**Envisioning Corpora: Root & Rule Infrastructure for Semantic Web and Topic Modeling**

Peter Zhang and Dr. T. N. Bhat

NIST Material Measurement Laboratory

**Introduction**

Approximately 2.5 million new scientific papers were published in 2018, a rate that increases each year. The sheer volume of papers motivates analysis on groups of these documents, or *corpora*. Corpus-level analyses can reveal trends, relations, and themes.

Recently, NASA released a “tagging service” trained on a corpus of 3.5 million manually tagged documents. While machine learning can achieve great accuracy in one domain, it is limited as a general tool. First, training a model on a new domain requires a great deal of training data. At this scale, manually generating data is unsustainable. Second, these methods struggle to deal with terminology that overlaps across multiple domains. For example, a model may struggle to differentiate the use of “cell” in “cell biology” and “battery cell.”

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.

*R&R deconstruction of an English sentence*

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.

**Method**

Semantic web denotes a web of objects connected via relations. It is premised on the RDF format, which stores information in subject-object-predicate triplets. If we consider the R&R terms as potential subjects and objects, then we can inquiry about the *predicates* that connect them. We devise a method to extract triplets from a given document processed by R&R terms.

*Automated extraction of triplets*

Describe…

Our second interest is in topic modeling. Given a corpus, we want to be able to infer the topics that comprise the documents, where a topic is a list of terms. The standard approach is Latent Dirichlet Allocation (LDA), which assumes that documents are “mixes” of topics are designs topics to better account for the documents. However, LDA is limited in that it requires the user to pre-specify the number of topics. We design a modified algorithm that infers the number of topics as well, called *Cyclic-LDA*.

*Cyclic-LDA*

We deploy two measures to evaluate our models. First, we examine the UCI coherence score. Coherence denotes how well the terms within a topic cohere with each other and is estimated as the likelihood of the terms co-appearing in a random document. Second, we examine the perplexity. Perplexity denotes how well the topics account for the corpus and is calculated as the ability for the model to predict a set of unseen documents.

*Equation*

To test the efficacy of the R&R topic model, we compare it to tests deployed on three other corpora. We train a model on a corpus of unigrams comprised of lemmatized nouns. We also train models on corpora of bigrams and trigrams, which are groups of collated unigrams.

**Results**

We evaluated our triplet builder on a corpus of 1233 crystallography papers. The program ran in [X] seconds and generated 6578 triplets. Below is an example of the Semantic Web generated by the program.

*Generated Semantic Web*

We observe that the generated Web successfully highlights meaningful and useful relations.

The topic models trained in an average of [X] per pass. After averaging the five trials, we arrived at the following results.

*Coherence graph*

*Perplexity graph*

**Conclusion**

Machine learning approaches to natural language processing can be effective but struggle to remain applicable across 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.