Evaluation Benchmarks for Spanish Sentence Representations

Vladimir Araujo[1,2]; Andrés Carvallo[1]; Souvik Kundu[5]; José Cañete ; Marcelo Mendoza[3]; Robert E Mercer[5]; Felipe Bravo-Márquez[4]; Marie-Francine Moens[2]; Alvaro Soto[1]

[1]Pontificia Universidad Católica de Chile, [2]KU Leuven , [3]Universidad Técnica Federico Santa María, [4]Universidad de Chile, [5]University of Western Ontario

Abstract

- Language models trained in languages other than English are still scarce.
- There are few ways to evaluate spanish language models'.
- We propose two benchmarks: Spanish SentEval and Spanish DiscoEval.
- We evaluate pre-trained Spanish language models.
- mBERT provides richer latent representations.

Introduction

Context

- Spanish is one of the most widely spoken languages.
- Proliferation of Spanish language models increases the need for annotated datasets to evaluate them.
- Most benchmarks focus on assessing word representations or basic linguistic knowledge.

Main benchmarks

- Two major benchmarks for evaluating language models in Spanish Language: adaptation of SentEval (Coneau et al, 2018) and DiscoEval (Chen et al, 2019).
- SentEval: evaluate quality of sentence representations.
- DiscoEval: evaluate discourse knowledge in sentence representations.

Contributions

- Compare available Spanish sentence encoders our proposed benchmarks.
- Expose the Spanish language models' current capabilities.

Methodology

Spanish SentEval:

- Sentence Classification (SC)
- Sentence Pair Classification (SPC)
- Semantic Similarity (STS)
- Linguistic Probing Tasks (LPT)

Spanish DiscoEval:

- Sentence Position (SP)
- Binary Sentence Ordering (BSO).
- Discourse Coherence (DC).
- Sentence Section Prediction (SSP).
- Discourse Relations (DR).

• Una historia policiaca que Scorsese la transforma en una memorable muestra del genero.

Figure 1. Example of SC task

Premise: Y yo estaba bien, ¡y eso fue todo!
Hypothesis: Después de que dije que sí, terminó.

Figure 2. Example of SPC task

• Un perro está con un juguete.
• Un perro tiene un juguete.

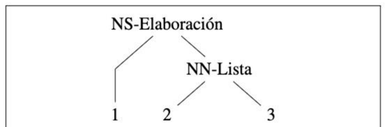
Figure 3. Example of STS task

• En enero participó en la infructuosa defensa de Forlí frente a César Borgia.

Figure 4. Example of LPT task

1) Se encontró que la adición de nanopartículas de sílice aumenta la rigidez del material. (2)
2) El objetivo de este trabajo es estudiar el efecto de la incorporación de nanopartículas de sílice en la rigidez de material.
3) Las Nanopartículas de sílice fueron sintetizadas utilizando el método sol-gel.
4) Las Nanopartículas de menor tamaño tienen un mayor efecto sobre las propiedades del material.
5) La rigidez del material aumentó hasta en un 80% con la adición de 30% de nanopartículas de silice.

Figure 5. Example of SP task.



[Se presentan algunos comentarios al Allgemeine Naturgeschichte und Theorie des Himmels escrito por Emmanuel Kant y publicado en 1755,]$_1$ [obra donde el pensador de Koenisberg dio a conocer sus principales ideas cosmológicas.]$_2$ [En particular se reseña su explicación cualitativa de cómo a partir de un material primordial tenue y difuso, la fuerza gravitacional produjo los cuerpos que forman el Sistema Solar.]$_3$

Figure 6. Example of RST spanish btree-bank for DR task.

Experiments

Models: Sent2Vec (Pagliardini et al, 2018), ELMo (Che et al, 2018), ELECTRA (Clark et al, 2021) RoBERTa-BNE (Gutierrez-Fandiño et al, 2021), BERTIN, BETO (Cañete et al, 2020) ,mBERT

Results

- Similar to their English language representations counterparts.
- mBERT obtain richer latent representation than models trained only in Spanish.

	SentEval				DiscoEval				
Models	**SC**	**SPC**	**SS**	**LPT**	**SP**	**BSO**	**DC**	**SSP**	**DR**
Sent2Vec	75.11	59.51	**76.05**	66.89	36.49	54.92	55.77	70.88	36.69
ELMo	71.50	**61.62**	62.06	69.90	37.13	55.13	58.68	72.60	45.14
ELECTRA	62.80	51.40	42.07	64.20	38.56	56.85	55.18	76.22	37.59
RoBERTa-BNE	72.51	54.57	41.34	68.22	41.82	57.02	56.31	76.83	39.21
BERTIN	73.54	55.47	32.53	67.72	41.66	56.66	55.54	**78.42**	45.86
BETO	**76.34**	58.17	55.37	69.38	41.43	57.53	60.89	75.33	47.84
mBERT	70.47	60.05	67.77	**71.41**	**43.21**	**57.97**	**63.45**	77.80	**51.08**

Table 1. Results for Spanish SentEval and Spanish DiscoEval by group.

Further analysis

- We perform a per-layer performance
- analysis of the representations learned by transformer-based models
- In general last layers learn best representations.

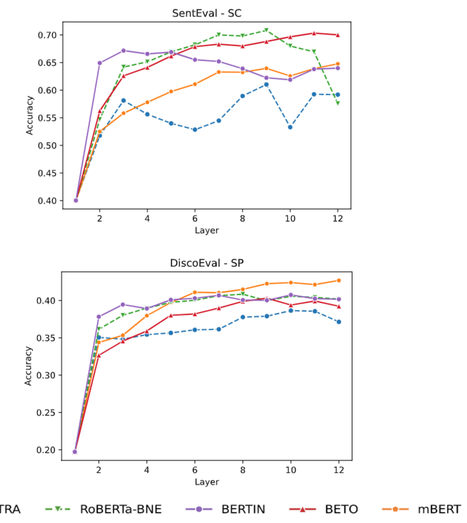


Figure 7: Per-layer performance evaluation on SentEval SC and DiscoEval SP tasks.

Conclusions

- We propose Spanish SentEval and DiscoEval to evaluate sentence representations.
- Results are consistent with English SentEval and DiscoEval.
- Multilingual BERT learns better representations compared to models trained in spanish.
- We observe last layers learn the best representations for all downstream tasks.

Future Work

- Include more tasks in these benchmarks.
- Perform other types of evaluations, namely stress or linguistic tests.

Contact Information

Vladimir Araujo
Pontificia Universidad Católica de Chile, KU Leuven
Email: vgaraujo@uc.cl
Website: https://vgaraujov.github.io/

References

1. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang,H., and Perez, J. (2020). Spanish pre-trained bert model and evaluation data. PML4DC at ICLR 2020.
2. Chen, M., Chu, Z., and Gimpel, K. (2019). Evaluation benchmarks and learning criteria for discourse-aware sentence representations. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International.
3. Devlin, J., Chang, M.-W., Lee, K., and Toutanova,K. (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
4. Gutierrez-Fandiño, A., Armengol-Estape, J., Pamies,M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., Gonzalez-Agirre, A., Armentano Oller, C., Rodriguez-Penagos, C., and Villegas, M. (2021). Spanish language mo

Acknowledgements

This work was supported by:
National Center for Artificial Intelligence CENIA FB210017,
Basal ANID. Felipe Bravo-Marquez was supported by ANID FONDECYT grant 11200290,
U-Inicia VID Project UI-004/20
ANID -Millennium Science Initiative Program - Code ICN17_002.

© Poster Template by Genigraphics® 1.800.790.4001 www.genigraphics.com