Semantic Graph Generation from a Large Document Corpus

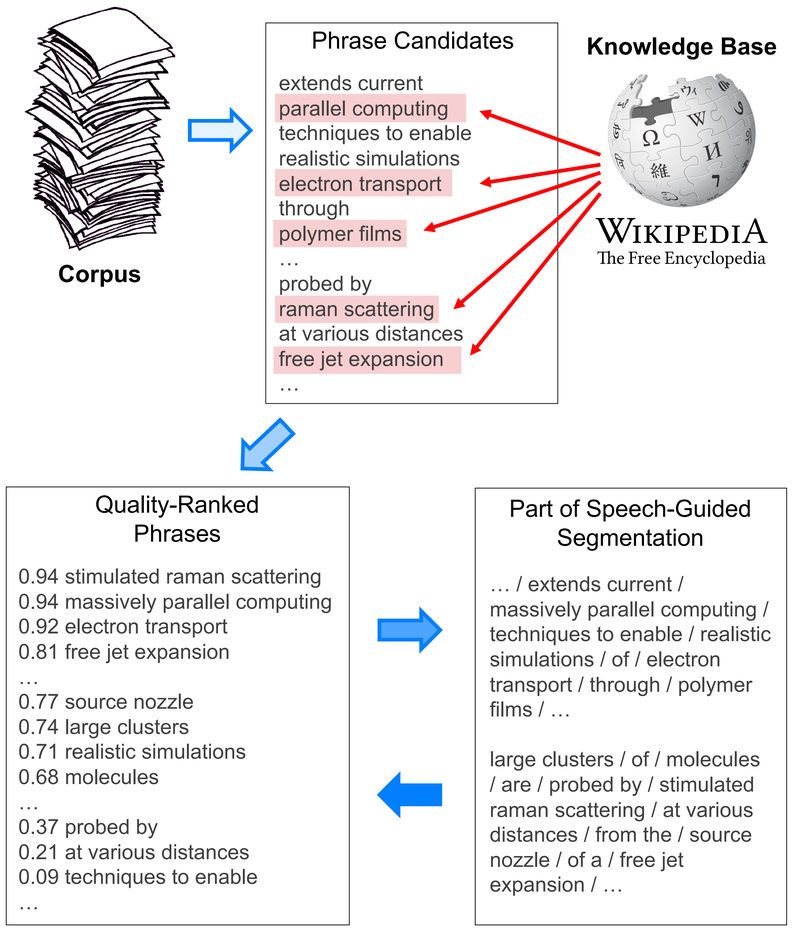
Anna Jurgensen (LLNL), David Buttler (LLNL)

Lawrence Livermore National Laboratory

Manual extraction and organization of information from a large volume of documents is a slow and expensive process. An approach that can extract and synthesize information with minimal user input will allow for more efficient use of large document collections.

Automated Process Extracts Entities from Text

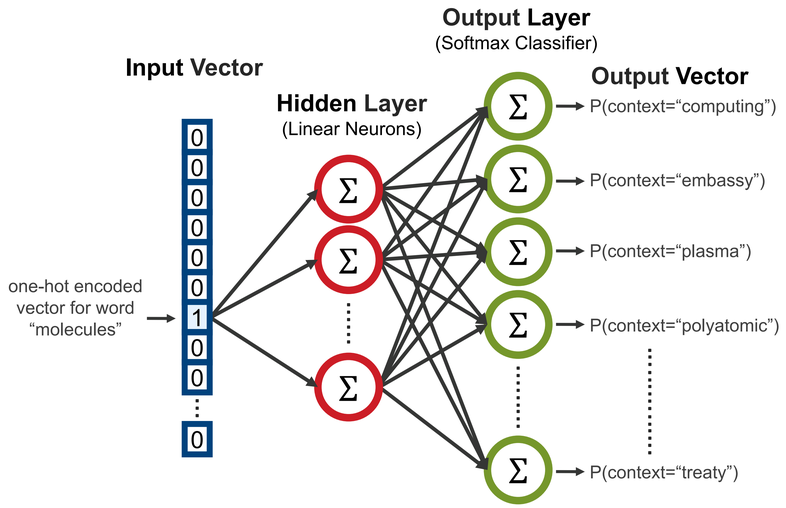
An automated process using part-of-speech tagging and cross-referencing of a knowledge base finds single- and multi-word 'phrases' in a corpus that are most likely to refer to a person, location, organization, or informative object or concept.



entity identification process, adapted from AutoPhrase [1] [2]

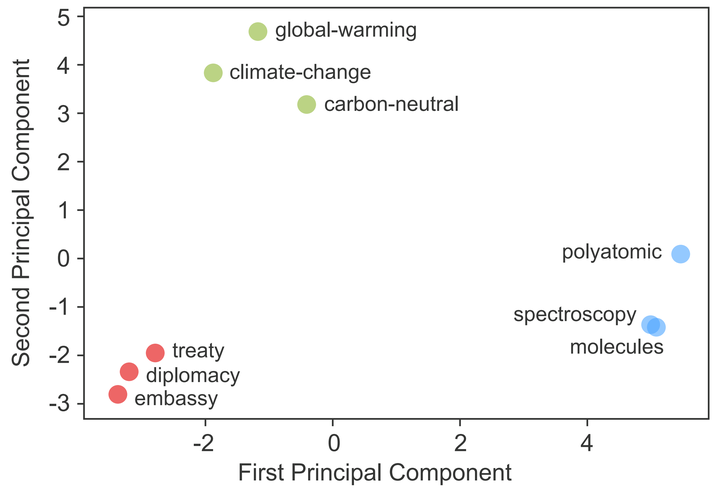
Related Information Found by Comparing Text Embeddings

Transforming words and phrases into vectors that represent their context are combined to represent sentences and larger units of text.



skip-gram architecture of word2vec [3]

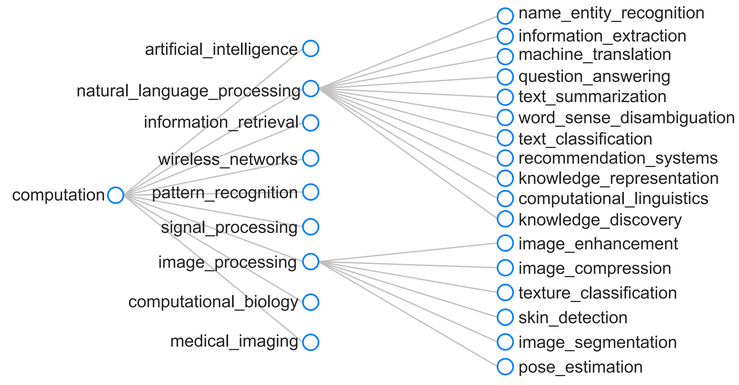
Vectors are compared to find related terms as well as sentences with similar information.



example of Principal Component Analysis of word vectors exhibiting topical clustering

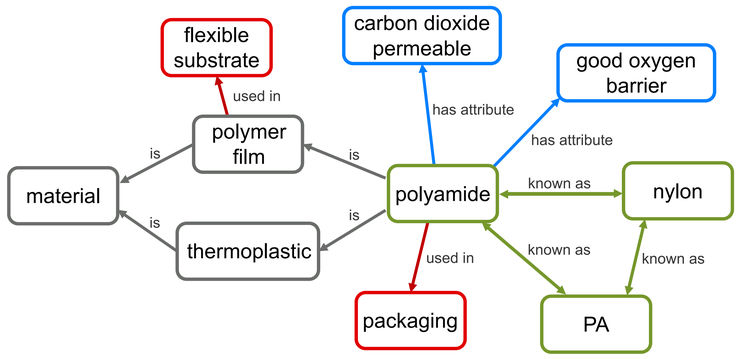
Entities Linked by Extracted Relationships

After determining physical and conceptual entities, a taxonomy is computed to classify the entities with respect to each other.



example of a generated taxonomy, adapted from HiExpan [4]

More complex relationships between entities are then found and used to construct a semantic graph.



demonstration of a semantic graph manually constructed from text embeddings, Probase [5] and AutoPhrase [1] [2] output

REFERENCES

[1] J. Shang, J. Liu, M. Jiang, X. Ren, C. R. Voss, J. Han "Automated Phrase Mining from Massive Text Corpora" in IEEE Transactions on Knowledge and Data Engineering, 2018.
[2] J. Liu, J. Shang, C. Wany, X. Ren, J. Han, "Mining Quality Phrases from Massive Text Corpora", Proc. of ACM SIGMOD, 2015.
[3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, "Distributed representations of words and phrases and their compositionality", Advances in Neural Information Processing Systems, 2013.
[4] J. Shen, Z. Wu, D. Lei, C. Zhang, X. Ren, M. T. Vanni, B. M. Sadler, J. Han, "HiExpan: Task-Guided Taxonomy Construction by Hierarchical Tree Expansion", Proc. ACM SIGKDD, 2018.
[5] W. Wu, H. Li, H. Wang, K. Q. Zhu, "Probase: Probabilistic Taxonomy for Text Understanding", Proc. of ACM SIGMOD, 2012.

LLNL-POSTER-XXXXXX
This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.