A summary of the paper - Nanomaterial Synthesis Insights from Machine Learning of Scientific Articles by Extracting, Structuring, and Visualizing Knowledge by Hiszpanski et al

Parth Shah

October 15, 2020

Due to their broad scope of applications, Nanomaterials are of interest to many researchers in the field of material sciences. However, the synthesis and creation of nanomaterials are often time-consuming and Edisonian Processes. Also, for them, knowledge accumulation and discovery is through tediously reading and digesting scientific literature.

To facilitate these problems, authors of the paper developed scientific article-processing tools that extract and structure information from the text and figures of nanomaterials articles. This enables the creation of a mineable personalized knowledgebase for nanomaterials synthesis.

The authors draw inspiration from research performed at IBM, Olivetti's group and Ceder & Jain's group. Through literature parsing, these researchers were successfully able to gain insights in material synthesis for applications like cooking recipes, metal oxide synthesis and synthetic inorganic recipes respectively.

In particular, the authors build Machine Learning Models to classify articles according to the nanomaterial composition and morphology, extract synthesis protocols from within the articles' text, and perform chemical entity recognition within synthesis protocols.

The authors create a corpus of 35k unique paper's from Elsevier's catalogue journals. They searched through the full text of articles and select those that contained the terms "X nanoY" and "synthesis", where X = nanomaterial compositions of interest and Y = the nanomaterial morphology of interest.

To identify the composition and morphology from the text, the authors implemented an unsupervised classification algorithm based on the TF-IDF statistic for each word in each article. The statistic determines the relevance of a word in direct relation to its frequency. The composition and morphology terms with the highest TF-IDF weights were assigned as the topical label for each document. Even though a topic modelling approach would be more

robust, the authors were able to get sufficient accuracy from their method. A set of 99 papers hand-labelled with the appropriate nanomaterial composition and morphology was used as the gold standard for evaluation, on which the unsupervised model yielded 100% accuracy on composition prediction and 95% accuracy on morphology prediction. Observing the combinations of composition and morphology can be used to identify hot/trending topics and areas ripe for exploration just from the number of times they appear comparatively.

To identify synthesis protocols, the authors trained a logistic regression classifier that classifies each sentence in a research paper as either relevant or irrelevant to nanomaterial synthesis. Through an iterative training and labelling routine, they were able to achieve an AUC-ROC score of 0.99, 52% precision and 90% recall. Through observing words that were given the highest and the lowest weightage for relevance in synthesis, the authors claimed a general explainability of their model.

Deconstructing and structuring synthesis protocols allow for comparisons of chemicals used or processing conditions. Such categorization of words and terms from text is known as Named entity recognition, and in the context of chemistry as chemical entity recognition (CER). The authors developed their CER model by identifying parts of speech or context. Their model performed on par with the best CER tool based on expert-defined rules.

Using their tool, the authors analyzed the synthesis of Ag nanowires, nanospheres, and nanocubes and Au nanorods, nanospheres, and nanocubes. Many chemicals appear commonly regardless of nanomaterial morphology or composition, like ethanol, which is often used for washing, and elemental silver and gold. However, polyvinylpyrrolidone (PVP) and hydrochloric acid (HCl) both appear nearly twice as frequently in the syntheses of Ag nanocubes than Ag nanospheres or nanowires, and manually searching the literature, they found several reports of PVP's and HCl's critical roles in directing Ag nanoparticle to a cube morphology. Within the Au nanomaterial articles, CTAB was found to be commonly occurring, particularly within Au nanorod articles when compared to nanosphere or nanocube articles, which was also validated by a targeted manual literature search.

SEM and TEM images were recognized and extracted using a transfer learning approach with a convolutional neural network model which are then analyzed to identify the nanomaterials morphology present.

To enable exploration of the database, the authors developed a complementary browser-based visualization tool that provides flexibility in comparing across subsets of articles of interest.