RatVec: A General Approach for Low-dimensional Distributed Vector Representations via Rational Kernels

Eduardo Brito       Bogdan Georgiev
Daniel Domingo-Fernández  Charles Tapley Hoyt
     Christian Bauckhage

Fraunhofer Center for Machine Learning, Germany
eduardo.alfredo.brito.chacon@iais.fraunhofer.de

1st October 2019
Motivation

Limitations of "classical" word embedding approaches

- Implicit similarity/relatedness concept only derived from co-occurrence (distributional hypothesis)
- We may force artificially our data to be sequential
  - Graph node embeddings for link prediction via random walks
Motivation

Limitations of "classical" word embedding approaches

- Implicit similarity/relatedness concept only derived from co-occurrence (distributional hypothesis)
- We may force artificially our data to be sequential
  - Graph node embeddings for link prediction via random walks

Main idea

1. Apply kernel PCA on a representative subset of the dataset
   - A suitable similarity function (rational kernel) substitutes the dot product.
   - Projection matrix from eigendecomposition of kernel matrix

2. Derive full computing similarity with representative vocabulary and using the projection matrix

3. Solve particular task with simple algorithms (e.g. $k$-nearest-neighbors).
Dutch Spelling Correction

Idea

1. Pick a suitable string similarity (e.g. n-gram similarity)
2. Generate a vector representation for each word in the vocabulary
3. Generate a vector for detected misspelling
4. Pick closest word in vocabulary from precomputed representations (1-nearest neighbor classification)

Evaluation

CLIN28 shared task on spelling correction in Dutch wikipedia texts

Our team achieved best F1 score among the participants
Dutch SpellingCorrection

Idea

1. Pick a suitable string similarity (e.g. n-gram similarity)
2. Generate a vector representation for each word in the vocabulary
3. Generate a vector for detected misspelling
4. Pick closest word in vocabulary from precomputed representations (1-nearest neighbor classification)

Evaluation

- CLIN28 shared task on spelling correction in Dutch wikipedia texts
- Our team achieved best F1 score among the participants
Dutch Spelling Correction
Idea

- Protein with similar aminoacid sequences belong to the same protein family
- Proteins can be modeled as aminoacid sequences (i.e. strings)
- String similarities applicable
Protein Family Classification

Idea
- Protein with similar aminoacid sequences belong to the same protein family
- Proteins can be modeled as aminoacid sequences (i.e. strings)
- String similarities applicable

Evaluation
- Binary classification on balanced datasets extracted from Swiss-Prot
- Weighted average accuracy: 0.93
Protein Family Classification

Visualizing the "7tm_6" and "ATE_C" protein families

- 7tm_6
- ATE_C
Conclusion

Main Advantages
- Framework for general (non-numeric) entities including text, proteins, and DNA
- Based on explicit similarity functions (rational kernels)
- Resource-aware (representative vocabulary size is adjustable)

Future Work
- Learning optimal similarity metric
- Determine optimal representative vocabulary
- Other uses cases: information retrieval, chemical interaction prediction
Thank you!

Check out our code (available as a PyPi package!):

🔗 https://github.com/ratvec/ratvec

Questions?
✉️ eduardo.alfredo.brito.chacon@iais.fraunhofer.de