A Knowledge-Graph-Based Intrinsic Test for Benchmarking Medical Concept Embeddings and Pretrained Language Models

Claudio Aracena[1], **Fabián Villena**[1,2], **Matías Rojas**[1,2], **and Jocelyn Dunstan**[1,2]
[1]Faculty of Physical and Mathematical Sciences, University of Chile
[2]Center for Mathematical Modeling, University of Chile
{claudio.aracena,fabian.villena,jdunstan}@uchile.cl
matias.rojas.g@ug.uchile.cl

Abstract

Using language models created from large data sources has improved the performance of several deep learning-based architectures, obtaining state-of-the-art results in several NLP extrinsic tasks. However, little research is related to creating intrinsic tests that allow us to compare the quality of different language models when obtaining contextualized embeddings. This gap increases even more when working on specific domains in languages other than English. This paper proposes a novel graph-based intrinsic test that allows us to measure the quality of different language models in clinical and biomedical domains in Spanish. Our results show that our intrinsic test performs better for clinical and biomedical language models than a general one. Also, it correlates with better outcomes for a NER task using a probing model over contextualized embeddings. We hope our work will help the clinical NLP research community to evaluate and compare new language models in other languages and find the most suitable models for solving downstream tasks.

1 Introduction

In healthcare, text plays a role of enormous importance. One of the media that a medical practitioner can persist is the text in clinical records (Dalianis, 2018). Text is one of the richest forms of information inside the electronic health record, so it is fundamental to develop tools to extract information from these text sources. To create these tools in this field, we must pay special attention to ensuring quality and reproducibility.

Analyzing unstructured texts written by humans is challenging since it is complex to formally understand and describe the rules governing human language, as it is ambiguous and constantly evolving. Natural Language Processing (NLP) is an interdisciplinary field of artificial intelligence that seeks to develop algorithms capable of understanding, interpreting, and manipulating these unstructured texts (Jurafsky and Martin, 2000).

In the medical context, using NLP helps to address tasks such as extracting medical entities, disease coding, text classification, and relation extraction, among others. However, one of the steps before solving any of these tasks is to create robust numerical representations of the text so that the computer can handle this data.

Word embeddings are dense, semantically meaningful vector representations of a word. These models have proven to be a fundamental building block of neural network-based architectures (Lample et al., 2016). Although these models have obtained excellent results for several NLP tasks, their main drawback is that they provide a single-word representation in a given document. This is not optimal since a word meaning may depend on the sentence in which it appears. This type of word embedding is known as static word embeddings.

Contextual representation models handle this issue by creating word representations based on sentence-level context. These representations are commonly retrieved from pretrained language models (PLM). Classic examples of these models are ELMO, BERT, RoBERTa, Flair, ALBERT, among others. However, contextualized word embeddings may not represent words as well as static ones, as results obtained in Reimers and Gurevych (2019) suggest.

Although contextualized word embeddings have these drawbacks, we can use these numeric representations of words to understand PLM representations. Specifically, we are interested in studying how domain-specific and general-domain PLM represent clinical and biomedical concepts. In this study, we aim to create a simple and efficient test for measuring concept embeddings' quality and comparing clinical and biomedical PLM performance using a relevant knowledge base and graph, the Unified Medical Language System (UMLS).

A knowledge graph is an extensive network of

Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI), pages 197 - 206

entities relevant to a specific domain. The network describes each entity's semantic types, properties, and relationships. Knowledge graphs represent real-world entities and their relations in a graph, define possible classes, and allow to relate arbitrary entities with each other (Ehrlinger and Wöß, 2016).

The UMLS is a knowledge graph that combines many clinical and biomedical vocabularies and standards to enable interoperability between computer systems (Bodenreider, 2004). The UMLS consists of multiple knowledge sources. One is the metathesaurus, a large, multi-purpose, and multilingual vocabulary database that contains information about biomedical and clinical-related concepts, their various names, and their relationships. Another source is the semantic network, a consistent categorization of all concepts represented in the metathesaurus, providing a set of valuable relationships between these concepts. In this work, we used both knowledge sources.

Two testing frameworks have been developed to measure the quality of language representations. First, an extrinsic test framework that uses the language representations to construct a more complex architecture to solve a specific downstream task. Second, an intrinsic test framework that measures the capacity of the language representation to resolve semantic questions regarding the language domain it represents (Zhai et al., 2016; Wang et al., 2019; Bakarov, 2018).

To construct intrinsic tests, we must compose questions based on a source of truth. This source can be expert knowledge, where we ask human experts to write each one of these questions manually, or we can use a knowledge base to compose these questions automatically. We used the UMLS knowledge graph to automatically derive a concept similarity intrinsic test using the length of the shortest path in the graph to compute a true similarity measure between concepts.

This intrinsic test will be used as a metric to check how good language representations are, but also as a comparison measure of whether clinical and biomedical PLM are better compared to general ones in downstream tasks such as Named Entity Recognition (NER).

2 Related work

PLM such as BERT (Devlin et al., 2019), ELMo (Peters et al., 2018), and GPT-2 (Radford et al., 2019) are able to produce contextualized word embeddings. It has been shown that contextualized word embeddings can achieve near state-of-the-art performance in tasks such as POS tagging or NER using probing models (Liu et al., 2019). Additionally, contextualized word embeddings from top layers of PLM produce more context-specific and anisotropic representations (Ethayarajh, 2019).

Regarding the clinical and biomedical domain in English, there are several models to obtain contextualized embeddings, such as BioELMo (Jin et al., 2019), Clinical BERT (Alsentzer et al., 2019), SciBERT (Beltagy et al., 2019), BioBERT (Lee et al., 2020), among others. However, there remains a significant lack of language models in Spanish. The only models available are SciELO Flair (Akhtyamova et al., 2020), Clinical Flair (Rojas et al., 2022b), and clinical and biomedical versions of RoBERTa (Carrino et al., 2022). Although these studies have shown that incorporating domain-specific contextualized embeddings significantly improves the models' performance in several extrinsic tasks, comparing their performances with intrinsic tests is still necessary.

Since PLM creates word-level contextual representations, it is necessary to define a method for combining these vectors to create sentence-level embeddings. For this purpose, a popular technique is the mean pooling of contextual word embeddings (Reimers and Gurevych, 2019). However, this method may lead to poor results if the PLM is not explicitly trained for similarity. Another study has proposed transforming the distribution of sentence-level embeddings to generate isotropic and smooth representations (Li et al., 2020). Creating these sentence-level representations is fundamental for testing the intrinsic tests proposed in this research.

Common approaches to evaluate biomedical PLM performance are benchmarks such as BLUE (Peng et al., 2019) and BLURB (Gu et al., 2021), which are built for the English language. There is no relevant benchmark in Spanish, and every author selects some annotated datasets to evaluate PLM performance on specific downstream tasks. Although the amount of annotated datasets in Spanish is growing, there is a lack of intrinsic tasks that can help to understand if a PLM is improving, and this research tries to fill that gap.

3 Methods

Our proposed method creates a semantic similarity intrinsic test with medical concept pairs and their semantic distances. We extracted these concept pairs from the UMLS[1] term graph and computed their distances as the length of the shortest directed path of parent relationships between the concepts. We measured the correlation of the knowledge-graph-derived distance to the cosine similarity of the terms string descriptions on an embedding space projected using different language representations. Finally, we compare these correlations with the performance on downstream tasks of each language representation.

3.1 Concept pair selection and its graph distances

In this vocabulary database, a concept is simply the meaning of a medical entity. Each concept in the metathesaurus has a unique and permanent concept identifier (CUI).

A UMLS concept can have multiple names because the same meaning can be described with numerous strings, for example, in different languages or source vocabularies. Each concept named description is called an atom and is identified by an atom identifier (AUI). To select a single concept description, we filtered out the atoms marked as non-preferred in the metathresaurus. With this filter and by only selecting atoms in Spanish, we assigned a single string describing each medical concept. In the UMLS Semantic Network, concepts are related using multiple relation types. The only relation type we used to connect the concepts was the parent relationship (PAR). We tried other relationship types but continued with PAR relationships because they are the most frequent. Child relationships (CHD) have the same frequency as PAR relationships, given they are the inverse relation type of PAR. Thus we can choose any of them.

After the previous step, we imported concepts and their PAR relations into a graph database[2]. Next, we queried the graph to select several random concepts and recursively extracted direct or related concepts at multiple distances. This means there is a path of one or more PAR relations of distance between pairs of concepts, as shown in Figure 1. Given that sometimes it is possible to find multiple paths between two concepts, we only used the shortest path between them. This process allowed us to extract the path length between two concepts. We select 20,000 concepts for this study to conduct the intrinsic tests rapidly. However, we can choose more concepts if necessary.

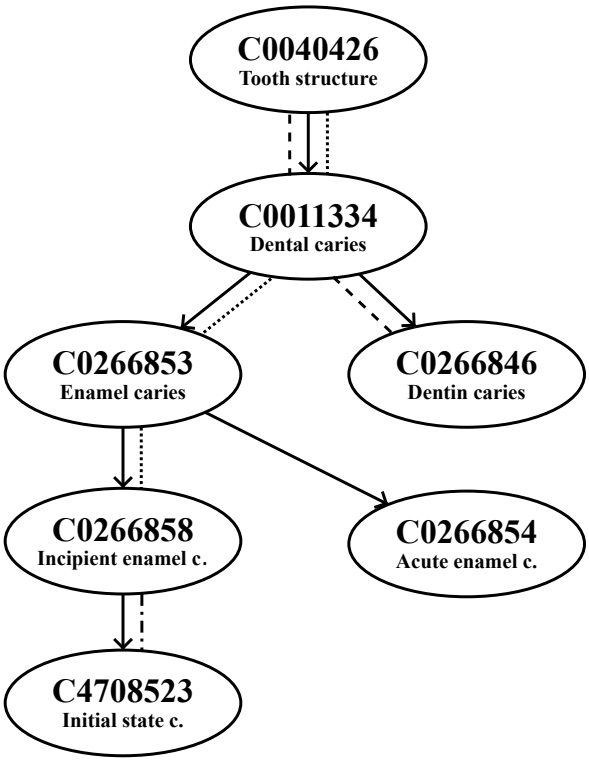

Figure 1: PAR-related concepts from C0040426 (Tooth structure). We highlight multiple paths,

- A dash-dot line represents the path between C0266858 (Incipient enamel caries) and C4708523 (Initial state caries) with a distance of 1 PAR edge.
- A dash-dash line represents the path between C0040426 (Tooth structure) and C0266846 (Dentin caries) with a distance of 2 PAR edges.
- A dot-dot line represents the path between C0040426 (Tooth structure) and C0266858 (Incipient enamel caries) with a distance of 3 PAR edges.

3.2 Generation of UMLS concepts' embeddings

After selecting the pairs of concepts and their descriptions, we generate concepts' embeddings using PLM. As UMLS concepts may contain more than one token, extracting embeddings that can represent the whole concept and not just one

[1] version 2022AA

[2] Neo4j (https://neo4j.com/)

word is essential. To do this, we used mean pooling of embeddings obtained for concept tokens from a PLM. For models hosted in the Huggingface Model Repository[3], we used the Python library `sentence-transformers`[4] (Reimers and Gurevych, 2019), and for models hosted in the Flair repository, we used the Python library `flair`[5] (Akbik et al., 2019).

All of our experiments were conducted for Spanish language datasets. We generated concept embeddings for several PLM of interest with different base architectures and domains. For the base architectures, we selected BERT, RoBERTa, and Flair. As for the domain, we chose, whenever possible, general, biomedical, and clinical models. As we did not find a publicly available BERT linguistic model for the clinical domain trained on Spanish text, we tuned a general domain model in Spanish (Cañete et al., 2020) with clinical text obtained from the Chilean Waiting List Corpus (Báez et al., 2020, 2022).

3.3 Implementation of intrinsic test

We build our intrinsic test as follows. First, we calculate the cosine similarity between concept embedding pairs. Then, we obtained the Spearman correlation between cosine similarity and path length, which we called ρ. This simple process allowed us to get our first metric. We expect that a greater path length between two concepts will result in a lower cosine similarity, given that they are farther semantically. Therefore, the Spearman correlation (ρ) between these two distances over all concepts pairs will be negative. If we compare embeddings generated by different PLM, we could expect that more domain-specific PLM will generate embeddings with more semantic differences between concepts within the domain, resulting in a more negative ρ. Thus, a more negative ρ indicates a PLM that can separate better semantically concepts within a domain.

As a part of our analysis, we calculated the average cosine similarity per path length. This step led us to obtain a complementary metric, the difference of mean cosine similarity for the shortest path length and the longest path length, that we called δ. The rationality behind this metric is similar to what we found in the previous one. However, in this case, a more positive δ indicates a PLM that can better separate concepts semantically within a domain.

3.4 Comparison with extrinsic test

Our intrinsic metrics were compared to extrinsic metrics using the F1 score in relevant biomedical and clinical NER datasets. The idea of incorporating extrinsic tests is to check if having better values of our intrinsic metrics will translate into better performance in downstream tasks for the selected PLM.

To build a reproducible extrinsic comparison for all PLM base architectures, we create a probing task for NER. In other words, we extracted contextualized embeddings from a PLM without fine-tuning for any downstream task, and those embeddings were input into a linear layer trained for NER.

The clinical and biomedical datasets in Spanish used for the NER probing task were:

- CANTEMIST[6] (Miranda-Escalada et al., 2020): annotated corpus with tumor morphology mentions in 1,301 oncological clinical case reports.
- PharmaCoNER[7] (Gonzalez-Agirre et al., 2020): annotated corpus with entities such as chemical compounds and drugs in 1,000 clinical case studies.
- CT-EBM-SP[8] (Campillos-Llanos et al., 2021): annotated corpus with UMLS entities in 1,200 texts about clinical trials studies and clinical trials announcements.
- NUBes[9] (Lopez et al., 2020): annotated corpus with negation and uncertainty entities in anonymised health records (29,682 sentences).

4 Results

We queried 20,000 pairs of random atoms to select UMLS concepts from the graph database. Figure 2 shows the histogram of those pairs by path length. We can see that pair frequency increases as path length increase until seven parent relationships of

[3] https://huggingface.co/models
[4] https://github.com/UKPLab/sentence-transformers
[5] https://github.com/flairNLP/flair
[6] https://zenodo.org/record/3978041
[7] https://zenodo.org/record/4270158
[8] http://www.lllf.uam.es/ESP/nlpmedterm_en
[9] https://github.com/Vicomtech/NUBes-negation-uncertainty-biomedical-corpus

distance. After that point, the frequency of pairs decreases until it reaches 14 relations of distance. We removed all path lengths containing less than 300 pairs of concepts to calculate the metrics ρ and δ.

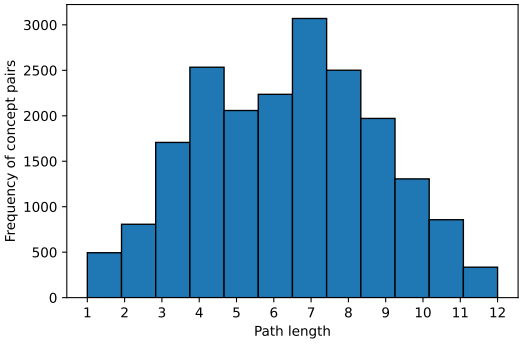

Figure 2: Histogram of UMLS concept pairs by path length

Then, we plot a boxplot of cosine similarity by path length for every PLM. Figure 3 shows such a boxplot for a general-domain BERT trained in Spanish text (Cañete et al., 2020)[10]. This plot allows us to understand how cosine similarity distributes along path length.

It is clear from the plot that average cosine similarity decreases as path length increases. However, the decline is near null or even negative from path length four onwards. Moreover, the average cosine similarity is not going near zero. We hypothesize this pattern is because all concepts are related to clinical and biomedical domains and also due to the anisotropic behavior of sentence embeddings obtained from PLM. As discussed in Ethayarajh (2019), contextualized embeddings obtained from PLM tend to distribute not evenly in the embedding space but in a small portion of it. Therefore, they still have a relatively high similarity when comparing dissimilar concepts.

To compare several PLM, we plot only average cosine similarity by path length for every language model, as shown in Figure 4. As we can see, average cosine similarity by path length varies for different base architectures and domains of PLM. However, they all repeat the same decline pattern as path length increases.

Similarly to Figure 3, Figure 4 does not show any average cosine similarity going near zero. However, the similarity level where each PLM stabilizes

[10]Other models' plots are included in the appendix

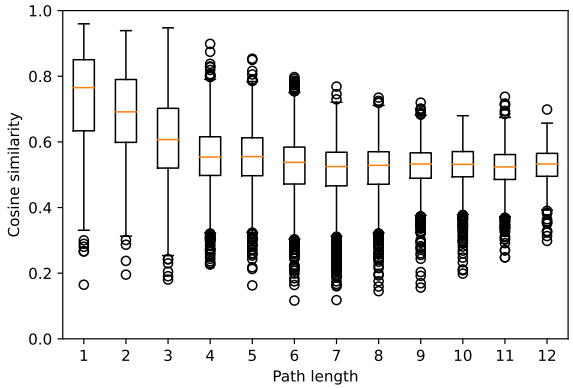

Figure 3: Boxplot of cosine similarity by path length for a general-domain BERT trained on Spanish text.

is different. Not surprisingly, language models trained on a similar corpus or being a fine-tuned version from another have comparable similarity levels. RoBERTa-es-clinical was trained with the same corpora as RoBERTa-es-biomedical plus a clinical corpus (Carrino et al., 2022), and BERT-es-clinical is a fine-tuned model from BERT-es-general over a clinical corpus.

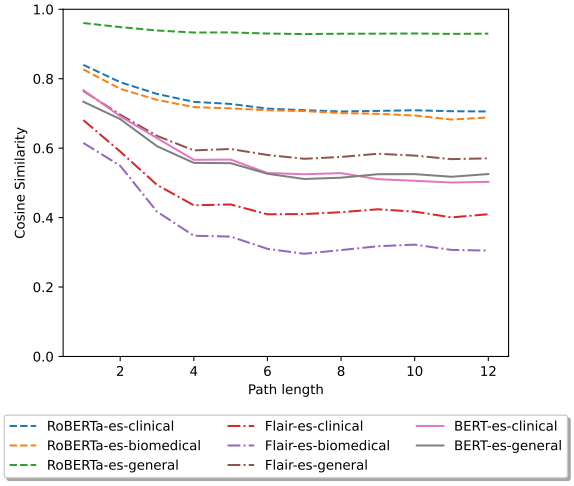

Figure 4: Average cosine similarity by path length for multiple language models

To measure the degree of the decline, we calculated the metrics ρ and δ for all the selected PLM, as shown in Table 1. We notice that ρ and δ are greater in absolute value for biomedical and clinical models than general ones within the same base architecture. This means that given a PLM base architecture, the degree of decline of the average cosine similarity is greater for domain-specific models than for general domain models. This finding suggests that domain-specific PLM and their concept embeddings better represent UMLS concepts;

Reference	Architecture	Domain	ρ	δ
Ours	BERT	Clinical	-0.38	0.25
(Cañete et al., 2020)	BERT	General	-0.30	0.18
(Akhtyamova et al., 2020)	Flair	Biomedical	-0.24	0.27
(Rojas et al., 2022b)	Flair	Clinical	-0.23	0.27
(Akbik et al., 2018)	Flair	General	-0.20	0.11
(Carrino et al., 2021)	RoBERTa	Clinical	-0.31	0.09
(Carrino et al., 2021)	RoBERTa	Biomedical	-0.28	0.13
(Gutiérrez-Fandiño et al., 2022)	RoBERTa	General	-0.23	0.03

Table 1: Correlations and differences for each language representation. The table is sorted ascending by ρ and then by base architecture. Every ρ is statistically significant.

Architecture	Domain	CANTEMIST	PharmaCoNER	CT-EBM-SP	NUBes
BERT	Clinical	0.739 (0.018)	0.577 (0.013)	**0.742 (0.012)**	0.791 (0.009)
BERT	General	**0.757 (0.004)**	**0.582 (0.007)**	0.714 (0.006)	**0.797 (0.013)**
Flair	Biomedical	**0.784 (0.006)**	**0.615 (0.013)**	**0.725 (0.008)**	0.792 (0.003)
Flair	Clinical	0.771 (0.009)	0.580 (0.021)	0.694 (0.000)	**0.802 (0.003)**
Flair	General	0.714 (0.013)	0.558 (0.002)	0.633 (0.002)	0.780 (0.005)
RoBERTa	Clinical	**0.794 (0.009)**	**0.633 (0.010)**	0.792 (0.012)	0.820 (0.004)
RoBERTa	Biomedical	0.784 (0.006)	0.626 (0.009)	**0.794 (0.014)**	**0.821 (0.005)**
RoBERTa	General	0.767 (0.014)	0.584 (0.006)	0.734 (0.005)	0.804 (0.003)

Table 2: F1 scores and standard deviations for NER probing task over four datasets in Spanish. The table is sorted according the same criteria as Table 1

hence the similarity pattern displayed. However, it is important to note that we do not find this behavior when comparing different base architectures.

We can see F1 scores for every NER probing task by PLM in Table 2. As expected, we can see a tendency to obtain better F1 scores for clinical or biomedical PLM than general ones. However, in the case of BERT architecture, results are mixed. We believe this behavior could be due to the creation of the clinical BERT model. Instead of being trained from scratch with clinical and biomedical data, it is a fine-tuned version of a general BERT. On the other hand, clinical and biomedical Flair and RoBERTa models were trained from scratch with domain-specific data.

Interestingly, when ρ metric is greater for a clinical model compared to a biomedical one, F1 scores for NER probing tasks are also greater, as we can see in the case of RoBERTa architecture for CANTEMIST and PharmaCoNER datasets. In the case of CT-EBM-SP and NUBes, there are no such differences, but F1 scores for clinical and biomedical are almost the same. On the contrary, when ρ metric is greater for a biomedical model compared to a clinical one, then F1 scores present a similar behavior, as we can see in the case of Flair architecture for CANTEMIST, PharmaCoNER, and CT-EBM-SP datasets. And as same as the previous situation, F1 scores for another dataset (NUBes) are almost the same. We do not observe this pattern for δ metric.

This finding suggests that ρ metric could be applied as a useful intrinsic test for comparing PLM within the same base architecture. However, it is important to note when comparing ρ metric for different base architectures, we do not find a clear relation with F1 scores. Consequently, we present the ρ metric as an intrinsic test to measure improvements for PLM within the same base architecture.

5 Conclusion and future work

Using domain-specific PLMs for downstream tasks has allowed reaching the state-of-the-art in several benchmarks. However, since these models are trained in large corpora, fine-tuning them or training from scratch is time-consuming. Therefore, before using these models to solve downstream tasks, it is crucial to create intrinsic tests that validate whether a domain-specific PLM yields better results than its base version.

In this study, we build an intrinsic test for clinical and biomedical PLM using contextualized em-

beddings and the UMLS knowledge graph. We suggest that our intrinsic test can help compare domain-specific PLM performance within its base architecture, which could be used to evaluate improvements when building PLM. Our experimental results show that this intrinsic test can capture improvements in clinical and biomedical PLM over general ones. Also, it correlates with better results in a NER probing task over four datasets in Spanish.

In future work, we can implement this study for other languages. Additionally, we can compare our intrinsic test with other probing tasks such as POS-tagging or coreference or even other clinical downstream tasks such as patient mortality or unplanned readmission. On the other hand, since our experimental datasets contain nested entities, but for simplicity, they were ignored, we would like to explore the use of contextualized embeddings in models that can address them, such as those proposed in Rojas et al. (2022a). Finally, we can compare several experimental settings, such as multiple numbers of concept pairs.

Limitations

We can group the limitations of our study in the ones related to the graph knowledge, the selected PLM, comparison with other embedding techniques, and language. First, regarding graph knowledge, we could have chosen several random subsets of concept pairs of different lengths and types of relations to check if our findings are still present. Second, we selected three base architectures, and all of them were of encoder type. Third, we could have compared our results with static embeddings. And finally, we could have selected more languages for comparison.

Ethics Statement

We state that our work complies with the ACL Code of Ethics. We believe that our work could help the research community with a new tool for their work in clinical and biomedical PLM. Our study was based on publicly available and anonymized data to avoid the privacy issues that clinical data may raise.

Acknowledgements

This work was funded by ANID Chile: Basal Funds for Center of Excellence FB210005 (CMM) and FB210017 (CENIA); Millennium Science Initiative Program ICN17_002 (IMFD) and ICN2021_004 (iHealth), Fondecyt grant 11201250, and National Doctoral Scholarships 21211659 (C.A.) and 21220200 (F.V.). Regarding hardware, the research was partially supported by the supercomputing infrastructure of the NLHPC (ECM-02) and the Patagón supercomputer of Universidad Austral de Chile (FONDEQUIP EQM180042).

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Liliya Akhtyamova, Paloma Martínez, Karin Verspoor, and John Cardiff. 2020. Testing Contextualized Word Embeddings to Improve NER in Spanish Clinical Case Narratives. *IEEE Access*, 8:164717–164726.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Pablo Báez, Felipe Bravo-Marquez, Jocelyn Dunstan, Matías Rojas, and Fabián Villena. 2022. Automatic extraction of nested entities in clinical referrals in spanish. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(3):1–22.

Pablo Báez, Fabián Villena, Matías Rojas, Manuel Durán, and Jocelyn Dunstan. 2020. The Chilean waiting list corpus: a new resource for clinical named entity recognition in Spanish. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 291–300, Online. Association for Computational Linguistics.

Amir Bakarov. 2018. A survey of word embeddings evaluation methods. *arXiv preprint arXiv:1801.09536*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270.

Leonardo Campillos-Llanos, Ana Valverde-Mateos, Adrián Capllonch-Carrión, and Antonio Moreno-Sandoval. 2021. A clinical trials corpus annotated with UMLS entities to enhance the access to evidence-based medicine. *BMC Medical Informatics and Decision Making*.

Casimiro Pio Carrino, Jordi Armengol-Estapé, Asier Gutiérrez-Fandiño, Joan Llop-Palao, Marc Pàmies, Aitor Gonzalez-Agirre, and Marta Villegas. 2021. Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario. *CoRR*, abs/2109.03570.

Casimiro Pio Carrino, Joan Llop, Marc Pàmies, Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Joaquín Silveira-Ocampo, Alfonso Valencia, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. Pretrained biomedical language models for clinical NLP in Spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 193–199, Dublin, Ireland. Association for Computational Linguistics.

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*.

Hercules Dalianis. 2018. Characteristics of Patient Records and Clinical Corpora. In *Clinical Text Mining*, pages 21–34. Springer International Publishing, Cham.

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.

Lisa Ehrlinger and Wolfram Wöß. 2016. Towards a definition of knowledge graphs. In *SEMANTiCS*.

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.

Aitor Gonzalez-Agirre, Antonio Miranda-Escalada, Obdulia Rabal, and Martin Krallinger. 2020. PharmaCoNER corpus: gold standard annotations of Pharmacological Substances, Compounds and proteins in Spanish clinical case reports. *Zenodo*. Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano-Oller, Carlos Rodriguez-Penagos, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. MarIA: Spanish language models. *Procesamiento del Lenguaje Natural*, 68(0):39–60.

Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. 2019. Probing biomedical embeddings from language models. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 82–89, Minneapolis, USA. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 1st edition. Prentice Hall PTR, USA.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234 – 1240.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. 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 Lan-*

guage Technologies, Volume 1 (Long and Short Papers), pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Salvador Lima Lopez, Naiara Perez, Montse Cuadros, and German Rigau. 2020. NUBes: A Corpus of Negation and Uncertainty in Spanish Clinical Texts. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC2020)*, pages 5772–5781, Marseille, France. European Language Resources Association.

Antonio Miranda-Escalada, Eulalia Farré, and Martin Krallinger. 2020. Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in Spanish, corpus, guidelines, methods and results. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings*.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Matias Rojas, Felipe Bravo-Marquez, and Jocelyn Dunstan. 2022a. Simple yet powerful: An overlooked architecture for nested named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2108–2117, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Matías Rojas, Jocelyn Dunstan, and Fabián Villena. 2022b. Clinical flair: A pre-trained language model for Spanish clinical natural language processing. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 87–92, Seattle, WA. Association for Computational Linguistics.

Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C.-C. Jay Kuo. 2019. Evaluating word embedding models: methods and experimental results. *APSIPA Transactions on Signal and Information Processing*, 8:e19. Publisher: Cambridge University Press.

Michael Zhai, Johnny Tan, and Jinho Choi. 2016. Intrinsic and Extrinsic Evaluations of Word Embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

A Appendix

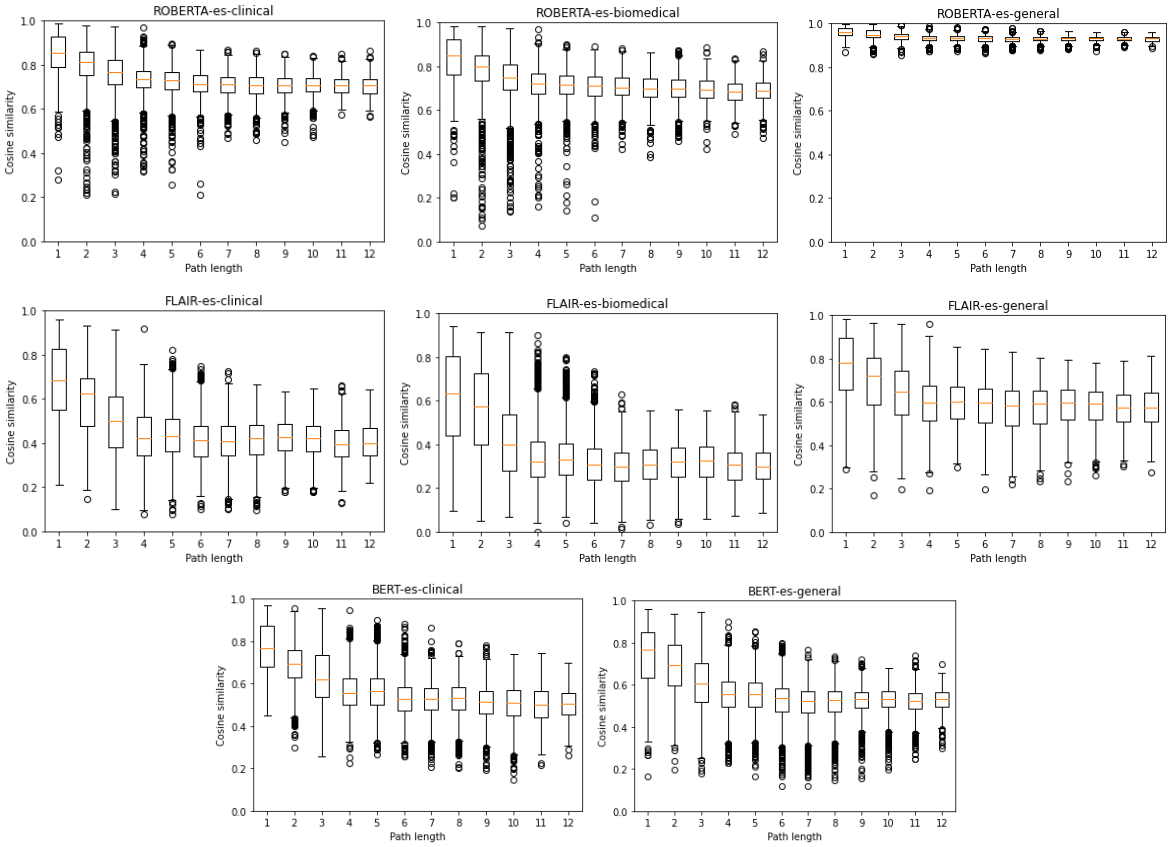

Figure 5: Boxplots of cosine similarity by path length for selected PLM trained in Spanish text