Portuguese Language Models and Word Embeddings

Ruan Chaves Rodrigues (UFG) * - ruanchaves93@gmail.com
Jéssica Rodrigues da Silva (UFSCar) ** - jsc.rodrigues@gmail.com
Pedro Vitor Quinta de Castro (UFG) * - pedrovitorquinta@inf.ufg.br
Nádia Félix Felipe da Silva (UFG) * - nadia@inf.ufg.br
Anderson da Silva Soares (UFG) * - anderson@inf.ufg.br

* : Institute of Informatics
Federal University of Goiás (UFG), Brazil
** : Department of Computer Science
Federal University of São Carlos (UFSCar), Brazil

March 2, 2020

DEEP LEARNING
BRASIL

Agenda

Introduction

- ASSIN datasets : Semantic Textual Similarity Task
  - ASSIN 1 (Fonseca et al. [2016])
  - ASSIN 2 (Real et al. [2020])

Text	Hypothesis	Similarity Score
Em comparação com o ano anterior, registaram-se menos 29 acidentes e menos duas vítimas mortais.	Feita a comparação com igual período do ano passado, registaram-se menos 29 acidentes e menos dois mortos.	5.0
O FC Porto renovou o contrato com o avançado colombiano Juan Quintero, até 2021.	O FC Porto confirmou nesta terça-feira a renovação com Juan Quintero.	4.0
A agência desceu a perspectiva do rating de Portugal, de "positiva" para "estável".	A agência elevou ainda a perspectiva do país de negativa para estável.	2.5
Este acessório prepara-se para dar um grande salto na nova versão do Surface Pro.	A Microsoft prepara-se para recolher os Surface Pro para a substituição dos respetivos cabos.	1.25

Related Work

- Portuguese word embeddings: Evaluating on word analogies and natural language tasks
  (Hartmann et al. [2017])
- Contextual Representations and Semi-Supervised Named Entity Recognition for Portuguese Language
  (de Castro et al. [2019])

Human error

- Conflicting replacement strategies for OOV words.
- Not applying the same preprocessing steps during evaluation and training.

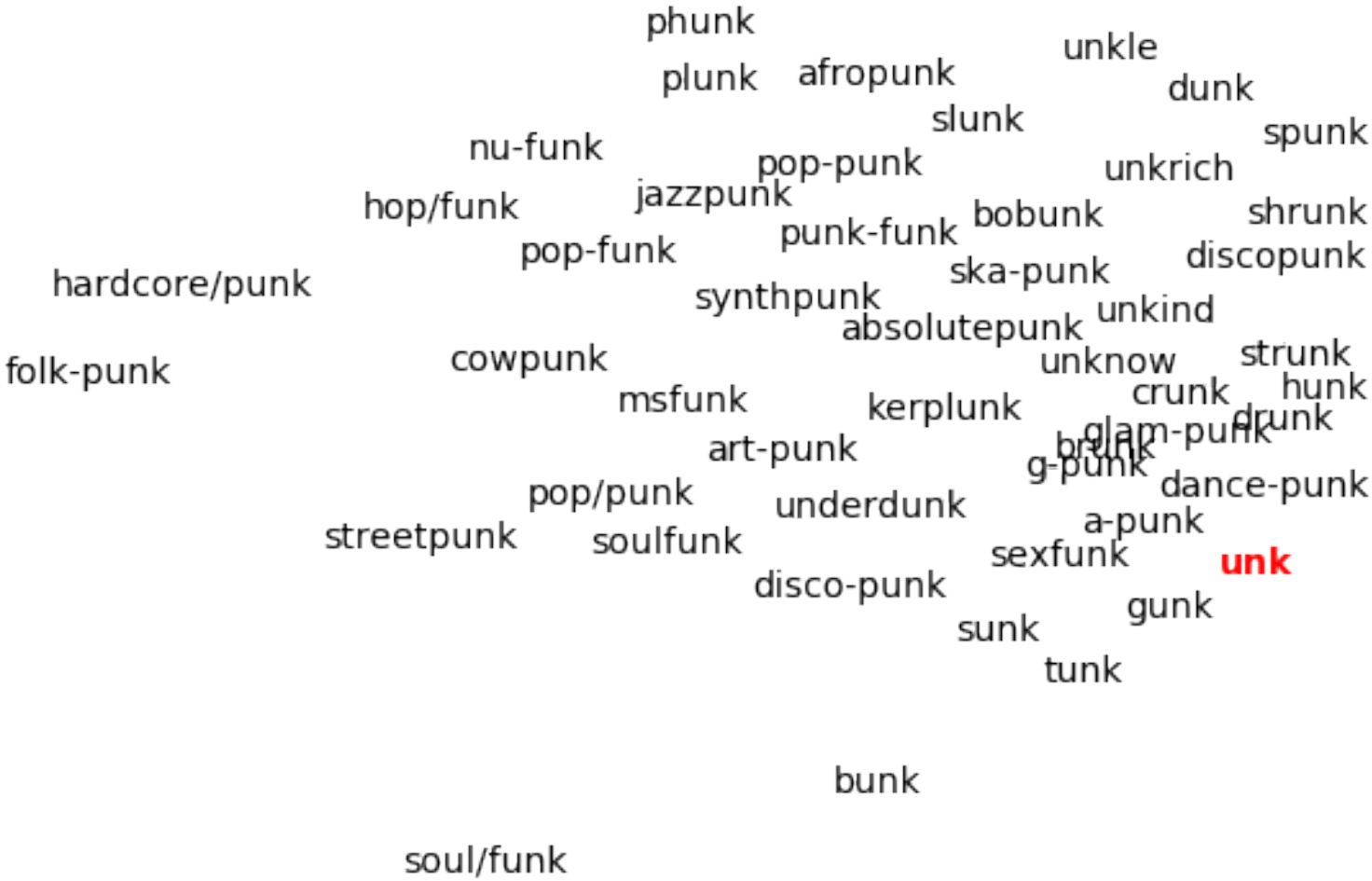
Example: FastText - OOV words replaced by "unk"

- Original approach by Hartmann et al. [2017]:
  - **Original phrase**:
    "Votaram contra a proposta 267 deputados, e 210 votaram a favor"
  - **Preprocessing**:
    "votaram", "contra", "proposta", "267", "deputados", "210", "votara", "favor"
  - **Replace unknown words by unk**:
    "votaram", "contra", "proposta", **"unk"**, "deputados", **"unk"**, "votara", "favor"

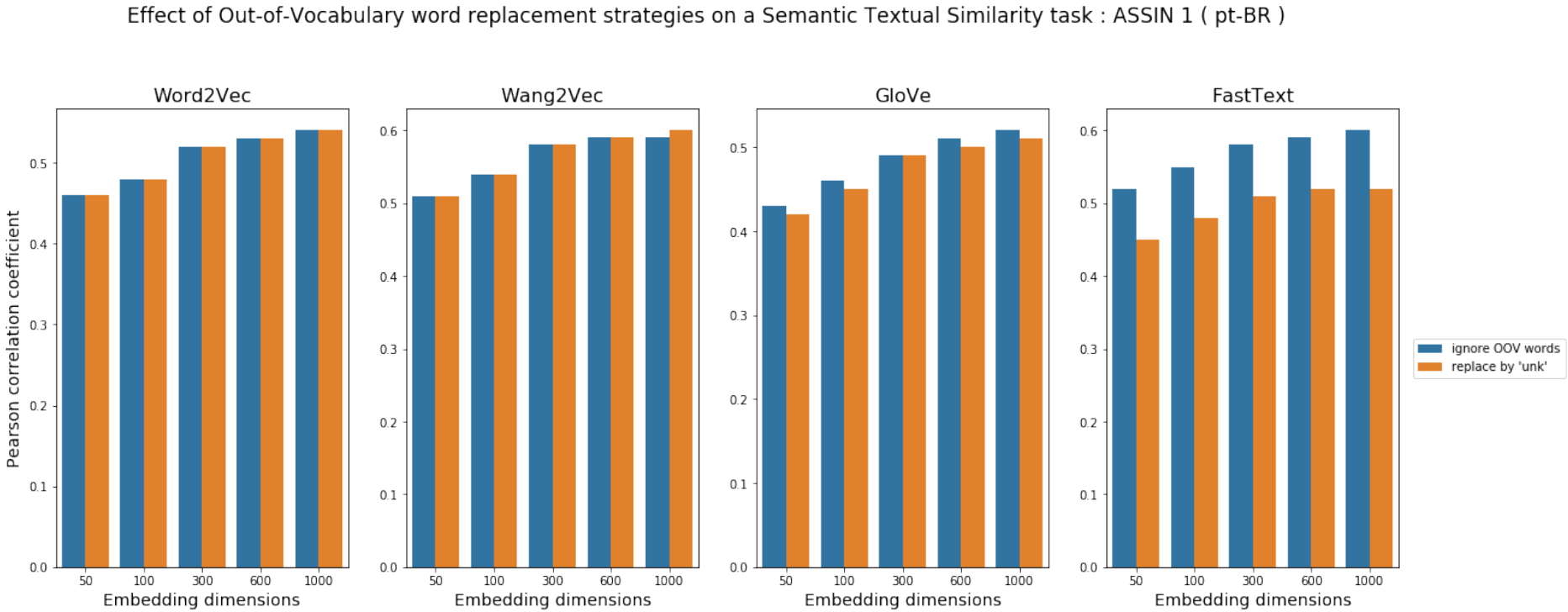
Example: FastText - OOV words replaced by "unk"

- Our approach:
  - **Original phrase**:
    "Votaram contra a proposta 267 deputados, e 210 votaram a favor"
  - **Preprocessing**:
    "votaram", "contra", "proposta", "000", "deputados", "000", "votara", "favor"
  - **Remove unknown words from the phrase**:
    "votaram", "contra", "proposta", "000", "deputados", "000", "votara", "favor"

Example: FastText - nearest neighbors of "unk"

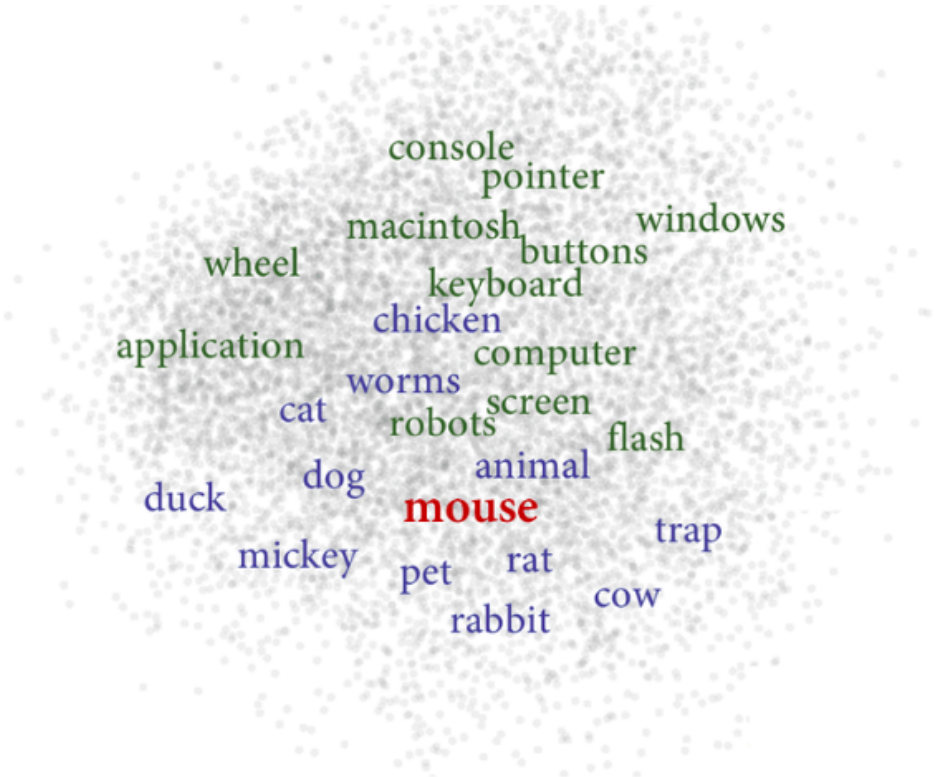


Example


Effect of Out-of-Vocabulary word replacement strategies on a Semantic Textual Similarity task : ASSIN 1 (pt-BR)
Word2Vec
Wang2Vec
GloVe
FastText
Pearson correlation coefficient
Embedding dimensions
50
100
300
600
1000
ignore OOV words
replace by 'unk'

Model-specific issues

- Out-of-vocabulary (OOV) words (Hu et al. [2019])
- Meaning conflation deficiency (Camacho-Collados and Pilehvar [2018])



ELMo (Peters et al. [2018])

- ~~Out-of-vocabulary (OOV) words~~
  - A convolutional neural network generates character-level embeddings.
- ~~Meaning conflation deficiency~~
  - A bidirectional language model generates contextualized word embeddings.

Results: ASSIN

Dataset	Model	Embedding	Architecture	Dimensions	PCC	MSE
ASSIN 1 (pt-BR)	ELMo - wiki (reduced)				**0.62**	**0.47**
	ELMo - wiki (reduced)	word2vec	CBOW	1000	0.62	0.47
	portuguese-BERT				0.53	0.55
	BERT-multilingual				0.51	1.94
ASSIN 1 (pt-PT)	ELMo - wiki (reduced)				0.63	0.73
	ELMo - wiki (reduced)	word2vec	CBOW	1000	**0.64**	**0.73**
	portuguese-BERT				0.53	0.88
	BERT-multilingual				0.52	0.90
ASSIN 2	ELMo - wiki (reduced)				0.57	1.94
	ELMo - wiki (reduced)	word2vec	CBOW	1000	0.59	1.88
	portuguese-BERT				**0.64**	**1.69**
	BERT-multilingual				0.51	1.94

Conclusions

- Our ELMo models have achieved acceptable performance both on named-entity recognition and semantic textual similarity tasks.
- Out-of-vocabulary word replacement strategies should be carefully considered during semantic textual similarity evaluation. (Hu et al. [2019])

- **Source code:**
  - `https://github.com/ruanchaves/elmo`

References I

Camacho-Collados, J. and Pilehvar, M. T. (2018). From word to sense embeddings: A survey on vector representations of meaning. Journal of Artificial Intelligence Research, 63:743–788.

de Castro, P. V. Q., da Silva, N. F. F., and da Silva Soares, A. (2019). Contextual representations and semi-supervised named entity recognition for portuguese language.

Fonseca, E., Santos, L., Criscuolo, M., and Aluísio, S. (2016). Visão geral da avaliação de similaridade semântica e inferência textual. Linguamática, 8(2):3–13.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., and Aluisio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. arXiv preprint arXiv:1708.06025.

Hu, Z., Chen, T., Chang, K.-W., and Sun, Y. (2019). Few-shot representation learning for out-of-vocabulary words.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations.

Real, L., Fonseca, E., and Gonçalo Oliveira, H. (2020). The ASSIN 2 shared task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese. In Proceedings of the ASSIN 2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese, CEUR Workshop Proceedings, page [In this volume]. CEUR-WS.org.

Acknowledgements

DEEP LEARNING
BRASIL

INSTITUTO DE
INFORMÁTICA
UFG

Acknowledgements