The Natural Language Qu Kit - NLQK for Quantum NLP and AI

Damir Cavar, Billy Dickson, James Bryan Graves, Shane A. Sparks, Koushik Reddy Parukola

Indiana University at Bloomington - NLP-Lab

Quantum Natural Language Processing

- Classical NLP Computing Environments
  - **Corpora:** dictionaries, text collections, NLP-annotated data
  - **Embeddings:** word and token-based vector models based on Distributional Semantics
  - **Language Models:** BERT (Devlin et al., 2019)
  - **Large Language Models and Generative AI:** OpenAI, Anthropic, VoyageAI, ...
  - **Libraries:** NLTK Bird and Loper (2004), spaCy, transformer, Pytorch, ...
- Quantum NLP Computing Environments
  - **Corpora:** ???
  - **Embeddings:** ???
  - **Language Models:** ???
  - **Large Language Models and Generative AI:** ???
  - **Libraries:** lambeq
  - **Generic Libraries - not NLP specific:** Qiskit Javadi-Abhari et al. (2024), Pennylane Bergholm et al. (2022), Cirq, ...
- Hybrid classical and quantum NLP environments are necessary, but:
  - Specification of Data formats, exchange, and sharing standards
  - Identification of Optimal encoding approaches
  - Hybrid algorithm specification
  - etc.
- Use-cases, for example:
  - Research and Experimental
  - Solutions Engineering
  - Education, Teaching, Training

General Criteria:

- Easy installation and use
- Multi-platform support (e.g., MacOS, Linux, Windows)
- State-of-the-art performance:
  - comparable to spaCy or Pytorch
  - connectivity to CuPy, CUDA and CUDA-Q, and common hardware providers

Our Goals

- Learning from excellent examples:
  - spaCy (`https://spacy.io/`)
  - Natural Language Toolkit (NLTK) (`https://nltk.org/`)
  - CuPy (`https://cupy.dev/`) Okuta et al. (2017)
  - CUDA and CUDA-Q (`https://developer.nvidia.com/cuda-q`)
  - ...

- **Mapping Algorithms for Embeddings:** Amplitude Encoding, Basis Encoding, Angle Encoding...
- **Similarity Measures:** SWAP test, Matrix Distances for Quantum Circuits (Frobenius Norm Distance, Symmetrized Frobenius Norm Distance, Minimized Frobenius Norm Distance, Eigenvalue Distance, Symmetrized Eigenvalue Distance)
- **Classical to Quantum Conversion:** Real-vectors to Complex-vectors conversion, Quantum Computing compatible language models and embeddings, etc.

- **Data Sets:** Similar to NLTK data (`https://www.nltk.org/nltk_data/`)
  - **Word 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
- Dictionaries
  - e.g., **SimLex-999**
- Models
  - e.g., **Language Models** and Complex-vector models
- Rich Documentation and Examples

Current Environment

- Core data sets:
  - **Wordlists:** SimLex-999, nouns, pairwise similarities
  - **Embeddings:** OpenAI and VoyageAI embeddings for all words
  - **Hamiltonians:** Quantum States for words and text stored as Hamiltonians
  - **Qauntum states:** word and text encodings as amplitudes (using complex numbers)
- Core functions:
  - Linear algebra functions
  - Functions for automatic data download and installation
  - Quantum embedding functions, e.g., complex embeddings, optimization of embeddings for classical and quantum environments
- Integration
  - Python >= 3.9
  - *Dependencies:* CuPy, CUDA and CUDA-Q, RAPIDS, Qiskit

Implementation

- NumPy or CuPy – automatically selected
- Full CPU and GPU support
- Nvidia CUDA and CUDA-Q integrated
- Currently, IBM Quantum and AWS Braket integration, expanding to other platforms
- Interaction with SOTA AI models:
  - OpenAI API
  - Anthropic API
  - VoyageAI API

Availability

- Data and Code available:
  - GitHub repo: `https://github.com/dcavar/nlqk`
  - PyPi module: `https://pypi.org/project/nlqk/`
  - Website: `https://nlqk.ai/`
- Documentation:
  - `https://nlqk.ai/documentation/nlqk.html`

Installation:

pip install nlqk

References

Ville Bergholm et al. Pennylane: Automatic differentiation of hybrid quantum-classical computations, 2022. URL `https://arxiv.org/abs/1811.04968`.

Steven Bird and Edward Loper. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL `https://aclanthology.org/P04-3031/`.

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.

Ali Javadi-Abhari, Matthew Treinish, Kevin Krsulich, Christopher J. Wood, Jake Lishman, Julien Gacon, Simon Martiel, Paul D. Nation, Lev S. Bishop, Andrew W. Cross, Blake R. Johnson, and Jay M. Gambetta. Quantum computing with Qiskit, 2024.

Ryosuke Okuta, Yuya Unno, Daisuke Nishino, Shohei Hido, and Crissman Loomis. Cupy: A numpy-compatible library for nvidia gpu calculations. In *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS)*, 2017. URL `http://learningsys.org/nips17/assets/papers/paper_16.pdf`.

Natural Language Processing Lab

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

https://nlqk.ai/ IEEE Quantum Week 2025 dcavar@iu.edu