Word Based Text Parser and Route Extractor for Chemistry Abstracts

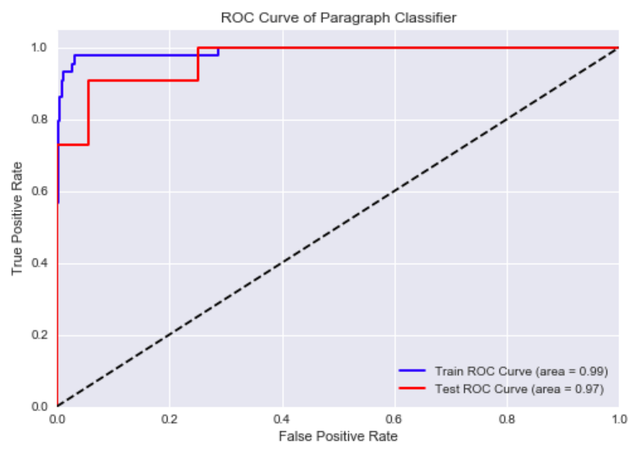
Yannan Yuan, Langkun Chen, Runbo Jiang
University of Pittsburgh

Objective

- Our goal is to build a model to parse and extract information from scientific articles of chemistry which is related to synthesis.
- The model will predict the categories of words in text, such as reactants, products, solvents, conditions, numbers, and units in a synthesis process, then extract this information to directly show the synthesis process.
- Work pipeline: data retrieval → paragraph classification → word labeling and modeling → word extraction.

Text Data Preprocessing

- APIs were used to obtain a large amount of abstracts (>10000) from Elsevier.
- A simple paragraph classifier was trained to get the abstracts that are related to chemical synthesis exclusively.
- In feature engineering, 152 features for each paragraph were created, where 100 features were word count features, 52 features w[illegible] and length of paragraph.
- Support Vector Machine was used to train the classifier, which gave 0.99 and 0.97 AUC scores on training and test set respectively.

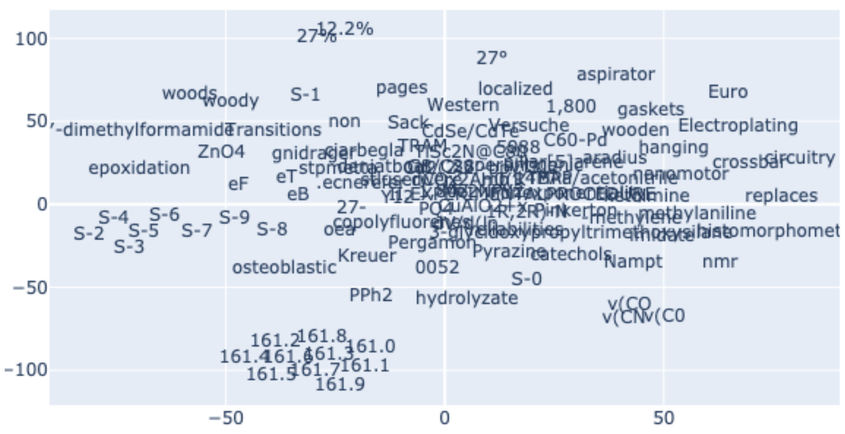


- Apply the well-trained classifier to larger data to get more synthesis abstracts as our further samples.

Feature Engineering

- Tokenize the paragraphs to get words as samples and label all words by our defined categories, which is listed in the table.
- Features were created based on words. The main feature is the word embedding which is obtained from a pre-trained **Word2Vec** model.

Word Embedding TSNE



0	Null
1	Reactant
2	product
3	Genetic material
4	condition
5	apparatus
6	operation
7	Synthesis category
8	Condition unit
9	Amount unit
10	Number
11	Condition description
12	Amount description

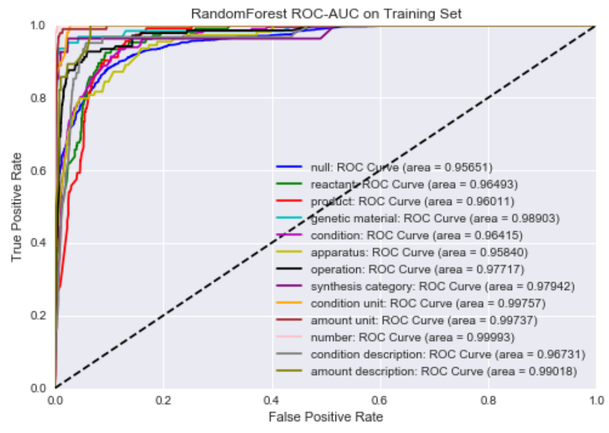
- Word embeddings are positioned in the vector space such that words that share common contexts in the corpus are located close to one another in the space.
- Some **heuristic features**, such as the part-of-speech and chemical entity were involved. ChemDataExtractor and NLTK were used for these rules. These results were transformed into binary feature.

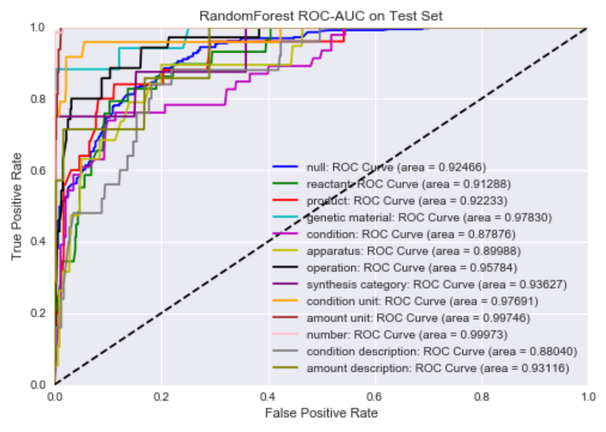
Models and Results

- We trained three models for label prediction: Support Vector Machine, Random forest, and LightGBM.

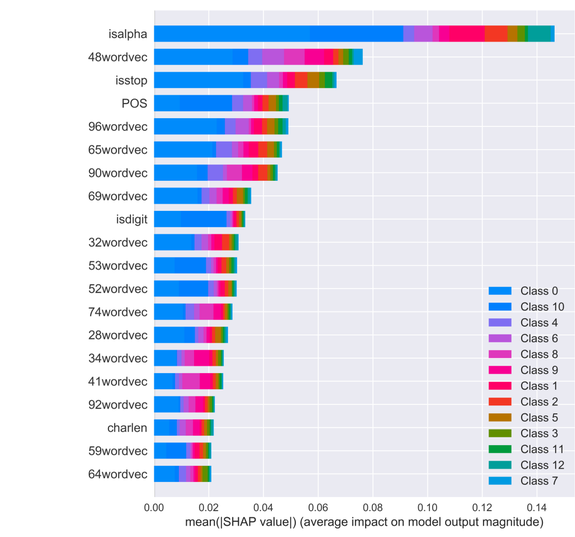
Models	Accuracy		Micro-AUC	
	Train	Test	Train	Test
SVM	0.671	0.671	0.977	0.958
RF	**0.766**	**0.740**	**0.986**	**0.969**
RF Hyperopt	**0.744**	**0.725**	**0.983**	**0.968**
LightGBM	0.806	0.750	0.989	0.967

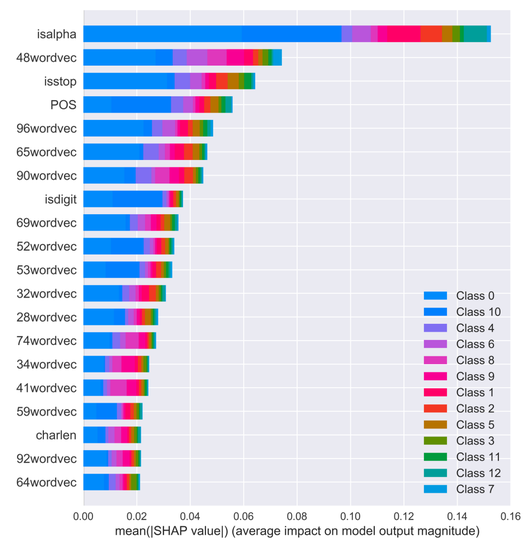
- Based on our best model (RF), we used hyperopt for hyper-tuning, the result became less overfitting. The ROC-AUC curve was plot below:





- SHAP summary plot for training and test set. In general, heuristic features take more effect for label prediction than word embeddings.





Extractor

Word	**Zn(NO3)2·6H2O**	0.297g	1	mmol	were	**dissolved**	in	70	mL
Label	**reactant**	null	number	amount unit	null	**operation**	null	number	amount unit
Word	of	deionized	**water**	and	sealed	in	the	steel	**autoclave**
Label	null	null	**condition**	null	null	null	null	null	**apparatus**
Word	and	**heated**	into	120	˚C	After	the	reaction	the
Label	null	**operation**	null	number	condition unit	null	null	null	null

Future Work

- Build context relevant models
- Find better rule features in feature engineering
- Learn patterns and knowledge from chemical synthesis

Reference

- Kim, E., et al. (2017). Machine-learned and codified synthesis parameters of oxide materials. Sci. Data, 4, 170127
- Swain, M. C. et al. (2016). ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature. *J. Chem. Info. Model*, *56*(10), 1894-1904.