# A Rank-Based Similarity Metric for Word Embeddings

**Enrico Santus, Hongmin Wang, Emmanuele Chersoni, Yue Zhang**

Computer Science and Artificial Intelligence Lab, MIT

University of California Santa Barbara

Aix-Marseille University

Singapore University of Technology and Design

ACL 2018, July 15-20 — Melbourne, Australia

## Motivation

- **Distributional Semantic Models** build *high-dimensional* and *sparse* representations from co-occurrence statistics
- **Semantic similarity** is measured by vector cosine, which treats all features equally
- **Rank-based metrics** have been successfully applied to DSMs, but not yet on *low-dimensional* and *dense* **Word Embeddings**

## *APSynP*: Rank-Based Metric

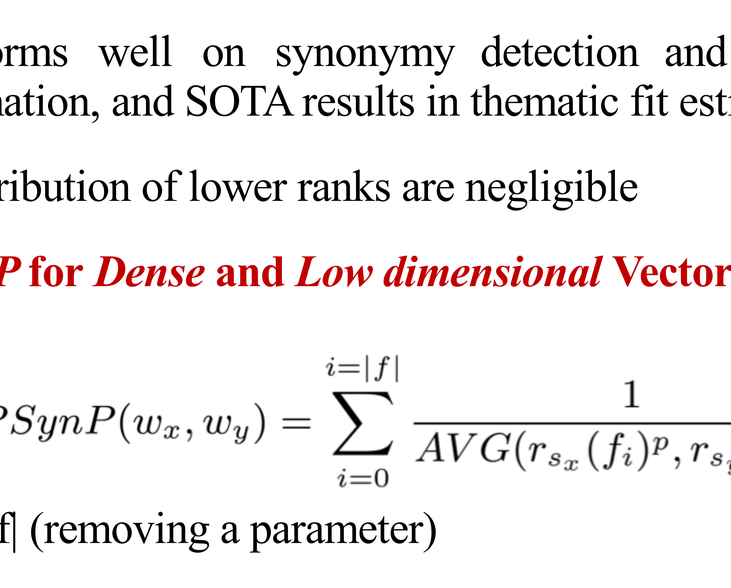
- **Hypotheses:**
  - Similarity = sharing a high number of relevant features
  - Dissimilarity = either non-sharing relevant features or sharing non-relevant features
  - Clustering = cluster members share their salient semantic dimensions to increase cluster cohesiveness
- **APSyn for *Sparse* and *High Dimensional* Vectors**:

$$APSyn(w_x, w_y) = \sum_{i=0}^{i=N} \frac{1}{AVG(r_{s_x}(f_i), r_{s_{s_y}}(f_i))}$$

  - Maps the average feature ranks to a non-linear function, emphasizing the contribution of top-ranked feature
  - Performs well on synonymy detection and similarity estimation, and SOTA results in thematic fit estimation
  - Contribution of lower ranks are negligible
- ***APSynP* for *Dense* and *Low dimensional* Vectors:**

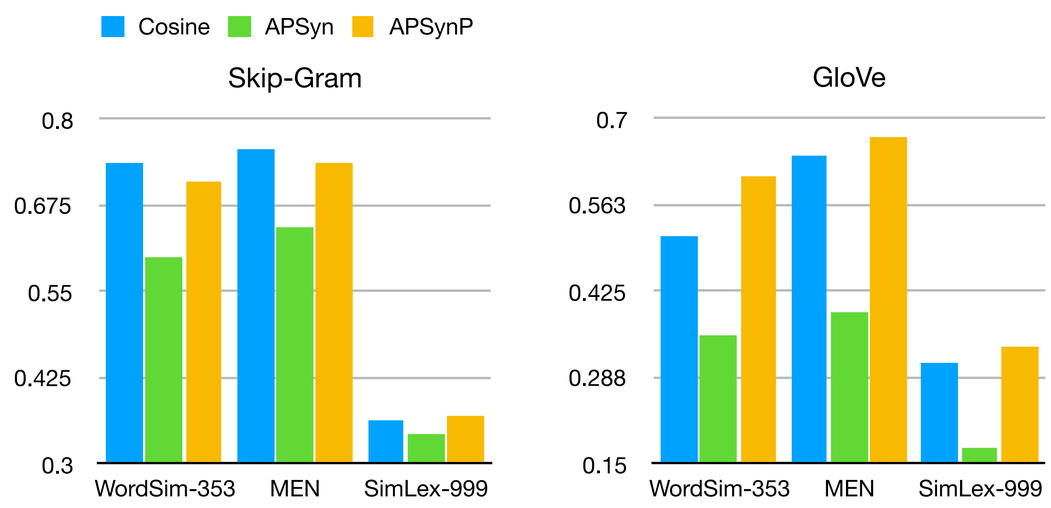
$$APSynP(w_x, w_y) = \sum_{i=0}^{i=|f|} \frac{1}{AVG(r_{s_x}(f_i)^p, r_{s_y}(f_i)^p)}$$

  - N = |f| (removing a parameter)
  - Adding a smoothing parameter, which can be tuned but tends to be constant (*p=0.1*) across all experiments
  - Preserving the non-linear weight allocation across the average feature ranks during the summation
  - Ranks of all features contributing to the final score




## Similarity Estimation

- **Spearman Correlation** between system-generated scores and human judgments
- **Benchmark:** WordSim-353, MEN, SimLex-999




## Outlier Detection

- **Benchmark:** 8-8-8 dataset
  - 64 sets of 8 words + 1 outlier for the evaluation.

$$OPP = \frac{\sum_{W \in D} \frac{OP(W)}{|W|-1}}{D} \times 100$$

$$Accuracy = \frac{\sum_{W \in D} OD(W)}{D} \times 100$$

- **Pairwise Comparisons:** outlier has the lowest average similarity score with the other words in the cluster
- **Cluster Prototype:** outlier has the lowest similarity score with the average vector of the other N-1 words

| | Skip-Gram | | | | GloVe | | | |
|---|---|---|---|---|---|---|---|---|
| | UMBC | | Wiki | | UMBC | | Wiki | |
| | OPP | Acc. | OPP | Acc. | OPP | Acc. | OPP | Acc. |
| $CC-Cos$ | 92.6 | 64.1 | 93.8 | 70.3 | 81.6 | 40.6 | 91.8 | 56.3 |
| **Pairwise** | | | | | | | | |
| $APSyn$ | 93.0 | 67.2 | 94.0 | 68.8 | 78.7 | 40.6 | 89.3 | 53.1 |
| $APSynP$ | **94.0** | **68.8** | **94.5** | **73.4** | **81.8** | **42.2** | **92.8** | **61.0** |
| **Prototype** | | | | | | | | |
| $PT-Cos$ | 93.4 | 65.6 | 93.8 | 68.8 | 80.3 | 40.6 | 90.6 | 54.7 |
| $APSyn$ | 92.6 | 70.3 | 91.0 | 62.5 | 81.6 | 40.6 | 88.7 | 54.7 |
| $APSynP$ | **94.0** | **70.3** | **94.9** | **73.4** | **82.2** | **43.8** | **92.0** | **60.9** |

## Contributions

- ***APSynP***, a rank-based similarity measure adapted with a smoothing parameter for word embeddings
- Setting N=|f| and using a constant parameter makes APSynP unsupervised
- Comparable or better performances than cosine and APSyn on similarity estimation, clustering and outlier detection
- Pilot studies suggest that other rank-based metrics can outperform vector cosine in multiple settings

## Reference

- Jose Camacho-Collados and Roberto Navigli. 2016. Find the Word that Does not Belong: A Framework for an Intrinsic Evaluation ofWord Vector Representations. In Proceedings of ACL Workshop on Evaluating Vector Space Representations for NLP
- Enrico Santus, Emmanuele Chersoni, Alessandro Lenci, Chu-Ren Huang, and Philippe Blache. 2016a. Testing Apsyn against Vector Cosine on Similarity Estimation. In Proceedings of PACLIC.

ACL 2018, July 15-20 — Melbourne, Australia