Assessing differences between general and medical specific neural word embeddings

Problem

There has been an increasing amount of studies that create numerical representation of text as a way to include unstructured data in machine learning algorithms. These representations, known as neural word embeddings (NWE), has been used in biomedical text analysis in English (1,2), but only few reports can be found for Spanish (3,4). In English, there is evidence that medical specific NWE perform better than general ones when tested in medical text analysis tasks, but there is no baseline reported for Spanish.

Solution

Create NWE trained over several clinical Spanish *corpora* using a variety of optimized state-of-the-art algorithms. Design test datasets to perform intrinsic evaluation, in the form of terms semantic similarity and analogy questions for the clinical Spanish language, as well as extrinsic evaluations in the form of machine learning text classification. Compare the general and medical specific NWE regarding their performance on the intrinsic and extrinsic tasks. Finally, publish trained NWE for developers and researcher community.

Method

Word2vec and fastText embeddings were used from the Spanish Billion Words *Corpus* Project (5) and we trained NWE using biomedical Chilean text using same algorithms. To compare NWEs the intersection vocabulary of all models was calculated. Performance of each model was assessed using analogy and similarity datasets obtained from SNOMED ontology.

Variable and Metrics

The accuracy of each model in semantic and analogy tests will be addressed by using cosine similarity between correct answers from the test datasets and the model’s outputs, counting as right if the word belongs to the k-nearest terms. For classification tasks, models will be compared based on F1-score.

Hypothesis

The NWE trained on medical-specific Spanish *corpus* performs better than a NWE trained on a general-topic Spanish *corpus* when tested using intrinsic and extrinsic medical text related tasks.

Objectives

1. Compare medical-specific and general NWE to assess their intrinsic and extrinsic performance.
2. Deploy currently available embeddings trained on Spanish general texts.
3. Train NWE on medical narratives in Spanish.
4. Select the best performing algorithm on the intrinsic and extrinsic task.

Preliminary Results

We obtained the 3 most similar words for 9 medical terms for each NWE model and plotted in 2D for each word and its neighbors within 1.5 units. Analogy tasks were translated from English, with procedure site analogy showing the best performance, similar to results reported for English.

Outlook

A numerical representation of free-text can impact clinical management and secondary use of information. The lack of available resources for Chilean clinical text represents an enormous opportunity.

References

1. Chen Z, He Z, Liu X, Bian J. Evaluating semantic relations in neural word embeddings with biomedical and general domain knowledge bases. BMC Med Inform Decis Mak [Internet]. 2018;18(Suppl 2). Available from: http://dx.doi.org/10.1186/s12911-018-0630-x

2. Wang Y, Sohn S, Liu S, Shen F, Wang L, Atkinson EJ, et al. A clinical text classification paradigm using weak supervision and deep representation. BMC Med Inform Decis Mak [Internet]. 2019 Dec 7 [cited 2019 Jan 22];19(1):1. Available from: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-018-0723-6

3. Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical Natural Language Processing in languages other than English: opportunities and challenges. J Biomed Semantics [Internet]. 2018 Mar 30 [cited 2018 Aug 23];9(1):12. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29602312

4. Weegar R, Perez A, Casillas A, Oronoz M. Deep Medical Entity Recognition for Swedish and Spanish. IEEE Int Conf Bioinforma Biomed. 2018;1595–601.

5. Cardellino C. Spanish {B}illion {W}ords {C}orpus and {E}mbeddings [Internet]. 2016. Available from: https://crscardellino.github.io/SBWCE/