Semantic Textual Similarity for Spanish Sentences

April 25th, 2017

Fiona Heaney & Matt Zielonko
WPI Class of 2017 - Computer Science

Fiona Heaney
WPI ‘17 - Computer Science
Cimpress - Software Engineer

Matt Zielonko
WPI ‘17 - Computer Science
Mathematical Sciences/Spanish & Latin American Studies Dual Minor

Outline

- Introduction
- Resources
- Methodology
- Results/Findings
- Conclusions/Future Research
- Acknowledgements
- Questions

Introduction

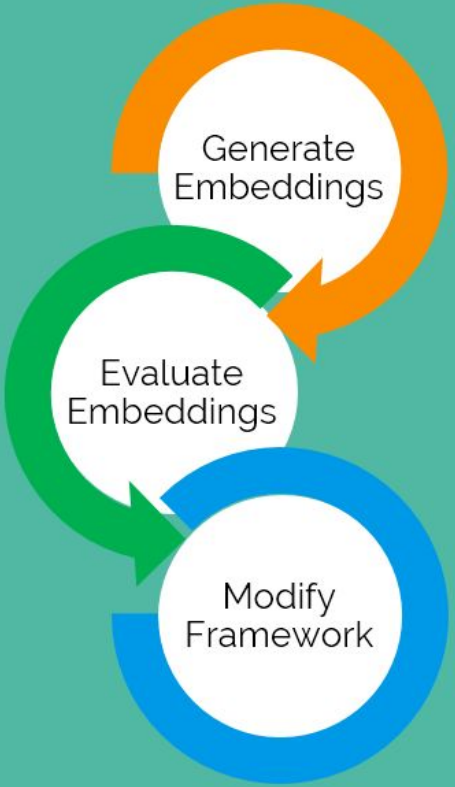
Natural Language Processing

Semantic Textual Similarity (STS)

SemEval Challenges

Our Project

- Locate and generate embedding sets
- Evaluate Spanish embedding performance
- Modify MathLingBudapest framework for English STS to accept Spanish



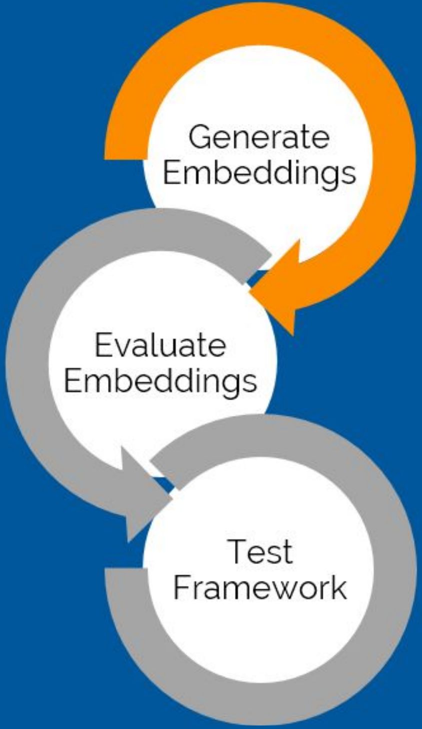
Resources

Corpus:
Spanish Billion Words

- 1.5 Billion words
- Covers:
  - Spanish novels
  - Parliament documents
  - Wikipedia
  - Other Corpora

Stemmer: NLTK Snowball

English Translation	Spanish Word	Spanish Stem
to talk	hablar	habl
we talk	hablamos	habl
zoo	zoológico	zoolog
quickly	rápidamente	rapid



Embedding Sets: Facebook, SBW, and GloVe

- SBW: Set of word2vec embeddings provided by corpus author.

- Facebook: fastText embeddings mined from Wikipedia

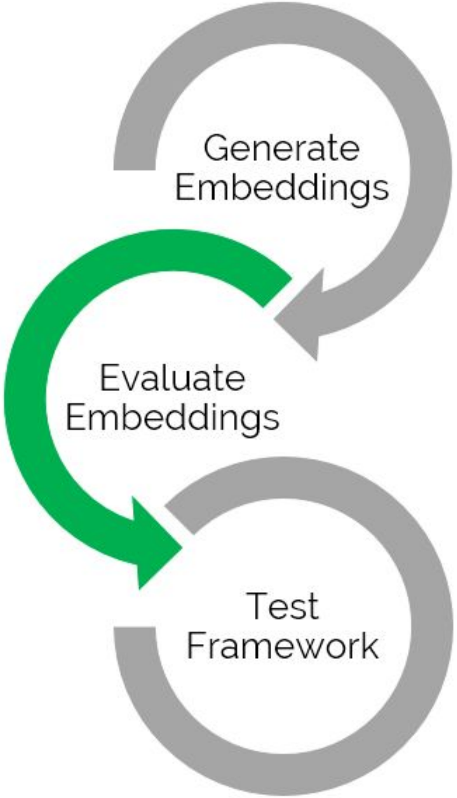
- GloVe: Stemmed SBW corpus passed through GloVe algorithm

Part of Speech Tagger: TreeTagger

- Pretrained model using Spanish Ancora Corpus

Word	The	sky	is	blue	today
POS	DT (determiner)	NN (noun)	VBZ (verb)	JJ (adjective)	NN (noun)

Methodology

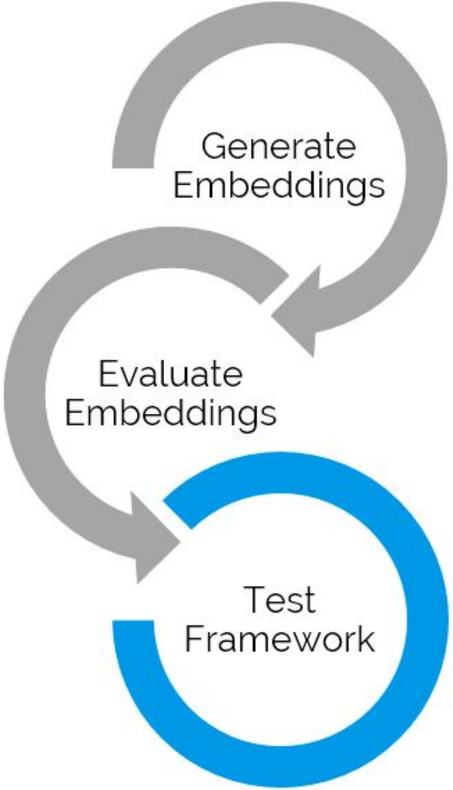


Embedding Evaluation

- Stem corpus and SimLex-999
- Map SimLex-999 word pairs to corresponding vectors
- Compute Cosine similarity

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2} \sqrt{\sum\limits_{i=1}^{n} B_i^2}}$$

- Compute Spearman correlation

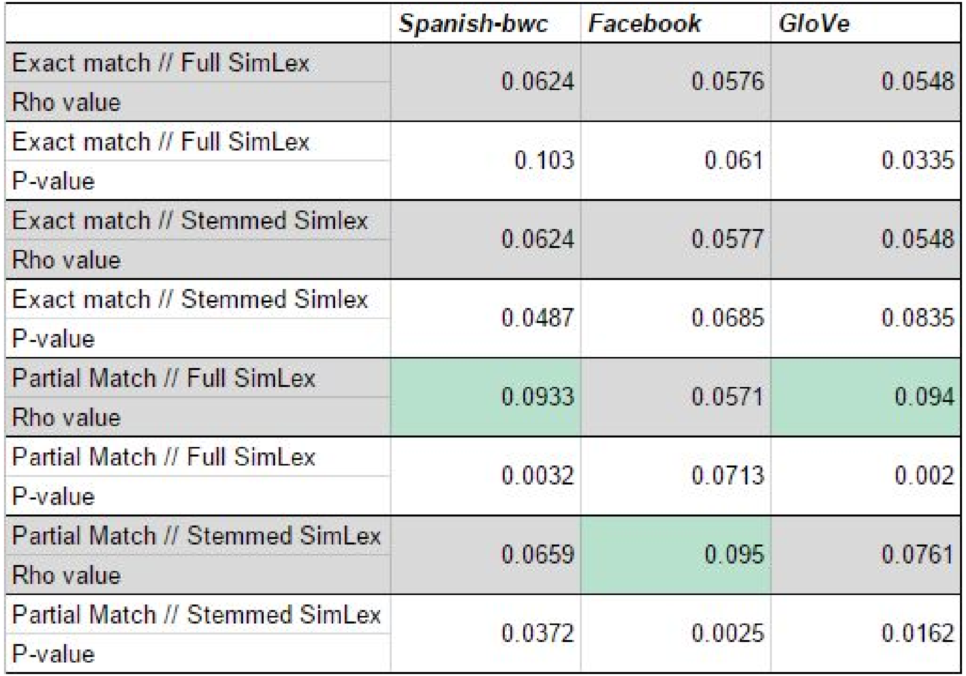


Framework modifications

- TreeTagger
- 2015 SemEval test data
- Hyperparameters tested
    - Modes
    - Stopwords
- Compare to Gold Standard values
    - Spearman Correlation

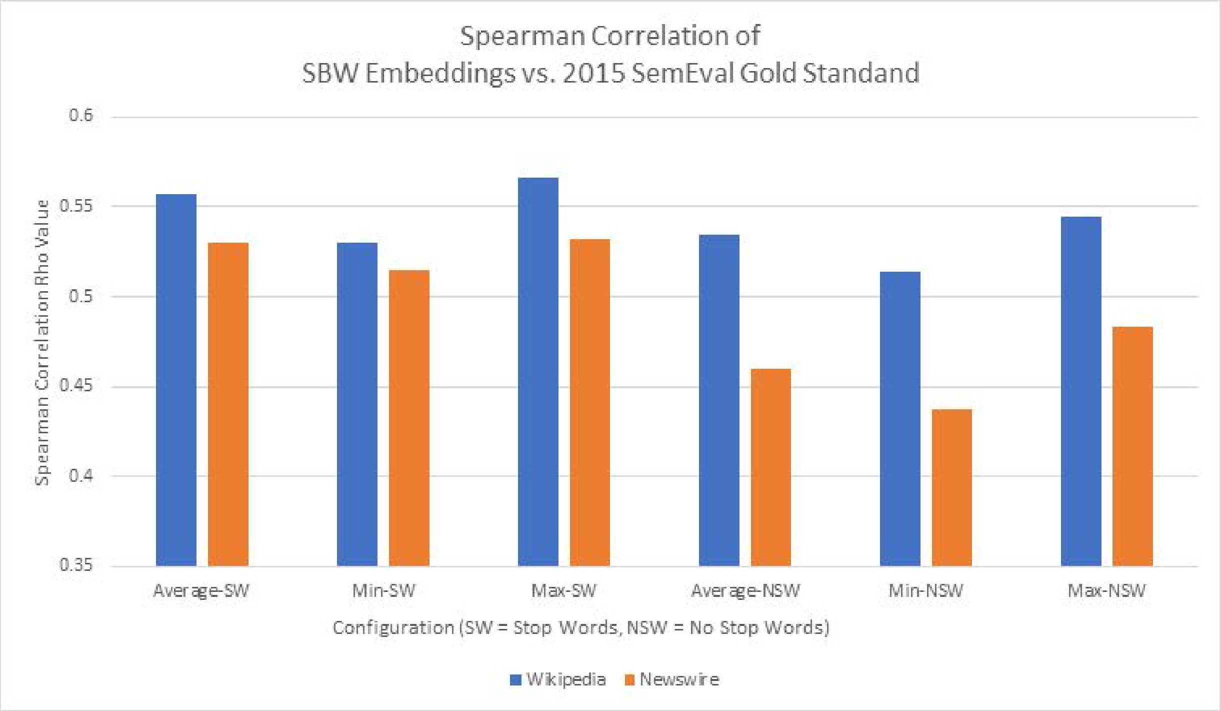
Results

Embeddings

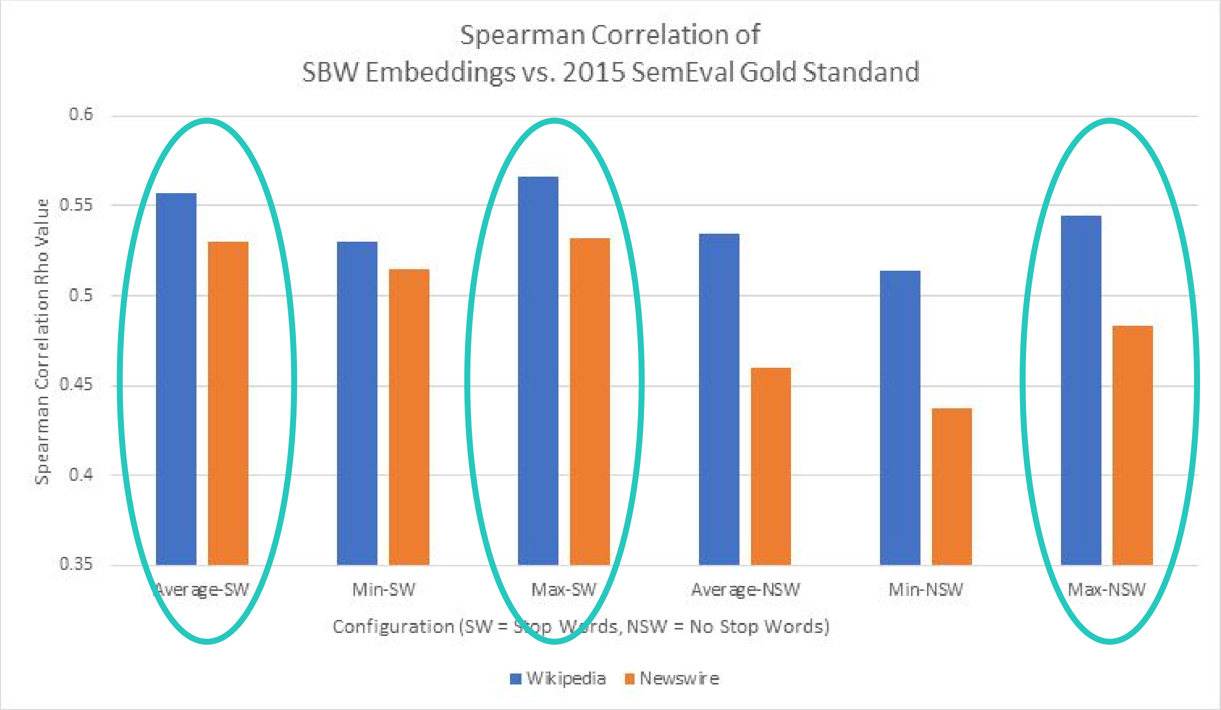


	Spanish-bwc	*Facebook*	*GloVe*
Exact match // Full SimLex Rho value	0.0624	0.0576	0.0548
Exact match // Full SimLex P-value	0.103	0.061	0.0335
Exact match // Stemmed Simlex Rho value	0.0624	0.0577	0.0548
Exact match // Stemmed Simlex P-value	0.0487	0.0685	0.0835
Partial Match // Full SimLex Rho value	0.0933	0.0571	0.094
Partial Match // Full SimLex P-value	0.0032	0.0713	0.002
Partial Match // Stemmed SimLex Rho value	0.0659	0.095	0.0761
Partial Match // Stemmed SimLex P-value	0.0372	0.0025	0.0162

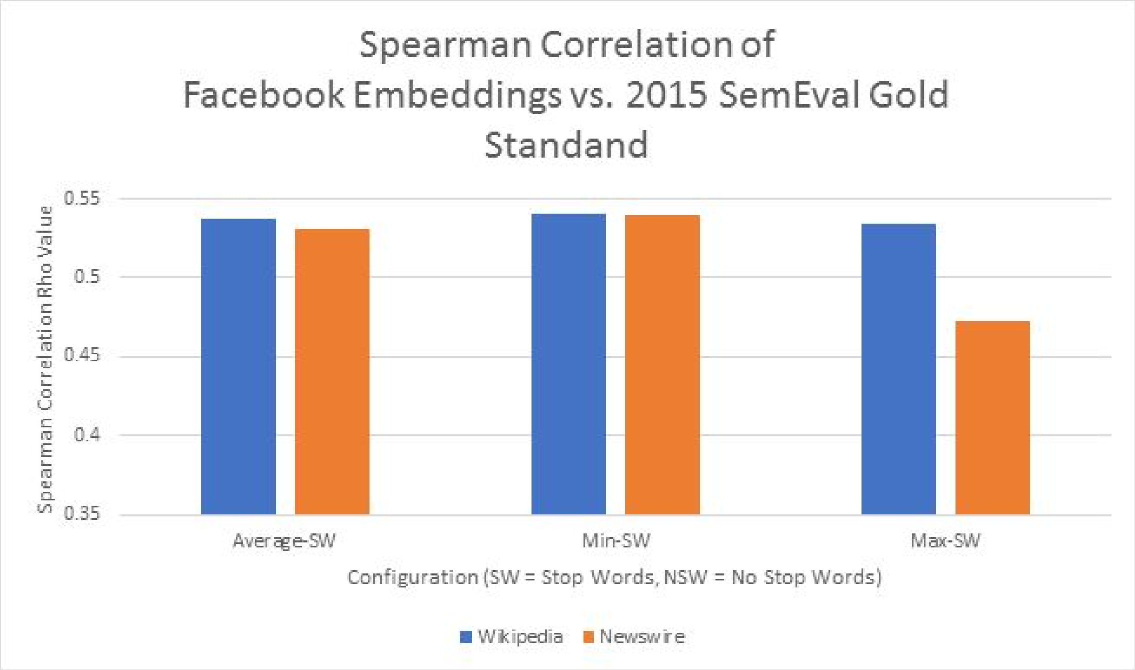
SBW Embedding Performance



SBW Embedding Performance



Facebook Embedding Performance



Conclusions

- Spanish SimLex data for further use
- Statistically satisfactory performance of all three embeddings
- Modified MathLingBudapest framework has satisfactory performance

Future Work

- Enhance Spanish SimLex data set
- Run tests with all 96 permutations of hyperparameters to find optimize configuration
- More in-depth selection of language processing resources

Special Thanks

- András Kornai
- Judit Ács
- Dávid Nemskey
- Gábor Recski
- Gábor Sárközy
- Worcester Polytechnic Institute
- MTA SZTAKI

Questions?

Thank You!