Procesamiento del Lenguaje Natural, Revista nº 74, marzo de 2025, pp. 305-319 recibido 05-12-2024 revisado 23-01-2025 aceptado 11-02-2025

Assessing lexical ambiguity resolution in language models with new WiC datasets in Galician and Spanish

Evaluación de la resolución de la ambigüedad léxica en modelos de lengua con nuevos conjuntos de datos WiC en gallego y español

Marta Vázquez Abuín, Marcos Garcia
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela
{martavazquez.abuin, marcos.garcia.gonzalez}@usc.gal

Abstract: Ambiguity resolution, particularly in addressing lexical phenomena such as polysemy, has been a long-standing challenge in NLP. From a computational point of view, this problem has traditionally been tackled through tasks such as word sense disambiguation and, more recently, with the appearance of Word-in-Context (WiC) datasets, which tackle polysemy resolution as a binary classification problem. These datasets play a crucial role in evaluating the lexical capabilities of vector models, but their availability is limited to only a few languages, creating a significant disadvantage for varieties lacking such resources. This paper introduces WiC datasets for Galician and Spanish, addressing the gap in the research on lexical ambiguity resolution for these languages. The datasets have a total of 4,300 instances, and their creation has followed the guidelines of the original English WiC. Besides introducing the datasets, we present a systematic evaluation of monolingual and multilingual transformer models across layers, exploring aspects such as data overlap, rogue dimensions, and cross-lingual transfer. The results reveal that (i) monolingual and multilingual models have comparable accuracy, (ii) vector normalization has little effect on the models' performance, and (iii) cross-lingual transfer between Galician and Spanish is not effective. Among the evaluated models, Llama 3.2 seems to be the most effective at solving the task.
Keywords: word-in-context, lexical ambiguity, language models.

Resumen: La resolución de la ambigüedad, sobre todo al abordar fenómenos léxicos como la polisemia, ha sido un reto de gran tradición en el PLN. Desde un punto de vista computacional, este problema ha sido tradicionalmente abordado mediante tareas de desambiguación del sentido de las palabras y, más recientemente, con la aparición de los conjunto de datos WiC, que abordan la resolución de la polisemia como un problema de clasificación binaria. Estos recursos desempeñan un papel crucial en la evaluación de las capacidades léxicas de los modelos vectoriales, pero su disponibilidad se limita a unas pocas lenguas, lo que supone una desventaja significativa para las lenguas que carecen de tales recursos. Este trabajo presenta datasets en formato WiC para gallego y español, abordando el vacío existente en la investigación de la resolución de la ambigüedad léxica para estas lenguas. Los datasets están formados por un total de 4.300 instancias, y su creación ha seguido las directrices del WiC original en inglés. Además de presentar los conjuntos de datos, presentamos una evaluación sistemática de los modelos transformer monolingües y multilingües entre capas, explorando aspectos como el solapamiento de datos, las dimensiones anómalas y la transferencia entre lenguas. Los resultados revelan que (i) los modelos monolingües y multilingües tienen una precisión comparable, (ii) la normalización vectorial tiene poco efecto en el rendimiento de los modelos, y (iii) la transferencia inter-lingüística entre el gallego y el español no es efectiva. En relación a los modelos evaluados, Llama 3.2 parece el más efectivo resolviendo la tarea.
Palabras clave: word-in-context, ambigüedad léxica, modelos de lenguaje.

ISSN 1135-5948 DOI 10.26342/2025-74-21

1 Introduction

One of the defining characteristics of natural languages is their ambiguity, particularly lexical ambiguity (Cruse, 1986). The recognition and resolution of lexical ambiguity is an intrinsic aspect of human communication, where speakers continually seek to navigate and mitigate any potential misinterpretations that may arise from this phenomenon (Tuggy, 1993). From a computational perspective, it presents a significant challenge for language models attempting to emulate human language understanding (Aina, Gulordava, and Boleda, 2019; Liu et al., 2023).

Among the different types of lexical ambiguity, polysemy (and homonymy) are worthy of particular mention. We can define them as a word form having multiple meanings depending on their context (Apresjan, 1974). To illustrate, the word *arma* (literally 'weapon') in Galician can have (at least) two distinct meanings: (i) *large but transportable armament*, and (ii) *a means of persuading or arguing* (Guinovart, 2011). Effective language models must be able to differentiate between these two meanings in context.

Lexical ambiguity resolution is a well-established problem with a long tradition in NLP, addressed primarily through tasks like Word Sense Disambiguation (WSD) (Agirre and Edmonds, 2007), where a model links a word-in-context to a synset in a lexical database, predominantly WordNet (Fellbaum, 1998). However, and despite the recent advances in language modeling, lexical ambiguity continues to present a significant challenge for linguistic technologies, particularly in the context of automatic and unsupervised approaches, and it is essential for different tasks such as machine translation, language understanding, or generation (Navigli, 2009; Loureiro et al., 2021; Liu et al., 2023; Ortega et al., 2024).

In recent times, tasks such as WiC (Pilehvar and Camacho-Collados, 2019) have approached lexical disambiguation as a binary classification problem, whereby a given form is observed to occur in two distinct contexts, in which it may or may not have the same meaning (see Table 1). This dataset has become a reference for the evaluation of lexical capabilities of vector space models, as evidenced by its inclusion in benchmarks such as SuperGLUE (Wang et al., 2019), and by its adaptation to other languages (Raganato et al., 2020; Bouma, 2024).

F	*Justify* the margins	The end *justifies* the means
T	*Air* pollution	Open a window and let in some *air*

Table 1: Example of a false (F) and a True (T) pairs from the original WiC (Pilehvar and Camacho-Collados, 2019) (target word in *italics*).

However, creating WiC-like datasets demands significant time and resource investment for compilation, organization, and revision. As a result, not only do under-resourced languages —like Galician— lack such datasets, but even languages with greater research attention —such as Spanish— still do not have WiC datasets available for evaluating language models in this task.

Taking the above into account, this paper introduces two novel WiC-like datasets for Galician and Spanish, developed in adherence to the original format, and utilizing Galnet (Guinovart, 2011) and SemCor (Miller et al., 1993) data. The Galician dataset is comprised of 3,300 instances, while the Spanish dataset contains 1,000 entries.

These datasets are used to systematically evaluate performance of monolingual and multilingual transformer models across their hidden layers, encompassing both encoder and decoder architectures, in Galician and Spanish. Furthermore, we explore the impact of seeing words during training (i.e., whether the model better solves instances in which the target word has been seen in training), the influence of rogue dimensions in the vector space (Timkey and van Schijndel, 2021), and cross-lingual transfer by means of using Galician data to train Spanish models.

As it will be showed in the results, the size of the models is not a determining factor in lexical disambiguation capabilities, and the cross-lingual transfer negatively affects performance. Regarding rogue dimensions, normalizing the vector space seems to have little to no noticeable impact on performance.

In sum, this paper contributes with two new datasets for Galician (GL) and Spanish (ES), which will be published under free licenses.[1] Additionally, it presents a systematic analysis of state-of-the-art models

[1] https://github.com/mrtva/wic-pln-2025

for these languages, establishing an initial benchmark for future evaluations of novel methods aimed at resolving lexical ambiguity in Galician and Spanish.

Section 2 reviews previous studies of lexical ambiguity and the creation of WiC datasets. Section 3 describes the process of creating Galician and Spanish datasets. Then, Section 4 details the experiments conducted and discusses the results. Finally, Section 5 concludes this study and outlines potential directions for future research.

2 Related work

Princeton WordNet (Fellbaum, 1998) serves as the primary repository for word senses in NLP, and is the foundation for most datasets and evaluation frameworks used for WSD, alongside BabelNet (Navigli and Ponzetto, 2012) and Wiktionary (Bevilacqua et al., 2021; Raganato, Camacho-Collados, and Navigli, 2017). This database of semantic relations groups words into synsets (synonym sets) and links them to other words according to their shared meaning, so a word can belong to multiple synsets, each representing a different meaning. The structured format of WordNet makes it particularly well-suited for the creation of WSD and WiC datasets, as its provides a well-defined inventory of word senses and semantic relationships.

Concerning Galician, resources including this language are scarce (e.g., the small subset of the XL-WSD dataset (Raganato et al., 2020)), or are focused on different semantic phenomena, such as the resources presented by Garcia (2021), designed for evaluating homonymy and synonymy.

Regarding the performance of language models in lexical ambiguity, contextualized models are effective in identifying homonyms and distinguishing between meanings (Apidianaki, 2022). However, when the contexts are similar, the accuracy of the results decreases, due to the fine-grained distinctions (Loureiro et al., 2020; Garcia, 2021).

For Spanish, recent research has explored the performance of pre-trained language models on lexical semantic tasks (Garí Soler and Apidianaki, 2021; Garcia, 2021; Rivière, Beatty-Martínez, and Trott, 2024). However, to the best of our knowledge, there is no WiC-like dataset for this language.

2.1 Word-in-Context datasets

The original English dataset was developed to evaluate context-sensitive word representations (Pilehvar and Camacho-Collados, 2019). Specifically, the aim was to assess how models interpret the diverse meanings of a word, without relying on predefined inventories (Navigli and Ponzetto, 2012). In this binary classification task, a target word w (either a verb or a noun) is presented in two different contexts (*c1* and *c2*), which have been assigned specific meanings for the target word w. The objective is to assess the capacity of vector models to ascertain whether both word's meanings are the same or not. To construct the original WiC dataset, WordNet, Wiktionary and VerbNet (Kipper Schuler, Korhonen, and Brown, 2009) were used. In the case of XL-WiC (Raganato et al., 2020), by incorporating different language-specific WordNets.

Following the original version, other Word-in-Context datasets were created: (i) the multilingual XL-WiC (Raganato et al., 2020); (ii) the cross-lingual MCL-WiC (Martelli et al., 2021) or (iii) the Mirror-WIC (Liu et al., 2021a). In addition, other monolingual datasets such as Swe-WiC (Bouma, 2024), WiC-ITA (Cassotti et al., 2023) or the Danish Word-in-Context (Pedersen et al., 2024) dataset were also released. Furthermore, other tasks were designed based on the same format, such as AM2ICO (Liu et al., 2021b), a binary classification between English and other 14 languages where each instance has the target word and a context in English and the target languages or the WiC-TSV (Breit et al., 2021), created to evaluate models in target sense verification tasks. Galician and Spanish have not been part of these datasets, due to insufficient data for their creation. Specially for Galician, the number of resources and datasets created for lexical disambiguation tasks and model evaluation is scarce (Garcia, 2021). Therefore, the creation of these datasets addresses the limitation of available resources, and provides comprehensive datasets for the assessment of lexical ambiguity in these languages. Consequently, it contributes to the assessment and potential improvement of vector models for both Galician and Spanish.

3 Building the Galician and Spanish WiC datasets

Here we outline the methodology employed for the retrieval of sentences and the construction of both datasets.

3.1 Galician WiC

We use the Galician WordNet, Galnet (Guinovart, 2011), as a basis for the Galician WiC. However, as its size is relatively limited —with only approximately 4,500 entries—, it was essential to develop additional methodologies to collect the necessary sentences for the dataset creation.

3.1.1 Data construction

The use of WordNet for this dataset is valuable because of the complete range of precise meanings. To illustrate, the Galician word *mancha* has ten distinct meanings, each one associated with a different synset. These meanings encompass a number of related concepts, including the following: (i) *a small contrasting part of something*, (ii) *a blemish made by dirt*, (iii) *an irregularly shaped spot*, and (iv) *a mark or flaw that spoils the appearance of something.*[2]

Sentence extension and extraction: The sentences that comprise this dataset were obtained via a multi-step process. In the initial stage of the process, all available Galician examples were extracted from Galnet, amounting to approximately 4,000. However, this number was insufficient for the creation process, as most examples were isolated instances of the target word. We opted to devise a methodology to augment the number of sentences, we translated all the English (around 20,000) and Spanish (1,200) examples that belong to a synset with a Galician word using a state-of-the-art neural machine translator (Gamallo et al., 2023). Afterwards, all the sentences were lemmatized to verify that the translated sentences contained the Galician word from the respective synset. Additionally, we expanded the resource by also translating WordNet English sentences whose target word did not exist in Galician with the same synset, resulting in approximately 20,000 examples. In these cases, we utilized bilingual word embeddings as probabilistic dictionaries to identify potential new Galician words. We included in the final dataset those cases in which a word translation provided by the bilingual word embeddings also appear in the sentence translation, resulting in around 13,000 sentences from English and 600 from Spanish. A random subset of 1,600 coincidences was reviewed by native speakers with background in Linguistics, resulting in a 98% success rate, ensuring the high quality of the translations at both word and sentence level. A detailed explanation of this process can be found at (Vázquez Abuín and Garcia, 2024).

3.1.2 Filtering

We obtained around 17,500 sentences before applying a filtering process. We compiled the three sets (train, dev and test) following the restrictions outlined in the original WiC (Pilehvar and Camacho-Collados, 2019) with minor modifications: (i) not having more than three instances for the same target word in each set, (ii) not having the same sentence more than once, (iii) for the negative instances, not belonging to the same supersense[3], and (iv) to not have less than a 0.3 Levenshtein distance between the two contexts, trying to avoid similar sentences. Additionally, we ensured that the number of true and false instances was balanced across all the sets. Test data was manually reviewed to ensure the quality of the instances.

3.2 Spanish WiC

The Spanish dataset comprises 1,000 instances, with 200 instances designated for development and 800 for training. Its creation relies on two sources: The extraction of Spanish examples from MCR 3.0 (Gonzalez-Agirre, Laparra, and Rigau, 2012), and the translated SemCor utilized in XL-WSD (Pasini, Raganato, and Navigli, 2021).

The MCR 3.0 (Gonzalez-Agirre, Laparra, and Rigau, 2012) for Spanish contains approximately 1,200 sentence examples, but these are limited to adjectives, which are not enough for the dataset creation. To address this limitation, we decided to use the sense inventory of SemCor translated to Spanish used for the XL-WSD (Pasini, Raganato, and Navigli, 2021). From this translated SemCor, we extracted the lexical forms associated with more than three synsets, resulting in over 17,000 sentences. The pruning

[2] https://ilg.usc.gal/galnet/galnet.php?version=dev\&lingua=GL\&variante=mancha\&compara=comeza\&lg=gl

[3] https://wordnet.princeton.edu/documentation/lexnames5wn

Target Word	Label	Sentence 1	Sentence 2
papel	F	a idea de que unha oficina funcione sen *papel* é absurda	cal é o seu *papel* no equipo?
aceptable	T	as actuacións varían entre *aceptables* a excelentes	niveis de radiación *aceptables*

Table 2: Examples of the Galician WiC. Each row is an instance containing a target word appearing in two sentences (in italics). Each instance is labeled as True if the word conveys the same meaning in both sentences, and False if it conveys a different meaning. Column indicating the POS-tag of the target word is not shown. Translations can be found in Table 8 (Appendix A).

and filtering of the dataset followed the same methodology used for the Galician dataset. The constraints were as follows: (i) no more than three instances of the same word, (ii) each sentence has to be limited to a single occurrence, and (iii) for negative splits, those that are part of the same supersense should be excluded. As was done for Galician, the 1,000 instances were manually reviewed to ensure their quality.

3.3 Statistics

Table 3 shows the statistics of Galician and Spanish sets with the information of the number of instances, unique words, and part of speech. The size of available datasets using WordNet is between 7,500 for English and 400 for Estonian (Raganato et al., 2020).

Splits	Inst.	Uniq.	V	N	R	A
Train	1500	1187	271	454	7	768
Dev	400	278	71	113	7	209
Test	1400	905	274	536	13	577
Dev	200	190	46	104	2	48
Test	800	641	168	451	4	177

Table 3: Size of Galician (top) and Spanish (bottom) splits. *Inst.* refers to the number of instances, and *Uniq.* to the unique words. The information about POS is also included: V refers to the number of verbs, N to nouns, R to adverbs and A to adjectives.

4 Evaluation

This section outlines the experimental setup, details the methodologies employed, and presents a discussion of the results.

4.1 Experiments

Inspired by the original WiC paper, we conduct a series of experiments to explore the off-the-shelf capabilities of current models, and to serve as an initial benchmark evaluation which may serve as a reference point for future improvements.

Models: For Galician, we evaluate four monolingual models: Both 'small' and 'base' variants of Bertinho (Vilares, Garcia, and Gómez-Rodríguez, 2021) and of BERT (Garcia, 2021). For Spanish, we compared BETO (Cañete et al., 2020), Bertin-RoBERTa-base (De la Rosa et al., 2022), and both variants ('base' and 'large') of RoBERTa-BNE (Fandiño et al., 2022). Additionally, we also included the following multilingual models for the evaluation in both languages: mBERT (Devlin et al., 2019) and XLM-RoBERTa ('base' and 'large') (Conneau et al., 2020) as encoders, and Llama 3.2 (1B) as auto-regressive.[4]

Method: Following Wang et al. (2019), we trained simple logistic regression classifiers which take as input the concatenated vectors of each target word in a pair, i.e., the word contextualized in both sentences (see examples in Table 2). To explore the impact of the contextualization process across layers, we trained a classifier for each layer of each model. For Galician, we only used the 'train' split, while for Spanish we compared the performance of models trained with the 'dev' set, with the Galician training data, and with both datasets. Word vectors were extracted using the `minicons` library (Misra, 2022)[5], which takes advantage of the models provided by HuggingFace's Transformers.[6]

Baselines: We compared the results of the LMs to two baselines based on cosine similarity threshold: For transformer models, a simple threshold-based classifier that uses the cosine distance between the contextualized vectors of the target word in both sentences, with steps of 0.02. Second, and for static vectors the same threshold approach using

[4] `https://huggingface.co/meta-llama/Llama-3.2-1B`

[5] `https://github.com/kanishkamisra/minicons`

[6] `https://github.com/huggingface/transformers`

sentence-level embeddings obtained by averaging the word vectors.[7] In both cases, and at each 0.02 step, we classified an instance as True if the cosine similarity was higher than the threshold, and False otherwise.

It is worth noting that transformer models tend to produce anisotropic spaces (Godey, Clergerie, and Sagot, 2024; Machina and Mercer, 2024), and that a few rogue dimensions seem to dominate similarity measures (Timkey and van Schijndel, 2021). Therefore, we compare —both in the logistic classifiers and in the threshold-based baselines— the performance of the original vectors extracted from the models (dubbed COS the cosine similarity thresholds, and EMB the logistic classifiers trained with embeddings) to normalized vectors using the z-score standardization (NORM) as proposed by Timkey and van Schijndel (2021).

4.2 Results

We present the results of the experiments conducted to evaluate the performance of the language models on the lexical ambiguity task for the original embeddings (the detailed results comparing with the normalized embeddings are provided in Appendix A).

4.2.1 Galician results

The results of the Baseline experiments are summarized in Table 4. The Bertinho-small, Bertinho-base, and XLM-large models exhibited the highest performance among the tested models, achieving an accuracy of 0.66 at medium thresholds, surpassing the original WiC benchmark (0.638).

In terms of model performance, Table 5 shows the accuracy of the logistics classifiers using both the original (EMB) and the normalized (NORM) embeddings, respectively. The results demonstrate that models with a larger number of hyperparameters, such as Llama 3.2 1B, exhibit enhanced performance in both configurations. Nevertheless, smaller models, such as BERT-small or Bertinho-small, demonstrate comparable accuracy, exhibiting a performance very similar to that of the larger and more complex models. Concerning the normalized vectors, there is also no notable difference in the results.

Figure 1 illustrates the performance of small models across layers, demonstrating that the highest accuracy is consistently achieved between layers 3 and 5, indicating that these smaller models generate their most effective embeddings at intermediate layers. For the 12-layer models, Figure 2 illustrates that the performance remains relatively stable from layer 4 onward, with no noticeable drop in performance in the final layers. Additionally, for larger models, Figure 3 shows that the peak accuracy is reached in the middle layers, with Llama 3.2 1B achieving its highest performance at layer 5. However, there is a gradual decline in performance in the upper layers.

In sum, the results suggest that model size does not significantly affect performance and smaller models demonstrate accuracy comparable to larger models. Regarding the layers, regardless of the size of the models, the intermediate layers provide the most effective accuracy.

Model	Acc.	Thr.	Layer
Bertinho-small	**0.66**	0.44	6 (C)
Bertinho-base	**0.66**	0.52	9 (C)
BERT-small	0.64	0.64	6 (C)
BERT-base	0.63	0.56	9 (N)
mBERT	0.63	0.78	8 (C)
XLM-base	0.61	0.50	9 (N)
XLM-large	**0.66**	0.46	18 (N)
Llama 3.2 1B	0.60	0.74	7 (C)
FastText	0.54	0.70	– (N)

Table 4: Baseline results for Galician. *Thr.* refers to the cosine similarity threshold, and *Acc.* to the accuracy. *Layer* includes, besides the transformer layer, whether it uses the Cosine from the original vector or the Normalized one.

Model	EMB		NORM	
	Acc.	L.	Acc.	L.
Bertinho-small	0.77	4	0.77	4
Bertinho-base	0.78	8-9	0.78	8-9
BERT-small	0.78	4-5	0.77	3-5
BERT-base	0.78	9	0.79	10
mBERT	0.78	5	0.79	5
XLM-base	0.80	12	0.79	12
XLM-large	0.79	12	0.78	11
Llama 3.2 1B	**0.81**	5	**0.81**	5

Table 5: Summary of the performance comparison using original *EMB* and normalized *NORM* embeddings. *Acc.* refers to the accuracy and *L* to the layer. For cases with more than one layer with the same accuracy, both layers are represented with a hyphen.

[7] For both Galician and Spanish, we used the official *CC* models provided by `fastText` (Grave et al., 2018).

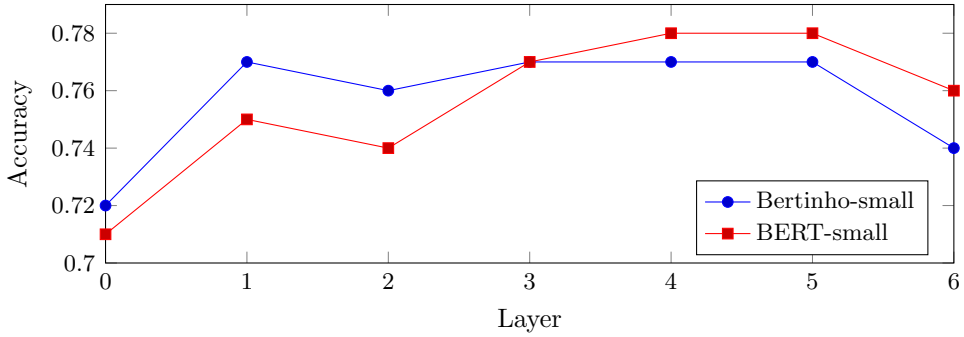

Figure 1: Accuracy across layers for small models in Galician using original embeddings.

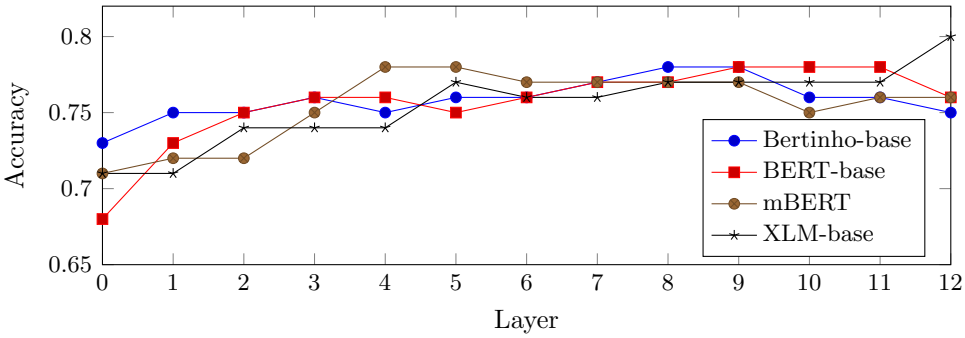

Figure 2: Accuracy across 12-layer models in Galician using original embeddings.

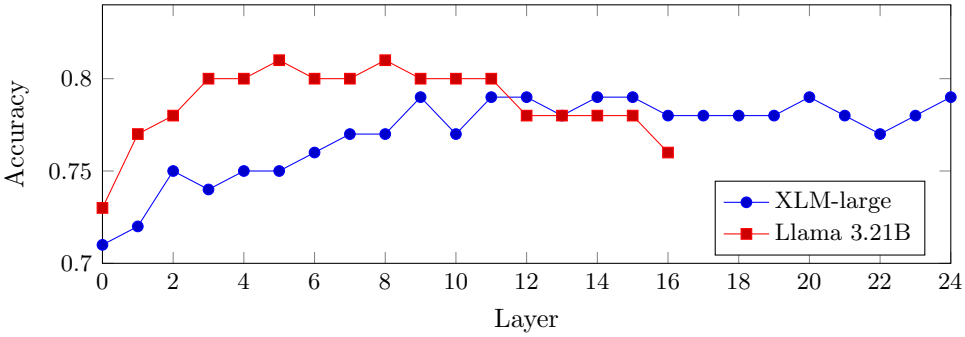

Figure 3: Accuracy across layers for large models in Galician using original embeddings.

Regarding the impact of words encountered during training on subsequent performance, we followed the approach of previous research (e.g., XL-WiC) to assess this effect. The hypothesis is that words in the test set that were seen during training (in different sentences) may be easier to disambiguate. As illustrated in Figure 4, the disambiguation performance of words seen during training (Seen) is better than those new (Unseen) words, where the XLM-base model achieves an accuracy of 0.85, representing the highest level within that category. However, in both the total (All) and the unseen categories, no clear differences were observed among the models, except for Llama, whose results for seen and unseen words are relatively similar.

4.2.2 Spanish results

The results of the Baseline experiments for Spanish are summarized in Table 6. The models that demonstrated the most favorable performance (0.63) in this experiment were Beto-base, XLM-large, and XLM-base. However, these results are inferior to those obtained with the Galician dataset (0.66) and original WiC (0.638). The discrepancy can be attributed to the limited size of the Spanish development dataset (200 instances), which was utilized for training, but further research is needed to understand the causes.

To assess the cross-lingual transfer capabilities of the LMs, three experiments were conducted with different training data (Ta-

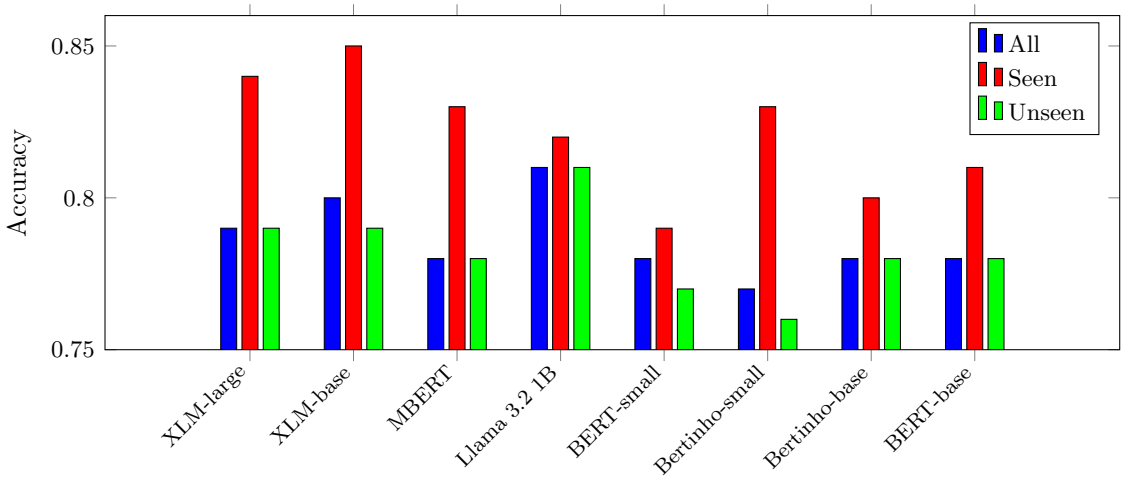

Figure 4: Accuracy depending on whether the target word was seen or not during the training for the Galician dataset.

Model	Acc.	Thr.	Layer
BETO-base	**0.63**	0.66	5 (N)
BERTIN-base	0.56	0.74	2 (N)
		0.58	12 (N)
RoBERTa-base	0.61	0.42	7 (N)
		0.58	10 (N)
RoBERTa-large	0.60	0.52	12 (N)
mBERT	0.60	0.68	6 (N)
		0.40	9 (N)
XLM-large	**0.63**	0.48	21 (N)
XLM-base	**0.63**	0.56	12 (N)
Llama 3.2 1B	0.57	0.62	5 (N)
fasttext	0.54	0.74	– (N)

Table 6: Baseline results for Spanish. *Thr.* refers to the cosine similarity threshold, and *Acc.* to the accuracy. *Layer* includes, besides the transformer layer, whether it uses the Cosine from the original vector or the Normalized one.

ble 7 shows the results of the accuracy using both the original and the normalized embeddings, EMB and NORM, respectively).

The initial experiment was carried out to evaluate the performance of both monolingual and multilingual models when trained on a limited Spanish dataset (ES column). In particular, 200 instances from the development dataset were utilized. Despite using only a small amount of training data, the supervised classifiers achieved the best performance for some models (e.g., Llama). However, it is worth noting that in several cases, the baseline approach outperformed the supervised classifiers. Concerning the models evaluated, there is no difference between monolingual and multilingual models. The second experiment involved training multilingual models using two sets: Galician training and Spanish development. However, contrary to expectations, this setup led to a significant drop in performance compared to the model trained on Spanish data alone (ES+GL column). This suggests that introducing Galician and Spanish data in a multilingual model may have introduced noise or interference, leading to worse results than when the model was exclusively trained on Spanish data. The last experiment aimed to train multilingual models using only the Galician training dataset (GL column in Table 7). This setup resulted in the worst performance. Concerning the normalized vectors, as in Galician, there is no notable difference.

Figures 5 and 6 show the evolution of accuracy along the layers when the Spanish train set was used. For the base models, the results are fairly balanced across the layers. Nevertheless, in Figure 6 a decrease is observed from layer 14 onwards, which then increases in the final three layers for the RoBERTa-large and XLM-large models.

Finally, Figure 7 shows that, as the observations made in Galician, the performance of the words that were used in training (Seen) have a better disambiguation compared to those unseen words.

In sum, while the supervised models for Galician clearly improved upon the baselines, this was not the case for Spanish, where performance dropped when using cross-lingual training data.

Model	ES				ES+GL				GL			
	EMB		NORM		EMB		NORM		EMB		NORM	
	Acc	L	Acc	L	Acc	L	Acc	L	Acc	L	Acc	L
BETO-base	0.62	4	0.61	2-3-4	–	–	–	–	–	–	–	–
BERTIN-base	0.60	3	0.60	1-3-4	–	–	–	–	–	–	–	–
RoBERTa-base	0.62	8	0.61	8	–	–	–	–	–	–	–	–
RoBERTa-large	0.62	19	0.62	0	–	–	–	–	–	–	–	–
mBERT	0.60	0	0.61	0	0.56	6	0.55	0-2	0.52	7	0.52	3-4-6
XLM-base	0.60	4-5	0.61	10	**0.57**	4	0.57	4-8	**0.53**	8	0.55	8
XLM-large	0.62	19	0.62	24	0.56	17	0.57	22	0.52	3	0.53	22
Llama 3.2 1B	**0.64**	12	0.63	7-11-13	0.54	4	0.54	4-12	0.51	2-15	0.51	1-13-15

Table 7: Summary of the performance comparison for the Spanish dataset using original embeddings *EMB* and normalized embeddings *NORM*. *Acc.* refers to the accuracy and *L* to the layer. For cases with more than one layer with the same accuracy, all layers are represented with a hyphen.

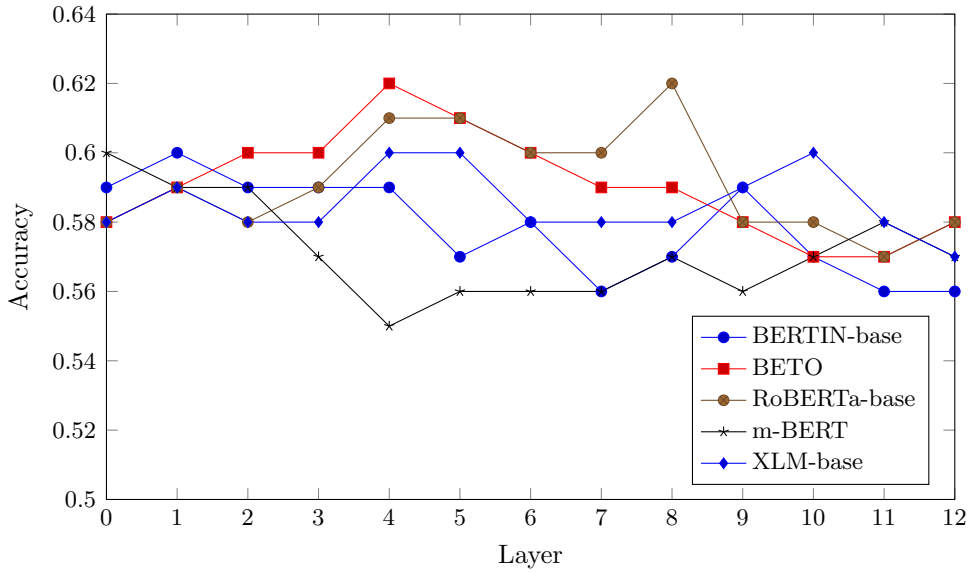

Figure 5: Accuracy across layers for base models in Spanish using original embeddings.

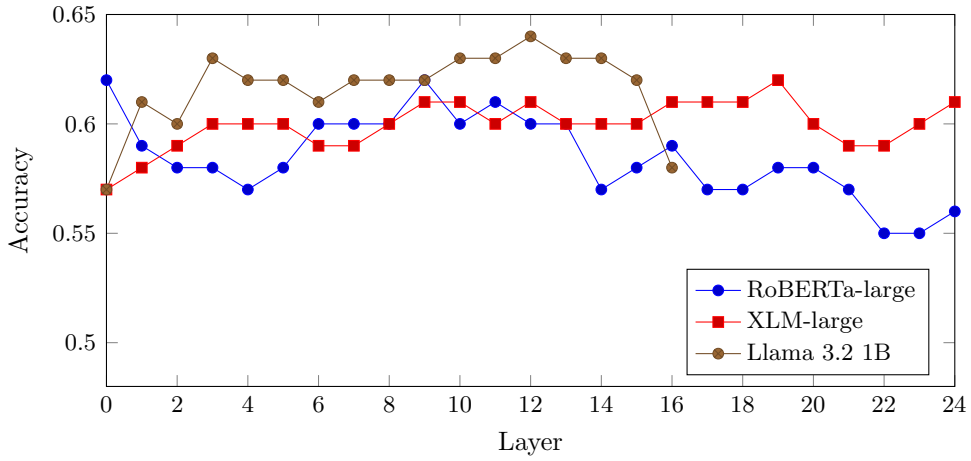

Figure 6: Accuracy across layers for large models in Spanish using original embeddings.

5 *Conclusions and further work*

This paper introduced Word-in-Context datasets for Galician and Spanish, designed to evaluate the lexical capabilities of vector models in this lexical disambiguation task. In this study, we conducted a systematic anal-

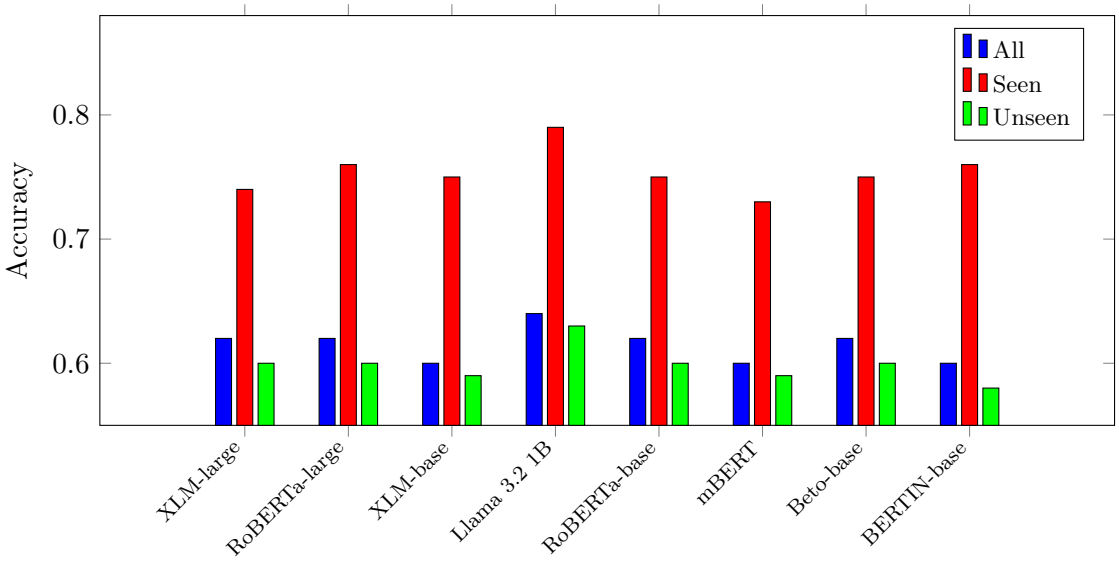

Figure 7: Accuracy depending on whether the target word was seen or not during the training for the Spanish dataset.

ysis of multilingual and monolingual transformer models in addressing lexical ambiguity. The results reveal that there is a comparable accuracy in the monolingual and multilingual models. Additionally, rogue dimensions seem to have little impact on the models' performance, and the use of Galician training data has negative effects on the classifiers for Spanish. This study highlights the potential of these datasets for evaluating new methods for resolving lexical ambiguity in Galician and Spanish. The datasets will serve as valuable resources for future research and model evaluation. Additionally, there is potential for expanding the datasets to other languages, such as Portuguese, or incorporating cross-lingual data. In further work, we plan to extend the cross-lingual transfer experiments including other languages, and to perform qualitative analyses on the results on each language. We aim to leverage the insights gained from this type of analyses to develop more advanced strategies for addressing lexical ambiguity in language models.

Acknowledgments

This work was funded by MCIN/AEI/10.13039/501100011033 (grants with references PID 2021-128811OA-I00, and TED 2021-130295B-C33, the latter also funded by "European Union Next Generation EU/PRTR"), by the Galician Government (ED481A-2024-070, ERDF 2014-2020: Call ED431G 2019/04, and ED431F 2021/01), and by a Ramón y Cajal grant (RYC2019-028473-I). This publication was also produced within the framework of the Nós Project, which is funded by the Spanish Ministry of Economic Affairs and Digital Transformation and by the Recovery, Transformation, and Resilience Plan -Funded by the European Union - NextGenerationEU, with reference 2022/TL22/00215336, and by the Xunta de Galicia through the collaboration agreements signed in with the University of Santiago de Compostela.

References

Agirre, E. and P. Edmonds, editors. 2007. *Word Sense Disambiguation. Algorithms and Applications*, volume 33 of *Text, Speech and Language Technology*. Springer.

Aina, L., K. Gulordava, and G. Boleda. 2019. Putting words in context: LSTM language models and lexical ambiguity. *CoRR*, abs/1906.05149.

Apidianaki, M. 2022. From Word Types to Tokens and Back: A Survey of Approaches to Word Meaning Representation and Interpretation. *Computational Linguistics*, 49(2):465–523, 06.

Apresjan, J. D. 1974. Regular polysemy. *Linguistics*, 12(142):5–32.

Bevilacqua, M., T. Pasini, A. Raganato, and R. Navigli. 2021. Recent trends

in word sense disambiguation: A survey. In Z.-H. Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization, 8. Survey Track.

Bouma, G. 2024. SweWiC 2.0.

Breit, A., A. Revenko, K. Rezaee, M. T. Pilehvar, and J. Camacho-Collados. 2021. WiC-TSV: An evaluation benchmark for target sense verification of words in context. In P. Merlo, J. Tiedemann, and R. Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1635–1645, Online, April. Association for Computational Linguistics.

Cassotti, P., L. Siciliani, L. C. Passaro, M. Gatto, P. Basile, et al. 2023. Wic-ita at evalita2023: Overview of the evalita2023 word-in-context for italian task. *EVALITA*.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

Cruse, D. 1986. *Lexical Semantics*. Cambridge Textbooks in Linguistics. Cambridge University Press.

De la Rosa, J., E. G. Ponferrada, M. Romero, P. Villegas, P. G. de Prado Salas, and M. Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68(0):13–23.

Devlin, J., M.-W. Chang, K. Lee, and K. 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, June. Association for Computational Linguistics.

Fandiño, A. G., J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, and M. Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. American Psychological Association (APA).

Gamallo, P., D. Bardanca, J. R. Pichel, M. Garcia, S. Rodríguez-Rey, and I. de Dios-Flores. 2023. Nos_mt-opennmt-en-gl. `https://huggingface.co/proxectonos/NOS-MT-OpenNMT-en-gl`.

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

Garí Soler, A. and M. Apidianaki. 2021. Let's play mono-poly: Bert can reveal words' polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics*, 9:825–844, 08.

Godey, N., É. Clergerie, and B. Sagot. 2024. Anisotropy is inherent to self-attention in transformers. In Y. Graham and M. Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 35–48, St. Julian's, Malta, March. Association for Computational Linguistics.

Gonzalez-Agirre, A., E. Laparra, and G. Rigau. 2012. Multilingual central repository version 3.0. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan,

B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2525–2529, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Grave, E., P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. 2018. Learning word vectors for 157 languages. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Guinovart, X. G. 2011. Galnet: Wordnet 3.0 do galego. *Linguamática*, 3(1):61–67, Jun.

Kipper Schuler, K., A. Korhonen, and S. Brown. 2009. VerbNet overview, extensions, mappings and applications. In C. Chelba, P. Kantor, and B. Roark, editors, *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts*, pages 13–14, Boulder, Colorado, May. Association for Computational Linguistics.

Liu, A., Z. Wu, J. Michael, A. Suhr, P. West, A. Koller, S. Swayamdipta, N. A. Smith, and Y. Choi. 2023. We're afraid language models aren't modeling ambiguity.

Liu, Q., F. Liu, N. Collier, A. Korhonen, and I. Vulic. 2021a. Mirrorwic: On eliciting word-in-context representations from pretrained language models. *CoRR*, abs/2109.09237.

Liu, Q., E. M. Ponti, D. McCarthy, I. Vulić, and A. Korhonen. 2021b. AM2iCo: Evaluating word meaning in context across low-resource languages with adversarial examples. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7151–7162, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Loureiro, D., K. Rezaee, M. T. Pilehvar, and J. Camacho-Collados. 2020. Language models and word sense disambiguation: An overview and analysis. *CoRR*, abs/2008.11608.

Loureiro, D., K. Rezaee, M. T. Pilehvar, and J. Camacho-Collados. 2021. Analysis and Evaluation of Language Models for Word Sense Disambiguation. *Computational Linguistics*, 47(2):387–443, 07.

Machina, A. and R. Mercer. 2024. Anisotropy is not inherent to transformers. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4892–4907, Mexico City, Mexico, June. Association for Computational Linguistics.

Martelli, F., N. Kalach, G. Tola, and R. Navigli. 2021. SemEval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (MCL-WiC). In A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, and X. Zhu, editors, *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 24–36, Online, August. Association for Computational Linguistics.

Miller, G. A., C. Leacock, R. Tengi, and R. T. Bunker. 1993. A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Misra, K. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Navigli, R. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Navigli, R. and S. P. Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Ortega, P., J. Luque, L. Lamiable, R. López, and R. Benjamins. 2024. Word sense disambiguation in native spanish: A comprehensive lexical evaluation resource.

Pasini, T., A. Raganato, and R. Navigli. 2021. Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13648–13656, May.

Pedersen, B. S., N. Sørensen, S. Olsen, S. Nimb, and S. Gray. 2024. Towards a danish semantic reasoning benchmark-compiled from lexical-semantic resources for assessing selected language understanding capabilities of large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16353–16363.

Pilehvar, M. T. and J. Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Raganato, A., J. Camacho-Collados, and R. Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In M. Lapata, P. Blunsom, and A. Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain, April. Association for Computational Linguistics.

Raganato, A., T. Pasini, J. Camacho-Collados, and M. T. Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online, November. Association for Computational Linguistics.

Rivière, P. D., A. L. Beatty-Martínez, and S. Trott. 2024. Evaluating contextualized representations of (spanish) ambiguous words: A new lexical resource and empirical analysis. arXiv preprint 2406.14678. To appear at NAACL 2025.

Timkey, W. and M. van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Tuggy, D. 1993. Ambiguity, polysemy, and vagueness. *Cognitive Linguistics*, 3(4):273–290.

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

Vázquez Abuín, M. and M. Garcia. 2024. WordNet Expansion with Bilingual Word Embeddings and Neural Machine Translation. In *Progress in Artificial Intelligence. 23rd EPIA Conference on Artificial Intelligence*, volume 14969 of *Lecture Notes in Computer Science*, pages 280–291. Springer.

Wang, A., Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

A Appendix

Target W.	Label	Sentence 1	Sentence 2
paper/role	F	the idea of an office withouh *paper* is absurd	what is your *role* in the team?
acceptable	T	the performances ranged from *acceptable* to excellent	*acceptable* radiation levels

Table 8: Translation of Table 2.

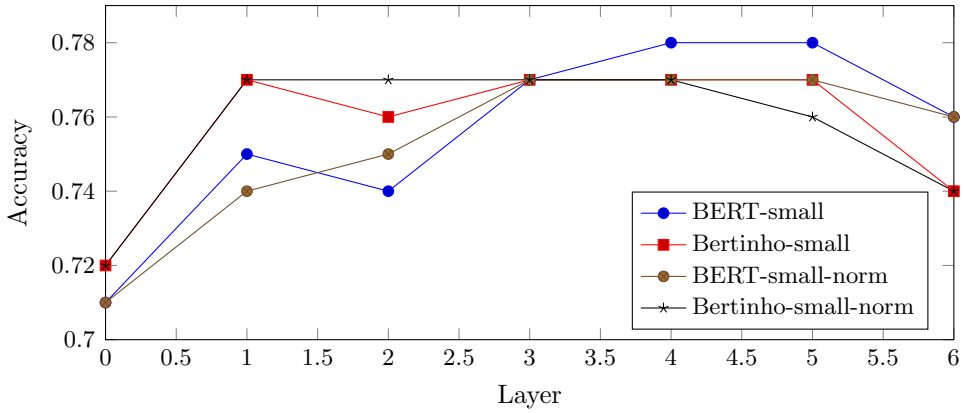

Figure 8: Accuracy across layers for small models in Galician, both original and normalized.

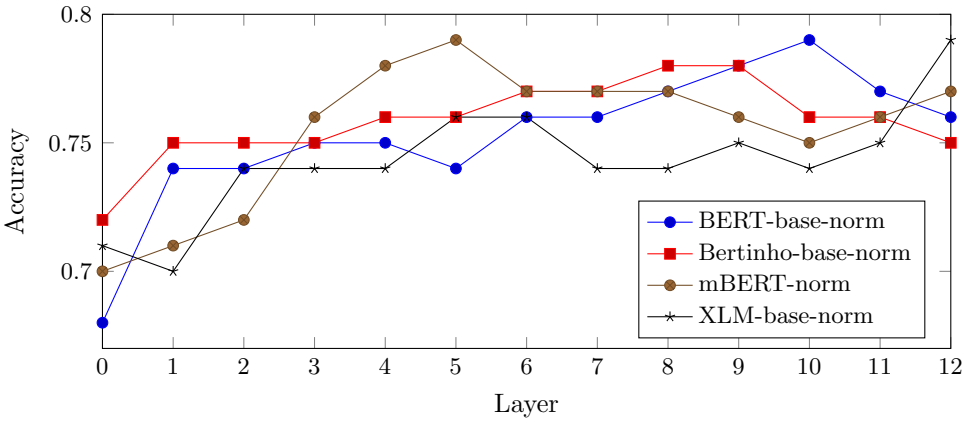

Figure 9: Accuracy across layers for 12-layer models in Galician, both original and normalized.

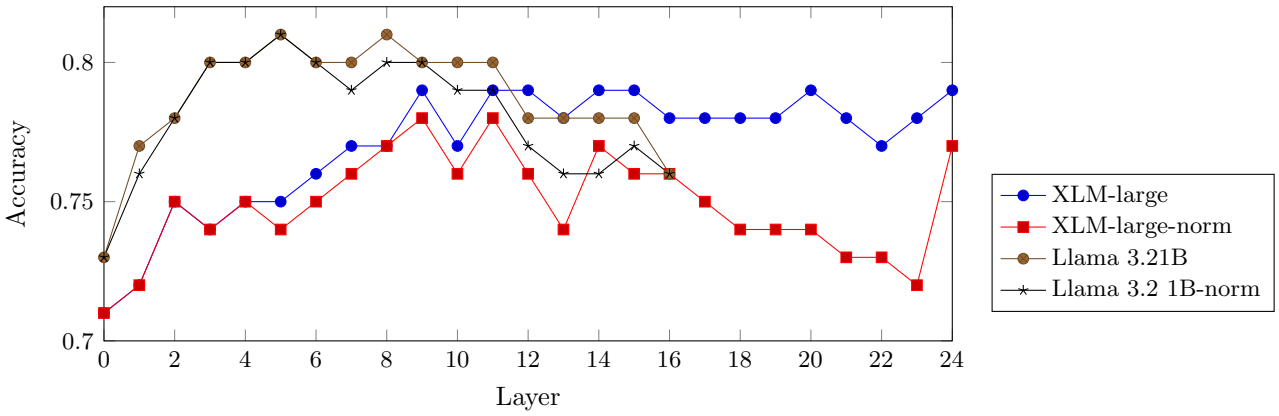

Figure 10: Accuracy across layers for large models in Galician, both original and normalized.

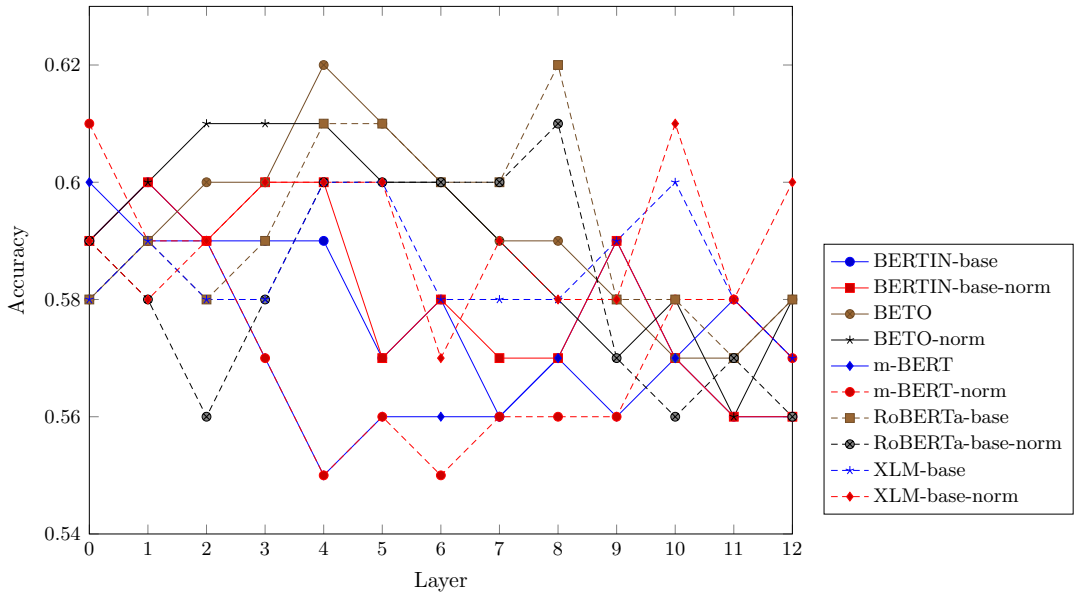

Figure 11: Accuracy across layers for base models in Spanish, both original and normalized.

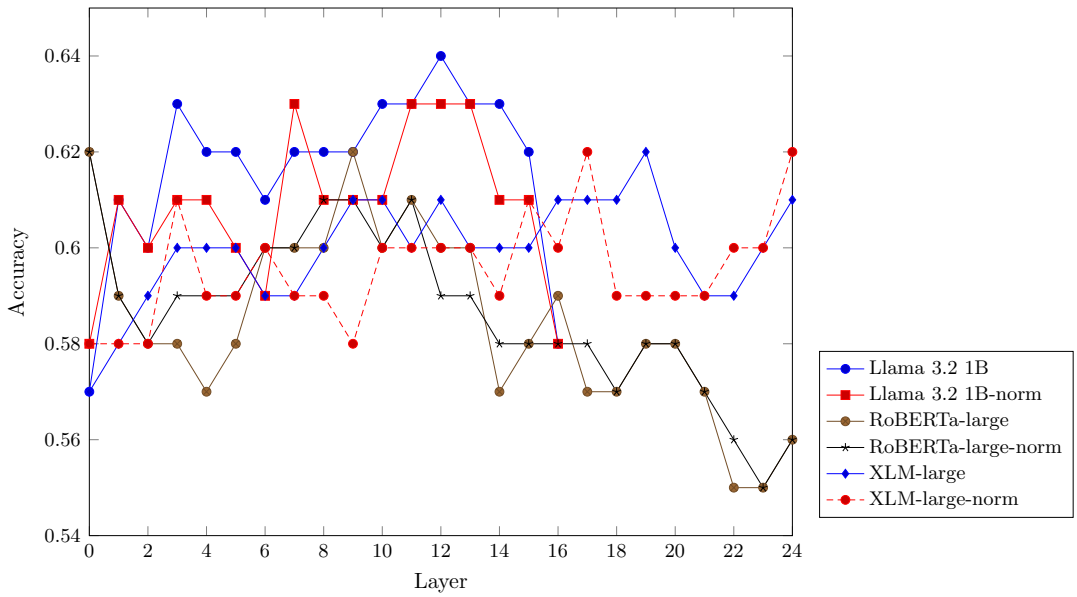

Figure 12: Accuracy across layers for large models in Spanish, both original and normalized.