Envisioning Corpora:

Root and Rule Infrastructure for Semantic Web and Topic Modeling

Peter Zhang and Dr. Talapady N. Bhat

NIST, Material Measurements Laboratory, Biosystems and Biomaterials Division

Motivation

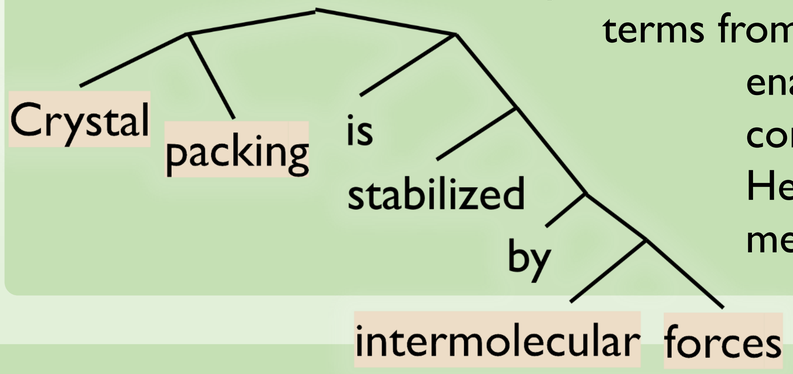
Over 2.5 million papers were published in 2018, a rate that increases each year. Analyses of groups of documents, or *corpora*, can reveal trends, relations, and themes.

Machine learning can achieve accuracy on a specific task, but fail to be effective across domains First, training a new model requires an unsustainable amount of manually generated training data. Second, machine learning struggles to deal overlapping terminology, such as the use of "cell" in "cell biology" and "battery cell."

Background

We take advantage of a novel natural language processing approach called root-and-rule (R&R). R&R is inspired by the Sanskrit method for constructing sentences, which starts from root terms and follows preset rules to build up to sentences. By conceptualizing English sentences in this format, we have developed a process to extract R&R terms from sentences.

Previously, we automated the process of extracting R&R terms from a PDFs of scientific articles. This enabled us to generate several R&R corpora of different scientific topics. Here, we develop and apply analytic methods on a corpus-level.



Methods

Semantic Web is a web of objects connected via relations. It relies on the RDF data format, which stores information in subject-object-predicate triplets. If we consider the R&R terms as potential subjects and objects, then our description of a document could be greatly enhanced by identifying the predicates that connect them.

We design a program that automates the search for predicate phrases. First, the program references Python NLTK library and Princeton's WordNet dictionary to identify verbforms. Then, it references dictionaries of prepositions and conjugations to collate neighboring phrases. Finally, the program builds a triplet with nearby R&R terms.

To conduct topic modeling, we compile four corpora (R&R, unigram, bigram, and trigram) and preprocesses with TF-IDF. We deploy Latent Dirichlet Allocation, which assumes documents are "mixes" of topics and adjusts topics to better explain the corpus. First, we measure the UCI coherence score, which denotes how well the terms within a topic cohere with one another. Second, we measure the perplexity, which denotes how well the topics account for the corpus.

"Crystal packing is stabilized by intermolecular forces."

Tag Verbs

"stabilized"

Collate

"is stabilized by"

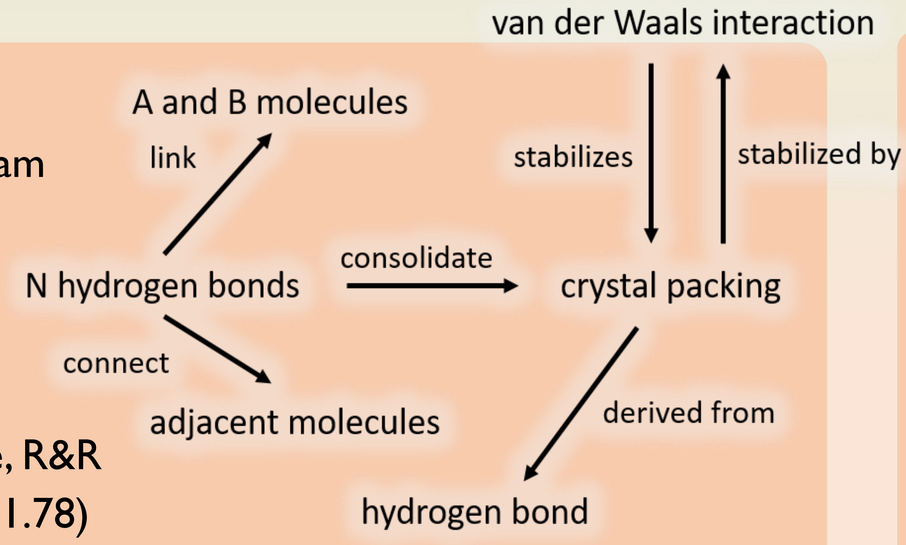
Search

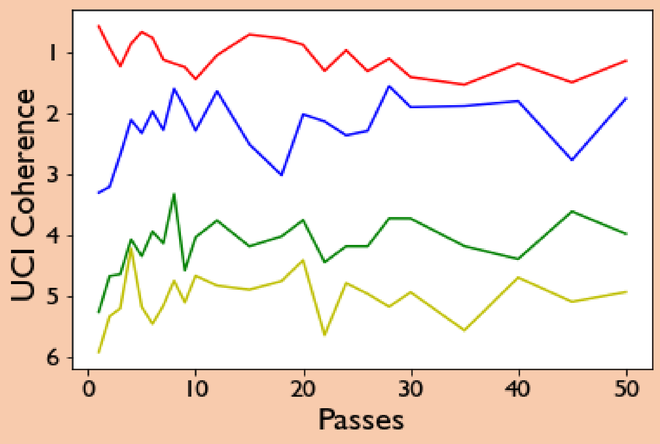
"crystal packing"
"is stabilized by"
"intermolecular forces"

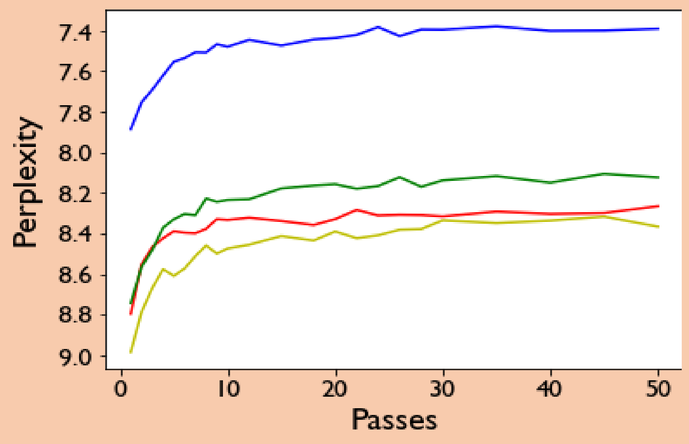
Results

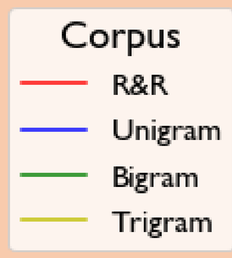
In less than a minute, the Semantic Web program generated 5585 triplets from a corpus of 1233 crystallography papers. These triplets formed a compact and meaningful semantic web.

We performed LDA training on the same corpus with seven topics. As the models converge, R&R achieves the best UCI coherence score (-1.13 vs. -1.78) and competitive perplexity (-8.27 vs. -8.37).









Conclusion

Machine learning approaches to natural language processing can be effective but struggle to remain applicable across topic domains. R&R can resolve these gaps by providing a generalized framework for the isolation of meaningful terms. Using R&R terms, we can automate the identification of subject-object-predicate triplets and the construction of Semantic Webs. Topic modeling with R&R terms can achieve improvements in complexity and perplexity, while also incorporating a richer diversity of terms. Future work should test these methods on different scientific fields.

References

Bhat, T.N., Bartolo, L.M., Kattner, U.R. et al. JOM (2015) 67: 1866. https://doi.org/10.1007/s11837-015-1487-4

Collard, J. , Bhat, T. N., et al. Washington (2018) 104: 31.

National Institute of Standards and Technology
U.S. Department of Commerce

SCIENCE • MATHEMATICS • COMPUTER SCIENCE • MAGNET PROGRAM • Montgomery Blair High School