Semantic Similarities using Classical Embeddings in Quantum NLP

Damir Cavar, Chi Zhang

Indiana University at Bloomington - NLP-Lab

Word Embeddings in Natural Language Processing

Distributional Semantics and Vector Models:

- Word and text meaning encoded in dense vectors:
  - **fastText** (Bojanowski et al., 2017; Joulin et al., 2018)
  - **GloVe** (Pennington et al., 2014)
  - **Numberbatch** (Speer et al., 2017)
  - **BERT** (Devlin et al., 2019)
  - **Large Language Models and Generative AI** (byte-pair encoding)
- Generating word embeddings and language models:
  - Costly and time-consuming computation
  - Large training and evaluation data sets
  - Many pre-computed models freely available
- Use-cases, for example:
  - Generic neural or probabilistic NLP methods for text classification, machine translation, ...
  - Lexical- or text-similarity computation in semantic search

Quantum NLP Questions:

- Can classical embeddings and language models be used in QC for Q-NLP/AI/ML?
- How reliable are encoding approaches for mapping classical to quantum embeddings?
- Is there information loss or deterioration of quality in different mapping approaches?

Our Goals

- Identify reliable mapping approaches from classical to quantum embeddings.
- Compare the similarity metric in classical with quantum similarity scores.

- **Mapping Algorithms:** Amplitude Encoding, Basis Encoding, Angle Encoding...
- **Similarity Measures:** SWAP test, Matrix Distances for Quantum Circuits (Frobenius Norm Distance, Symmetrized Frobenius Norm Distance, Minimalized Frobenius Norm Distance, Eigenvalue Distance, Symmetrized Eigenvalue Distance)

Quantum Word Similarities

- Classical embeddings → Quantum embeddings
  - **fastText**, 300-dimensional word vectors, 2.5 mil. words
  - **GloVe**, 840 billion tokens, 300-dimensional word vectors, 2.1 mil. words
  - **Numberbatch**, 300-dimensional vectors, 516,783 words
  - **BERT**, 768-dimensional word vectors
  - **OpenAI GPT Embeddings**, large 3072-dim. and short 1536-dimensional word vectors
- Amplitude Encoding
- **SWAP Test** (Buhrman et al., 2001)
  - Two circuits S and T with the same number of qubits
  - Measures the difference between S and T
  - Given qubits in S as $s_0, s_1, \ldots, s_{n-1}$ and qubits in T as $t_0, t_1, \ldots, t_{n-1}$ and an ancillary qubit q_0:
    - perform first the Hadamard gate, then
    - the controlled SWAP gate from q_0 to s_i and $t_i, i = 0, \ldots, n-1$ and
    - again the Hadamard gate
  - Measures the value of q_0

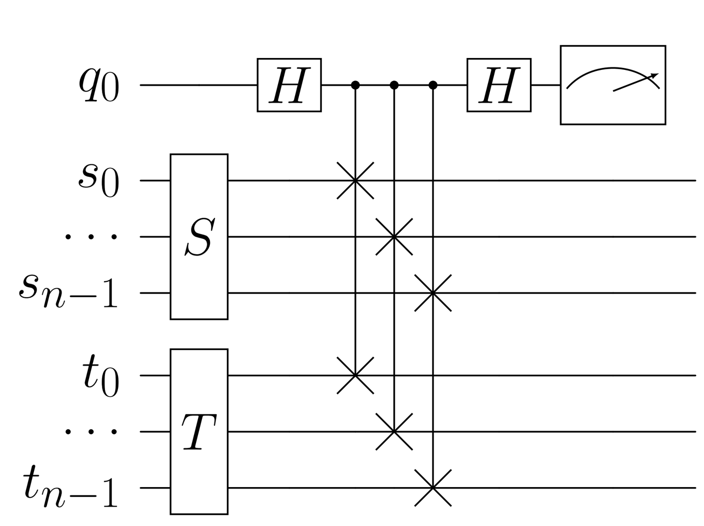

Figure 1. Quantum circuit for the SWAP test between two circuits S and T

Data Selection

- Language: English
- No function words
- Only nouns filtered by NLP pre-processing
- Random selection of 100 words in the vector models
- Manual selection of 100 words covering 20 different topics or semantic fields:
  - *teacher professor student school university college*
  - *game sport ball team player*
  - *computer laptop tablet phone*
  - *music song band singer guitar piano*
  - *movie actor actress director producer*
- Generation of all possible word pairs for each of the two sets
- Computation of Cosine Similarity for each word pair
- Environment: `qiskit` and `qasm_simulator`

Issues

- Computational complexity of conversion classical to quantum embeddings.
- Costly computation of similarity scores.
- Word embedding models are not cleaned, contain ambiguities, and contain non-lexical data.

Conclusion

- **Result: classical vector similarity using Cosine Similarity** and quantum embedding similarity using Quantum similarities
  - **Correlation Coefficient for 4400-word pairs approx. 0.90 on average for the pre-computed vector models using the `qasm_simulator`**
- **There is minimal information loss in the encoding process.**
- **Classical word embedding models can be used in Q-NLP/ML/AI tasks.**

Availability

- **Data and Code available:** GitHub repo, URI TBA in final poster version

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. ISSN 2307-387X.

Harry Buhrman, Richard Cleve, John Watrous, and Ronald De Wolf. Quantum Fingerprinting. *Physical Review Letters*, 87(16):167902, September 2001. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.87.167902.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein et al., editor, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, June 2019.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL `http://www.aclweb.org/anthology/D14-1162`.

Robyn Speer, Joshua Chin, and Catherine Havasi. ConceptNet 5.5: An open multilingual graph of general knowledge. pages 4444–4451, 2017. URL `http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972`.

Natural Language Processing Lab

The NLP-Lab (`https://nlp-lab.org/quantumnlp/`):