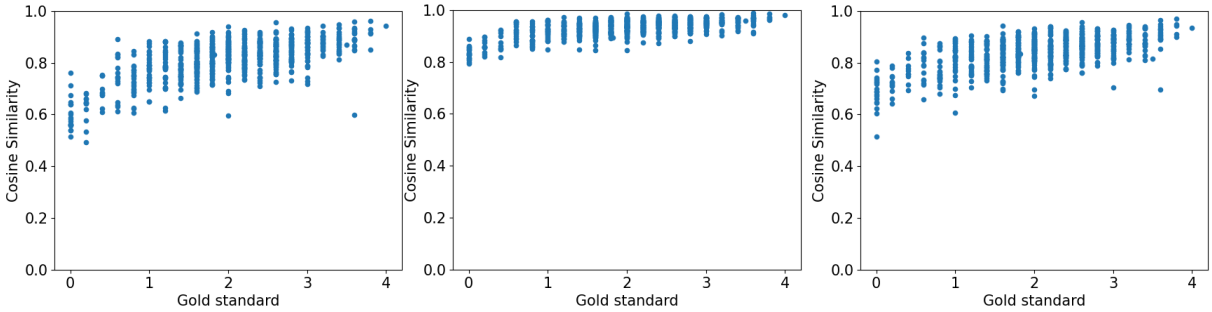
5.2.1. Agent vs Patient

Before analyzing the results, it is important to note that the agent takes the role of the subject of the sentence in the active voice, while it takes the role of the complement in the passive voice. In the same way, the patient shifts from being the complement in the active voice to the subject in the passive voice. This distinction is important because changes in the subject affect the verb to maintain grammatical agreement.

While there are good results for Number of patient in active voice, as seen in the top rows of Table 6, this is not seen in passive voice (middle rows of the table). In the same way, the Number of agent gets somewhat good results (albeit not as good as Number of patient) in passive voice while getting poor results in active voice. In a similar trend we get more balanced results in the active to passive voice test, as seen in the bottom rows. These results indicate that the models are better at changing the number of the complement of the sentence in comparison than changing the number of the subject of the sentence. This makes sense since changing the number of the subject of the sentence also changes the conjugation of the verb, something that does not happen when changing the number of the complement.

The results for gender transformations are better compared to the number transformations. Changing the gender of the subject does not impact the verb in the active voice, and has a smaller impact in the passive voice compared to changing the number of the subject. This also explains why we see better results in the active voice compared to the passive voice.

The result that we find the most interesting is how the models have the best performance in the



(a) **BETO Cased SEP LHS** (b) **mBERT uncased CLS + SEP LHS** (c) **ROUBERTa CLS + tokens LHS**

Figure 3: Each point represents a pair of sentences. In the X axis we have the human-assigned similarity score while in the Y axis we have the cosine similarity between the vectors.

Experiment		beto cased	beto uncased	mbert cased	mbert uncased	xlm roberta base	xlm roberta base sp	xlm roberta large	xlm roberta large sp	rouberta
Active voice	Negation	14.70	47.45	34.90	28.35	**52.80**	47.00	52.35	44.05	45.30
	Gender of patient	11.60	10.00	**47.20**	8.33	2.00	19.13	12.40	18.87	10.47
	Number of patient	42.80	49.65	30.55	37.90	20.65	20.95	16.85	33.95	**61.60**
	Gender of agent	23.60	**39.45**	17.85	7.15	4.90	13.30	11.05	10.80	12.05
	Number of agent	0.15	1.40	0.05	0.05	1.45	2.15	**2.80**	1.05	0.00
Passive voice	Negation	1.65	41.80	14.70	13.60	**53.15**	49.95	53.10	41.70	28.00
	Gender of patient	0.07	5.93	**30.47**	0.53	0.07	11.13	3.73	1.87	0.87
	Number of patient	2.70	**7.15**	3.65	1.65	0.15	2.55	2.25	1.00	0.25
	Gender of agent	31.55	**46.45**	31.75	30.85	4.30	10.60	22.35	25.60	25.70
	Number of agent	42.85	33.95	12.25	26.00	4.35	14.10	26.20	**43.70**	28.05
Mixed active-passive	Negation	8.45	52.25	24.20	22.45	**53.70**	47.45	53.35	41.40	37.70
	Gender of patient	3.47	10.20	**60.27**	4.93	1.07	18.80	8.53	14.00	9.53
	Number of patient	22.30	**34.20**	18.00	21.70	7.80	16.75	8.60	17.45	31.05
	Gender of agent	28.35	**48.20**	26.65	16.10	4.85	15.25	19.50	22.15	15.35
	Number of agent	9.20	18.20	4.80	7.60	2.45	7.20	15.10	**19.70**	6.75

Table 6: Top-5 results for each model for experiments using only active voice, only passive voice, or two in active and two in passive voice. The values show the percentage of times a sentence appeared in the top 5 among its nearest vector neighbors. Results were obtained in different layers depending on the model and task.

mixed-voices tests as seen when comparing results in the three sections of Table 6 for the gender change tests. This goes against what we see in the number change tests. Another surprising result is that the only model that performs well for the gender of patient test is the Multilingual-BERT cased model. Firstly, this model was not fine-tuned for Spanish and we expected it to be one of the least performing models. Secondly, it did not matter if the transformation was in active, passive, or active to passive voices, since, as previously stated, the patient becomes the object or subject of the verb based on which voice we are using. This was also noted for the gender of the agent, where BETO uncased was the model that got the best results regardless of the voice transformation being used. These results prompt further analysis of these kinds of transformations in the future to get a better understanding of the models.

When we analyze these results at the layer level we can see how the information is best represented in the last hidden state (top rows in Table 7). In fact, the best results use SEP token, followed by CLS token. This indicates that this semantic information is better summarized there for all the models we tested, coinciding with the results seen in the STS test.

5.2.2. Negation

The results for the negation of the verb were consistently low for all of the models. This indicates that negation may not be linearly represented in sentence embeddings. However, the best results were seen in the first layers, be it the input layer or the first hidden state as shown in the middle rows of Table 7. This indicates that the changes were captured primarily at word level rather than at the semantic level. It would be interesting to see if newer BERT-like models perform better since new models have been released for the specific purpose of embedding generation in RAG (Wang et al., 2024).

5.2.3. Temporal modifier reordering

The bottom rows of Table 7 show both BETO and Multilingual-BERT got 100% in their uncased vari-

	Representation	beto cased	beto uncased	mbert cased	mbert uncased	xlm roberta base	xlm roberta base sp	xlm roberta large	xlm roberta large sp	rouberta
Gender of Agent task	tokens il	0.12	0.13	1.50	0.02	0.10	0.17	0.10	0.28	8.28
	tokens fhs	0.02	0.08	0.82	0.00	0.02	0.07	0.02	0.13	3.58
	tokens sep fhs	0.02	0.03	0.73	0.00	0.03	0.08	0.03	0.15	3.53
	cls lhs	21.20	41.82	21.90	10.63	4.10	5.92	12.15	17.53	**14.33**
	sep lhs	**26.98**	**44.70**	22.68	16.70	4.05	11.67	17.08	11.08	9.72
	tokens lhs	21.43	38.42	22.47	5.62	3.43	8.43	16.52	16.97	4.83
	cls sep lhs	25.75	43.50	22.18	**16.78**	4.23	**13.05**	**17.20**	15.37	11.98
	cls tokens lhs	22.53	39.45	22.70	5.97	3.75	8.90	16.15	**17.83**	6.82
	tokens sep lhs	22.25	39.47	22.70	6.12	3.80	8.92	16.55	16.95	6.18
	lhs	23.02	40.08	**22.87**	6.32	4.20	9.47	16.38	17.65	7.30
	pooler output	21.72	42.18	18.22	4.50	**4.25**	5.90	11.65	17.47	13.93
Negation task	tokens il	4.32	13.40	**24.45**	18.43	**53.22**	**48.13**	**52.93**	**42.38**	0.00
	tokens fhs	**8.22**	46.17	22.20	**21.47**	22.92	20.35	22.75	21.85	36.68
	tokens sep fhs	7.43	**47.17**	19.65	19.42	21.80	17.80	21.82	20.85	**37.00**
	cls lhs	0.02	0.30	0.28	0.35	7.17	7.75	0.00	1.33	0.38
	sep lhs	0.85	0.38	0.23	1.70	4.78	14.55	0.00	1.78	1.08
	tokens lhs	0.02	0.83	0.97	1.90	26.53	14.25	0.00	0.70	1.03
	cls sep lhs	0.13	0.28	0.18	1.63	5.85	11.68	0.00	1.48	0.58
	cls tokens lhs	0.00	0.67	0.85	1.68	16.10	12.63	0.00	0.80	0.98
	tokens sep lhs	0.07	0.70	0.85	1.82	14.38	13.07	0.00	0.70	1.07
	lhs	0.10	0.60	0.82	1.80	7.65	11.50	0.00	0.78	1.05
	pooler output	0.02	0.22	0.70	0.57	5.32	8.07	0.00	1.58	0.43
Temporal Modifier Reordering task	tokens il	6.43	**100.00**	**8.27**	**100.00**	**85.20**	**85.17**	**76.87**	**62.57**	42.93
	tokens fhs	**6.90**	**100.00**	1.10	**100.00**	57.00	57.03	67.97	50.87	**91.00**
	tokens sep fhs	6.60	**100.00**	1.10	**100.00**	57.30	57.40	68.03	51.47	**91.00**
	cls lhs	0.20	0.43	0.20	0.20	12.40	0.17	20.17	0.07	0.00
	sep lhs	0.67	0.40	0.00	6.80	10.03	0.00	24.23	0.00	0.00
	tokens lhs	0.00	6.57	0.03	0.60	37.63	1.53	32.63	0.00	0.00
	cls sep lhs	0.60	0.43	0.07	5.77	11.23	0.03	23.37	0.00	0.00
	cls tokens lhs	0.03	5.90	0.03	0.67	38.83	1.57	32.77	0.00	0.03
	tokens sep lhs	0.03	6.20	0.07	0.93	39.27	1.10	34.07	0.00	0.03
	lhs	0.07	5.50	0.10	0.93	42.03	1.27	34.03	0.00	0.03
	pooler output	0.03	0.30	0.63	1.90	11.63	0.20	19.30	0.10	0.00

Table 7: Top-5 accuracy for Gender of Agent, Negation, and Temporal Modifier Reordering tasks. Scores of all the experiments: the active-voice, passive-voice, and mixed-voice evaluations. In bold is the best result for each model. The best result across all the variants is underlined.

ants. This is expected since they got those results in the input layer and first hidden state. In these layers, embeddings are the same for both models since the sentence is composed of the same words and word order does not matter, making the task trivial. Interestingly, the capitalization of words was enough to make the cased versions of these models perform poorly. This result was unexpected since the sentences maintained the same meaning and we expected them to have similar representations in higher layers. RoBERTa models on the other hand got pretty good results. These are cased models, with the exception of ROUBERTa, so their much better performance in the task when compared to BETO and Multilingual-BERT is something to note. However, as already seen for the negation tests, the best results were obtained in the input layer and first hidden state, with performance deteriorating in the last hidden state and output. Interestingly, both Multilingual-RoBERTa models got good results in the last layers. As they are not fine-tuned for Spanish, we did not expect them to be the only models to be able to perform well.

An interesting analysis would be to test all other hidden states to determine whether performance deteriorates gradually across layers or drops suddenly after a specific one. Another possible analysis would be to test whether we can change one temporal modifier to another — for example changing “in the morning” to “at night” or adding a temporal modifier to a sentence without one.

6. Conclusions and Future Work

We proposed a new intrinsic approach to test and compare BERT-like language models with respect to their capabilities to retain syntactic and semantic information at the sentence level. The tests yield interesting results that enable us to compare different models in terms of their ability to preserve specific syntactic and semantic information across sentences.

Our experiments show that different types of transformations are not equally captured. While gender and number changes were partially captured, the negation transformation was not well represented in later layers.

However, this framework is only an initial proposal and needs to be further developed to provide

a more robust comparison of models. Nonetheless, we expect this approach to raise awareness of the need for improved methods to evaluate transformer-based language models, thereby enabling more accurate comparisons in their current applications.

Future work includes expanding the range of semantic transformations, such as verb tense variants, adjective modifications, synonym-antonym substitutions, and active to passive voice. We also would like to explore sentence embedding models and how those compare to the results we found. Additionally, we plan to extend the dataset with more naturalistic sentences constructed from real-world examples.

7. Bibliographical References

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

Matteo Bonino, Giorgia Ghione, and Giansalvo Cirrincione. 2025. The geometry of bert.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Hyunjin Choi, Judong Kim, Seongho Joe, and Youngjune Gwon. 2021. Evaluation of bert and albert sentence embedding performance on downstream nlp tasks. In *2020 25th International conference on pattern recognition (ICPR)*, pages 5482–5487. IEEE.

R. Collobert and J. Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *International Conference on Machine Learning, ICML*.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

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.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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 55–65, Hong Kong, China. Association for Computational Linguistics.

Juan Pablo Filevich, Gonzalo Marco, Santiago Castro, Luis Chiruzzo, and Aiala Rosá. 2024. A language model trained on uruguayan Spanish news text. In *Proceedings of the Second International Workshop Towards Digital Language Equality (TDLE): Focusing on Sustainability @ LREC-COLING 2024*, pages 53–60, Torino, Italia. ELRA and ICCL.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Lukas Lange, Heike Adel, and Jannik Strötgen. 2021. Boosting transformers for job expression extraction and classification in a low-resource setting. In *Proceedings of The Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Advances in Neural Information Processing Systems, 33:9459–9474.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual 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 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. 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. Association for Computational Linguistics.

Hariom A. Pandya, Bhavik Ardeshna, and Dr. Brijesh S. Bhatt. 2021. Cascading adaptors to leverage english data to improve performance of question answering for low-resource languages.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. *Advances in Neural Information Processing Systems*, 32.

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

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024. Text embeddings by weakly-supervised contrastive pre-training.

Zeyu Yun, Yubei Chen, Bruno Olshausen, and Yann LeCun. 2021. Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 1–10, Online. Association for Computational Linguistics.

8. Language Resource References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

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