Computers and Chemical Engineering 141 (2020) 107026

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Using word embeddings in abstracts to accelerate metallocene catalysis polymerization research

David Ho, Albert S. Shkolnik, Neil J. Ferraro, Benjamin A. Rizkin, Ryan L. Hartman*

New York University, Department of Chemical and Biomolecular Engineering, 6 MetroTech Center, Brooklyn NY, 11201, United States

ARTICLE INFO

Article history:
Received 21 April 2020
Revised 17 June 2020
Accepted 14 July 2020
Available online 15 July 2020

Keywords:
Machine learning
Metallocene catalysis
Word embeddings
Polymerization
Natural language

ABSTRACT

Natural language processing (NLP) and word embeddings trained neural networks were investigated as a more efficient method to extract useful information on catalytic polymerizations. Thousands of abstracts on metallocene-catalyzed polymerizations were accessed through journal Application Programming Interfaces. These abstracts were then used to create a group of related models to produce word embeddings, making use of the word2vec algorithm. This algorithm turns vocabulary into high dimensional vectors using unsupervised training. These vectors can then be used to show relationships between chemicals, suggest catalysts and activators combinations, understand acronyms, and categorize chemical compounds based on their reagent classification. We hypothesize that one can determine which areas of metallocene catalysis are understudied by comparing the predicted abstract and catalysts combinations with those found in existing abstracts, thereby guiding research to major breakthroughs as scientific literature continues to grow.

1. Introduction

Data mining scientific research is an emerging technique used to quickly gather relevant data as a prerequisite for starting a new study. It was reported in 2015 that 2.5 million scholarly papers were being published per year, with that number expected to rise continuously (Ware and Mabe, 2015). Therefore, it is becoming increasingly difficult to manually parse through even relatively narrow areas of research. An increasing volume of text and data does, however, benefit neural network learning algorithms. The application of neural networks to chemical and materials research has been greatly expanded over recent years (Butler et al., 2018). Methods that take advantage of large quantities of text in order to understand and manipulate it for useful purposes are known as Natural Language Processing (NLP) (Chowdhury, 2003). Word embeddings is a technique in which all words in a set of text are ascribed to a high-dimensional vector generated by a neural network which can learn from their syntactical context as well as the semantics in which they are used (Schnabel et al., 2015). Embedding models can be created from several algorithms, chiefly word2vec (Goldberg and Levy, 2014). Within these algorithms, several machine learning techniques with varying degrees of accuracy are used to generate models. Two competing word2vec algorithms are Continuous Bag of Words (CBOW) and Skip-gram (Mikolov et al., 2013). CBOW operates by using the projection layer for all words, considering the distributed representations of context. In other words, surrounding words are used to predict which word should be put in that context. For this method of classification, term frequencies are not necessarily a good representation as common words have a greater appearance rate in a normal body of text; thus using the term-frequency inverse-document-frequency (TF-IDF) mitigates this issue. TF-IDF is a numerical statistic in order to determine the importance of any term or word in a set of documents.

The TF-IDF is calculated as follows:

$$tfidf(t, D, D^*) = tf(t, d) * idf(t, D^*)$$

The term frequency (tf) is used to identify the theme of a document, where t is the term in document D. This calculates the frequency of a specific term in any given document. The inverse document frequency (idf) is used to show a term's impact on the document by comparing it to the other terms amongst all documents. The term t is once again the term, but D^* is the total number of documents in the corpus divided by the number of documents where term t appears. This means that a high idf shows a word which is not used frequently amongst many documents,

Abbreviations: CBOW, Continuous Bag of Words; NLP, Natural Language Processing; API, Application Programming Interface; TF-IDF, Term Frequency – Inverse Document Frequency; PCA, Principal Component Analysis; t-SNE, t-Distributed Scholastic Neighbor Embeddings.

* Corresponding author.
E-mail address: ryan.hartman@nyu.edu (R.L. Hartman).

https://doi.org/10.1016/j.compchemeng.2020.107026
0098-1354/

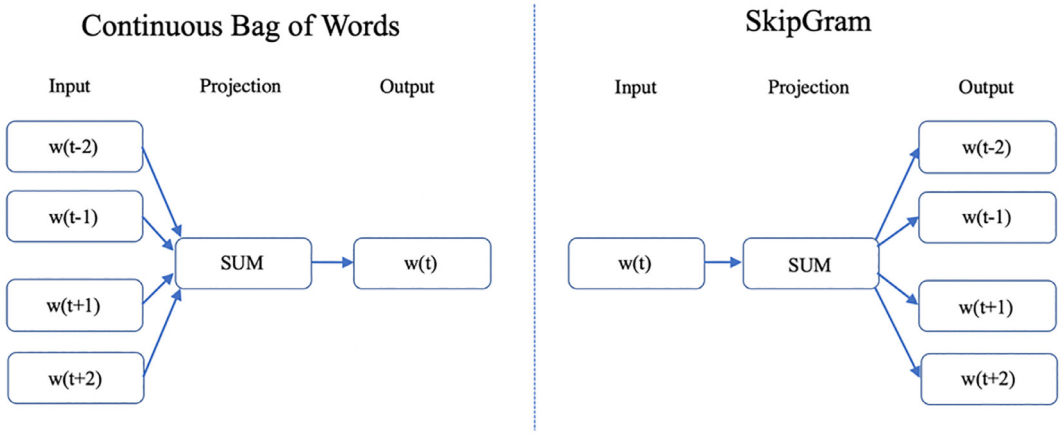

Fig. 1. Continuous bag of words and SkipGram black-box mathematical visual representation of inputs and outputs.

thus showing its significance to its specific document (Salton and Buckley, 2002).

It is well known that the Skip-gram algorithm is more accurate than CBOW, but much slower. That being said, CBOW's accuracy is better for more frequent words; since we are focusing on catalysis amongst catalysis research papers, the CBOW algorithm was tested rather than simply relying on Skip-gram. In contrast to CBOW, Skip-gram works by attempting to maximize the probability of predicting the surrounding words or context of a given word. The embeddings models are generated using mat2vec, a modified version of the word2vec algorithm that has been used to study material science (Weston and Shitoyan, 2019). The two methods can be described by Fig. 1 above.

Typically, a large body of text is required as an input to begin generating meaningful and accurate embeddings models (Roberts, 2016). This comes with the hurdle of obtaining, storing, and processing this large volume of data. By focusing solely on a small window of research and only the keywords associated with the topic, we hope to bypass the shortcomings of using a statistical method like this on a relatively small (~3000 abstracts) body of text.

NLP techniques have been successfully applied before in the context of science and chemistry. A study done in 2019 by Tshitoyan et al., demonstrated the predictive capability of embeddings models by inputting a relatively large amount generic scientific and chemical literature and making accurate predictions about chemical and thermoelectric properties of existing and yet to be discovered compounds. Without any prior chemical knowledge, their word embeddings model was capable of finding complex material science concepts including the underlying structure of the periodic table as well as properties from different materials. From their work in word embedding models being able to understand material science theories, we pursued using word embedding models in order to understand catalytic polymerizations (Tshitoyan and Dagdelen, 2019). Pei-Yua et al. made use of word embeddings in order to study chemical-protein interactions from scientific text faster than if done manually. Their machine learning model specifically extracted chemical-protein interaction pairs and triplets from text where the model identified the chemical compound and its corresponding protein. Being able to extract valuable words and information from text is one of the first stepping stone for future works of developing models capable of extracting high-level information from sentences, such as the method used in identifying catalytic polymerizations in this research (Lung et al., 2019). This however only scratches the surface of what is possible from using NLP to analyze scholarly articles. Furthermore, the application of other machine learning tactics, including and expanding upon NLP, have been discussed in order to predict atomistic potentials, understand catalytic properties, and find cheaper ways to solve Kohn-Sham equations in DFT (Kitchin, 2018). There are great strides to provide methodology, usage and impacts of machine learning in catalysis, and even future work that could shift traditional catalysis research. The work by Toyao et al. shows the design, synthesis, and characterization of catalytic materials as well as contributes to this discussion of the potential in this newly growing field (Toyao et al., 2020). This paper puts these theories into application, but also broadens it by creating a template possibly useful to all material science fields through the usage of NLP. The field of catalysis is enormous, making it attractive for application of cutting-edge data informatics tools.

This paper applies the word embedding algorithm to a relatively small, highly focused dataset pertaining to metallocene-catalyzed polymerizations. This was done in order to generate lists of catalysts, activators, monomers, and their shared context as a way of datamining literature from previously studied reactions. Although this is a specific application, the underlying neural network and algorithms can be trained for any focused dataset or topic of research. The goals of this research were to demonstrate an accurate grouping of similar categories of chemicals, visualize the word embedding space of metallocene-catalyzed polymerization research as whole, and autonomously generate previously unresearched combinations of catalyst/activator/monomer predicted by embedding vectors of the chemical names alone.

2. Methods

2.1. Abstract handling and manipulation

Approximately three thousand abstracts were obtained pertaining to metallocene polymerization chemistry. The abstracts were downloaded using Elsevier's Scopus Application Programming Interface (API) (Elsevier, 2020) and Springer Nature's API (Springer Nature API Portal, 2020). From the APIs, abstracts that contained a metallocene and the keyword polymer were downloaded and duplicates were removed. Spelling and nomenclature of chemical names varied so the corpus was preprocessed to standardize and decrease the number of unique vocabulary words. Furthermore, to gain the strongest correlation of related words, copyright symbols, watermarks, and other forms of identification from authors were stripped. Using mat2vec, a modified word2vec algorithm containing useful preprocessing steps for chemistry papers, the beginnings of sentences were made lowercase while preserving the capitalization of acronyms, chemical formulas, and roman numerals for oxidation states. All numbers were replaced by

the token <num> which further reduced the size of the vocabulary. Nomenclature was standardized across texts; for example, "polypropene" and "polyethene" were translated to "polypropylene" and "polyethylene", respectively. Approximately 50 chemical compounds were renamed and standardized. Overall, the size of the vocabulary was reduced from 27,361 unique words to 25,945 unique words.

2.2. Unsupervised learning and training

Similarly to other word embedding algorithms and NLP techniques used in catalysis research, this learning algorithm is unsupervised; meaning data was gathered through specific searches but models used all data collected. This was used rather than a binary relevance network which filters data through a neural network. The data obtained from the API searches were relevant and the usage of binary relevant networks seemed unnecessary to achieve the same results (Berger, 2015; Huang and Ling, 2019). A machine learning model is only as capable as its data, and separating the useful data from its useless counterparts is imperative (Rothenberg, 2008). The usage of the API queries in order to find abstracts were used as a filter in order to find research papers relating to the keywords used: metallocene and some variation of polymer. This returned all abstracts that the APIs found relevant; from there, there was no neural network used to filter unrelated articles amongst these results. Rather, every article was used in the training of our model. Trying to train binary relevant/irrelevant screening network using TF-IDF proved to be too inaccurate for the relatively small size of the corpus and too time consuming to produce training data. Furthermore, using an unsupervised learning method allows the process to be applicable towards any topic with minimal tuning rather than just for metallocene catalysts.

Using the preprocessed corpus, mat2vec was used to generate a series of embeddings models using Skip-gram and Continuous Bag of Words. Further adjustments were also tested such as varying the number of embedding dimensions and using an unprocessed corpus, a preprocessed corpus, and a preprocessed corpus with phrasing enabled. The unprocessed corpus contained the plaintext of all the abstracts while the other two normalized the entire text file; normalizing the file consisted of the preprocessing techniques mentioned earlier: making capitalization across sentences, chemicals, and acronyms uniform; replacing numbers with <num>; and standardizing nomenclature of chemical compounds. In addition, the preprocessed corpus with phrasing enabled grouped common words appearing in a similar context together as a phrase and treated them similarly to a single word. These models are compared against each other by their accuracy in solving a set of metallocene polymerization related analogies. These analogies were prepared from both domain expertise as well as a previously made set of chemical analogies. The analogies from domain expertise focused on catalytic polymerization and was weighted more than the preset list of analogies. The total list of analogies amounted to 201,350 analogies. Furthermore, the chemical names were also defined from domain expertise and manual inspection rather than from an information extraction tool.

The preset analogies come from the previously stated paper from Tshitoyan et al. and a copy can be found in our source code. The model that solves the most analogies correctly was selected for further analysis. In total, 18 different models were tested in order to determine the best model for this specific algorithm. The results of these models were normalized, returning a percentage of the analogies solved over the analogies solved by the best model. It was quickly found that the processed corpus Skip-gram model consistently yielded the best results for completing analogies. Analogies clearly shows grouping, clustering, and understanding of the material; so the set of two-hundred thousand analogies was used as the testing process. By being able to relate certain terms to solve analogies, these models are capable of clustering and identifying trends, patterns, and related materials. In the Figs. 2, and 3 below, the relative analogy of each model was tested and shown; 100% relative analogy accuracy resulting in 64.3% overall accuracy, correctly solving 129,453 analogies out of 201,350.

The embeddings of chemical names are collapsed down to two dimensions using Principle Component Analysis (PCA) in order to

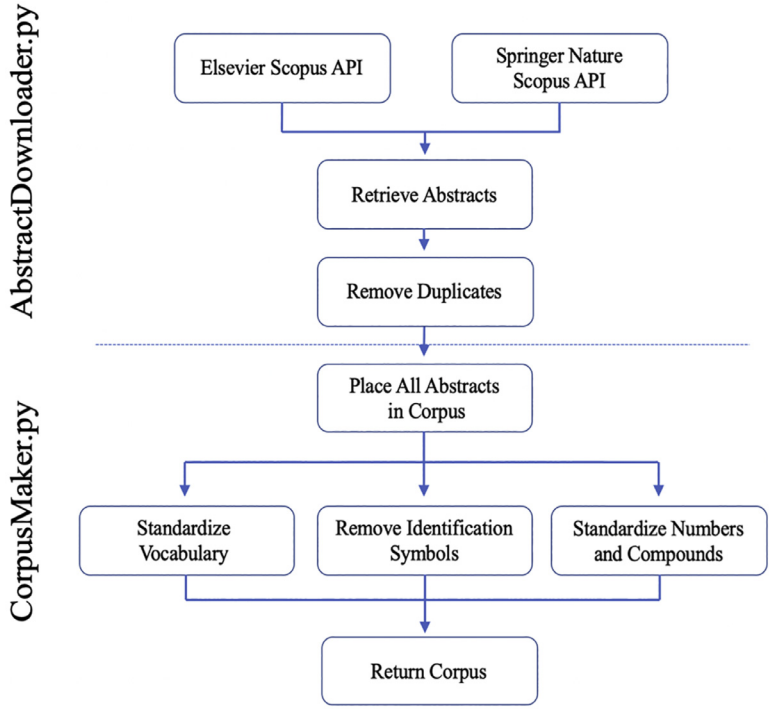

Fig. 2. Flow diagram of abstract collection to corpus generation.

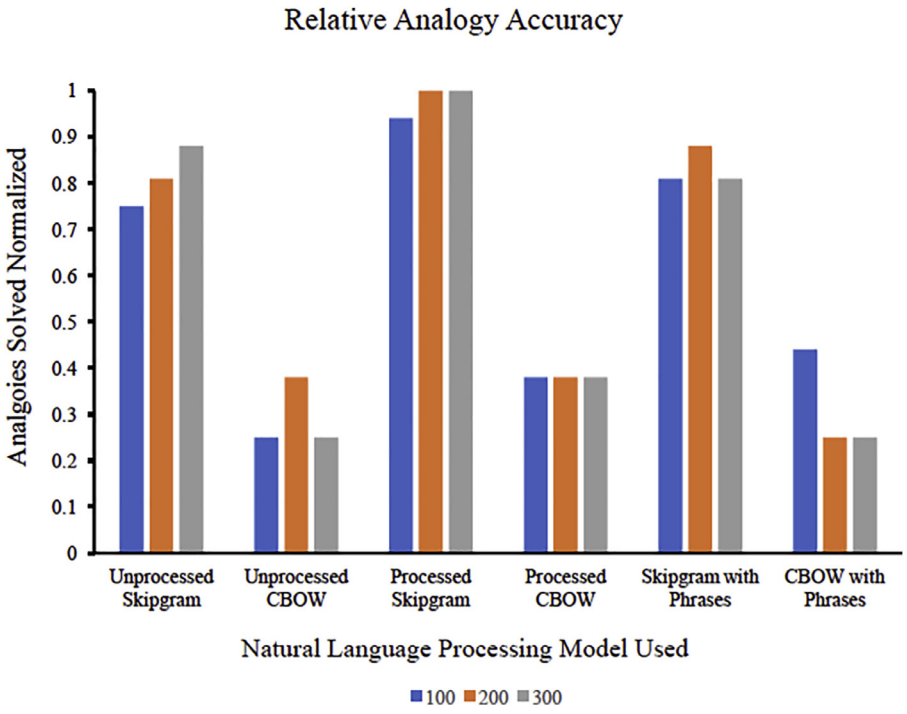

Fig. 3. Grouped bar chart of relative analogy accuracies amongst different Natural Language Processing (NLP) training models.

see how the model groups these chemicals. PCA works by mathematically reducing the high dimension vectors created in the model for each set of words by finding correlations between similar dimensions to minimize the variables of each vector while keeping the variance, distribution, and information of the original vectors. PCA forms the linear combination attempting to maximize the standard deviation and Rayleigh Quotient. PCA is extremely useful for finding these linear correlations. Although studies such as Chen et al. (2013) have shown that potentially useful information is lost when the dimensions are reduced, it can still help a researcher make qualitative decisions based on the results of PCA.

A larger set of vocabulary found by looking at the most similar words to some common chemicals used in metallocene catalysis was plotted using t-Distributed Stochastic Neighbor Embeddings (t-SNE). T-SNE is a technique to visualize high dimensional data by translating a data point to a two or three dimensional map. It does this by first converting the high dimensional data into conditional probabilities that show similarities between two points. Then it uses a Student-t distribution to compute similarities between two points in a low dimension space rather than the high dimensional space like before. Ultimately, the algorithm minimizes the Kullback-Leibler divergence between the two probability distributions and applies a gradient descent to form its graph (Maaten and Hinton, 2008). We use this algorithm to take our 200-dimensional embeddings vector of each word, apply it in order to group words deemed similar and plot them in two dimensions. T-SNE uses probability in order to find similarities and correlations between the dataset in a high-dimensional space and the same dataset in a lower dimensional space, while PCA uses a linear dimension reduction technique to preserve larger distances between pairs. PCA geometrically projects the data points to a lower dimension. These data points are called principal components; PCA attempts to obtain the best summary of the data while attempting to minimize the number of principal components and keeping the variance of data points (Lever et al., 2017). In other words, PCA finds the linear correlation within the dataset to give a lower-dimension vector attempting to maintain the variance while t-SNE attempts to find multiple patterns and correlations amongst a high-dimension dataset and finds the similar correlations in a low-dimensional dataset. Both of these graphs are then generated and shown in order to show the clustering and determinations the NLP algorithm has produced.

3. Results and discussion

3.1. Analogy determination

The embeddings model is shown to be capable of creating meaningful word analogies first outlined in Mikolov et al. (Weston and Tshitoyan, 2019). These analogies are created by inputting two words with a known relationship, such as the category of catalyst, zirconocene, and its metal element name, Zr, then another word such as Ti, to find the word with the most similar relationship as the first two, which in this case is titanocene. A list of example analogies predicted by the embeddings model is show in Table 1. The full list of all analogies used can be found in the code.

General formula of the analogies is (a-$b + c = d$)

3.2. Principal component analysis

Principal component analysis (PCA) was then used to visualize the high dimensional vector representations of words by comparing the dimensions with the greatest variation between the inputted words. A more complete set of the acronyms and chemical names PCA can be found in figure S1 in the supplementary files.

As can be seen in Fig. 4, the acronyms appear relatively close to their full name which shows their embeddings are correctly drawing a close association between them. TMA is one exception as it appears as an outlier in dimension 1 but is still close to its full name trimethylaluminum in dimension 2. This suggests that the acronym TMA is used somewhat differently to its full name compared to the other activators. In the t-SNE plot later which is initially seeded by PCA, TMA and trimethylaluminum are grouped

Table 1
Examples of word embeddings analogies used in comparisons for determining best Natural Language Processing (NLP) model.

Positive 1[a]	Negative[b]	Positive 2[c]	Result[d]
zirconocene	Zr	Ti	titanocene
zirconocene	Zr	Hf	hafnocene
titanocene	Ti	Zr	zirconocene
titanocene	Ti	Hf	hafnocene
hafnocene	Hf	Zr	zirconocene
hafnocene	Hf	Ti	titanocene
MAO	methylaluminoxane	triethylalmunium	TEA
MAO	methylaluminoxane	triisobutylaluminum	TIBA
MAO	methylaluminoxane	trimethylaluminum	TMA
TEA	triethylalmunium	methylaluminoxane	MAO
TEA	triethylalmunium	triisobutylaluminum	TIBA
TEA	triethylalmunium	trimethylaluminum	TMA
TIBA	triisobutylaluminum	methylaluminoxane	MAO
TIBA	triisobutylaluminum	trimethylaluminum	TMA
TIBA	triisobutylaluminum	triethylalmunium	TEA
TMA	trimethylaluminum	methylaluminoxane	MAO
TMA	trimethylaluminum	triisobutylaluminum	TIBA
TMA	trimethylaluminum	triethylalmunium	TEA
methylaluminoxane	MAO	TEA	triethylaluminoxane
methylaluminoxane	MAO	TIBA	triisobutylaluminum
methylaluminoxane	MAO	TMA	trimethylaluminum
triethylaluminoxane	TEA	MAO	methylaluminoxane
triethylaluminoxane	TEA	TIBA	triisobutylaluminum
triethylaluminoxane	TEA	TEA	triethylaluminoxane
triisobutylaluminum	TIBA	MAO	methylaluminoxane
triisobutylaluminum	TIBA	TMA	trimethylaluminum
triisobutylaluminum	TIBA	TEA	triethylaluminoxane
trimethylaluminum	TMA	MAO	methylaluminoxane
trimethylaluminum	TMA	TIBA	triisobutylaluminum
trimethylaluminum	TMA	TEA	triethylaluminoxane
polyethylene	ethene	propene	polypropylene
polypropylene	propene	ethene	polyethylene
heterogeneous	supported	unsupported	homogeneous
homogeneous	unsupported	supported	heterogeneous
unsupported	homogeneous	heterogeneous	supported
supported	heterogeneous	homogeneous	unsupported
Cp2ZrCl2	Zr	Hf	Cp2HfCl2
Cp2ZrCl2	Zr	Ti	Cp2TiCl2
Cp2HfCl2	Hf	Zr	Cp2ZrCl2
Cp2HfCl2	Hf	Ti	Cp2TiCl2
Cp2TiCl2	Ti	Hf	Cp2HfCl2
Cp2TiCl2	Ti	Zr	Cp2ZrCl2

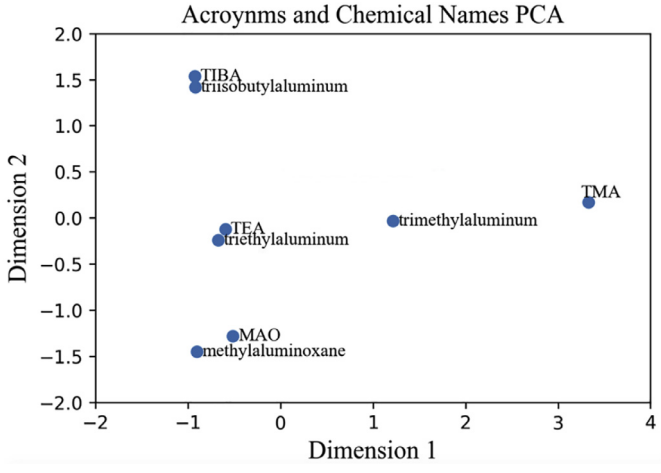

Fig. 4. Principal component analysis plot of metallocene activators and their acronyms.

closely together when other types of chemical names are plotted together. This suggests that the embeddings model is correctly predicting which chemical names have similar meanings and also that the acronym TMA might be being used differently than other acronyms in relation to is full name.

The PCA plot provides an insightful visualization of the high-dimensional word embeddings and their relative closeness to one another. Fig. 5 shows how Skip-gram can identify when reagents belong to distinct categories. Catalysts, co-catalysts/activators, and monomers each can be seen grouped in three corners of a triangle, maximizing the distance between themselves. The PCA serves as a seed for a greater number of word representations.

3.3. T-SNE visualization

T-SNE has proven useful when plotting a greater number of word representations than PCA. The t-SNE plot shown in Fig. 6 was seeded using PCA initially.

The t-SNE plot shows how words' similarities are interpreted using word2vec embeddings. Cosine similarity was used in order to determine the 50 closest words; cosine similarity measures the likeness between two words by measuring the cosine of the angle of the two vectors. This works well when translating a high dimension vector space into a lower one, as done with NLP. This concept is more effective in analyzing semantic similarity over Euclidean distance as the Euclidean distance calculates the raw distance between two points rather than the angle of the two vectors which can get obscured during translation. Word2vec inherently calculates the cosine similarity when looking for most related words (Rehurek, 2019; Han et al., 2000; Yang et al., 2019). Words that belong to categories such as monomers, polymers,

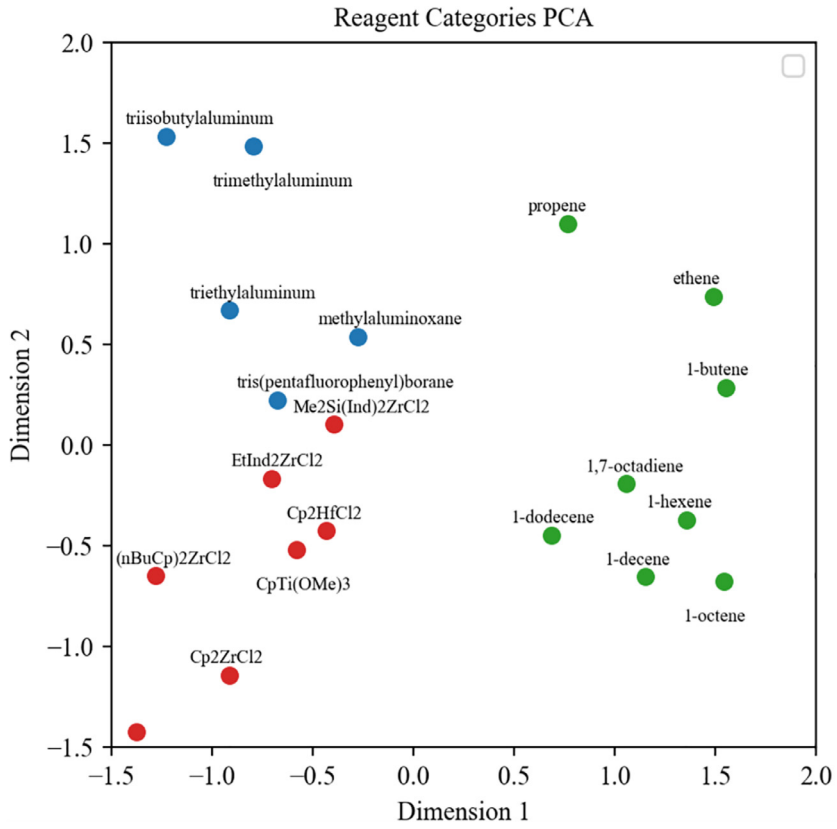

Fig. 5. Principal component analysis (PCA) plot of embeddings for: monomers (green), catalysts (red), and activators (blue).

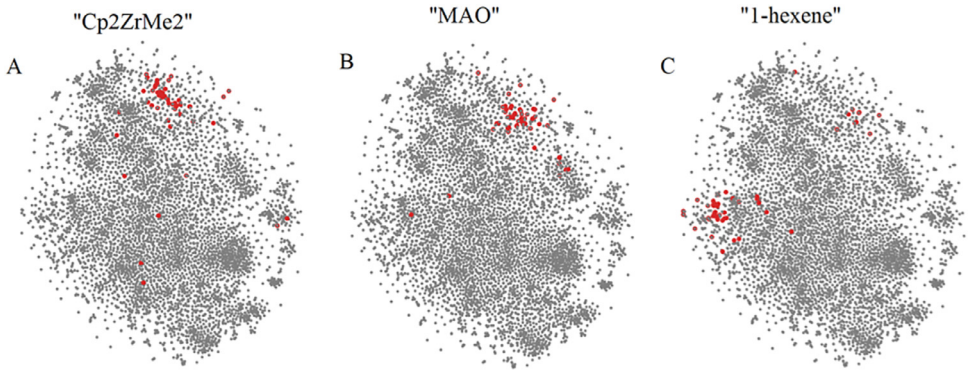

Fig. 6. t-SNE plot of the entire vocabulary highlighting the top 50 closest words by Cosine Similarity in the embedding space to the words A) Cp_2ZrMe_2, B) MAO, and C) 1-hexene.

activators, catalysts, and physical properties group together while dissimilar words spread apart. Other groupings observed are adjectives for physical properties, words relating to crystallography, and heterogeneous catalyst supports. The 50 closest words as well as their cosine similarities can be found in the supplementary files (1Hexene.csv, MAO.csv, and Cp2ZrMe2.csv). Exploring the outlying groups also revealed vocabulary pertaining to unrelated abstracts that slipped by the initial filtering, showing that this visualization can be used for quality control in obtaining a corpus for the generation of embeddings models.

3.4. Under-researched topics

A list of catalysts and activators was created by looking through the top similarity matches of sixteen common catalysts and activators. The corpus was checked to see how many times specific catalysts and activators appeared in the same abstract. The top result, Cp_2ZrCl_2 and MAO were used as a template catalyst-activator combination analogy and each other catalyst was checked to see what the top activator results would be (MAO – Cp_2ZrCl_2 + Other catalyst = Predicted Activator). The predicted activators for each catalyst compared to the number of abstracts containing this combination are presented in Table 2.

Searching through the corpus, it became evident that most catalyst + activator combinations had not been studied for various reasons including costs, availability, and predicted catalytic activity. That being said, it should be noted that this search of under-researched catalyst and activator combinations has not been thoroughly tested, but is a feature that can be passably seen from the results of our embeddings model. As expected, when predicting catalyst-activator combinations, analogies to Cp_2ZrCl_2: MAO display other, well studied reactions. This shows that the vector

Table 2
Predicted catalyst-activator pairs based on analogies.

Catalyst	Activator	Number of Papers in Corpus
Cp*Ti(OBz)3	triisobutylaluminum	4
Cp*Ti(Obz)3	methylaluminoxane	7
Cp*Ti(Obz)3	triethylaluminum	None
Cp*Ti(Obz)3	CPh3B(C6F5)4	None
Cp2ZrCl2	methylaluminoxane	121
Cp2ZrCl2	triethylaluminum	5
Cp2ZrCl2	CPh3B(C6F5)4	2
Cp2ZrCl2	TIBAO	1
Cp2ZrCl2	triisobutylaluminum	7
Cp2HfCl2	TIBAO	None
Cp2HfCl2	CPh3B(C6F5)4	None
Cp2HfCl2	methylaluminoxane	4
Cp2HfCl2	triethylaluminum	None
Cp2TiCl2	AlEtCl2	4
Cp2TiCl2	AlEt2Cl	3
Cp2TiCl2	triethylaluminum	1

representations of these chemicals are not random. Upon investigation of the word2vec source code, it is apparent that these words are indeed being correctly grouped as opposed to a random output. During the training phase, word2vec takes every word in order to increase or decrease its cosine similarity, or relatedness, with the words around it. However, it not only brings the words' vectors closer together if the two words appeared next to each other, but also uses two phases, hierarchical and negative sampling, in order to push away a sample of words. Negative sampling randomly pushes away a random set of words to its given word whenever that word is brought closer together with another word. Hierarchical sampling takes the neighboring words of a given word and compares it with its own subset of words chosen from a tree data structure relating to more frequent words. Through this training process, not only are words that appear together in the corpus related, but word2vec brings the words that are likely to appear together while disregarding extremely common words such as prepositions. That is why chemical components, catalyst and activator combinations, and even advanced materials are able to be brought together as the corpus contains enough occurrences of these complex phrases, and similar occurrences of these phrases though the words may not be the same, in order to recognize their connection. This is similar to the popular recognition word2vec is able to make: man – king and woman – queen (Mikolov and Sutskever, 2013; Mikolov et al., 2013). Combinations predicted but not actually observed in the corpus could be potential areas of future study. Overall this process presents a method of selecting under researched combinations for a series of activator screenings that is independent of any physical properties of the chemicals themselves.

The "Holy Grail" of applying NLP to academic research would be the ability to accurately predict the values of physical properties of the words being represented as embeddings. This would require data that is usually locked away inside figures and tables of the full articles to be put in a format that is friendlier for a learning algorithm to decipher. Currently some general chemical and physical properties can be predicted to varying degrees of accuracy using a corpus encompassing millions of abstracts about thermoelectric materials, but for highly focused areas of research with relatively few papers, this becomes very difficult. It is conceivable that an unsupervised NLP working as part of a larger neural network seeded with a large volume of physical data will be able to make predictions about the chemical properties of untested polymerization catalysts and activators. Neural machine translation models have been applied by Nam and Kim (2016) to predict organic chemical reactions. Furthermore, named entity recognition has been applied to extract information from material science literature in order to drastically improve the efficiency between researchers and literature. However, there are still some drawbacks such as a lack of entity-relation extraction (Corbett and Boyle, 2018; Weston et al., 2019). Regardless, work like this shows how promising these techniques can be to reaction chemistry.

The biggest challenge while creating embeddings models and searching through the results was the number of fragmented chemical formulas and identical chemicals named differently across different abstracts. These inconsistencies of nomenclature and formating necessitated the preprocessing step to correct common spelling differences (such as aluminum versus aluminum). However, many other difficult to standardize inconsistencies, such as chemical formulas, slipped through. One could spend a great deal of time manually accounting for each inconsistency until every chemical name and every formula is completely standardized, but that would completely negate the speed advantage data mining and NLP provides. In order to aid human and nonhuman readers of scientific journals, authors and publishers should make more strict formating rules when it comes to chemical names and formulas. Not only will this aid future data scientists, but also the amateur researcher who might not know the many different ways the same catalyst can be represented as.

4. Conclusion

Making use of an embeddings algorithm when researching metallocene polymerizations can be a fast way for a researcher to be introduced to the topic. Lists of catalysts, co-catalysts/activators, and monomers can be compiled from the wealth of past research based on closeness to such chemicals the researcher might already know. Important properties to look out for that have been studied in the past in regard to certain polymerizations can also be found by looking for words similar to the word property or using a known property of interest as a seed word. Synonyms and acronyms can be found by looking for very closely matching word vectors or by using analogies. Analogies can also be used to generate predicted catalyst/activator combinations that may be of interest to research. Used in conjunction with other data sources, there is potential for unsupervised embeddings to be applied in larger neural networks to expand catalysis research by making use of the wealth of existing knowledge.

The use of data visualization tools can make it feasible for an amateur researcher to quickly learn about the range of different catalysts/activators/monomers that exist for metallocene catalyzed polymerization. Chemical names that appear frequently with similar context words are automatically grouped together and appear as the closest neighbors in the high-dimensional space. Traditionally, researcher trying to substitute a reagent may need to read dozens of papers to create a list of possible substitutes. This can take many hours to do manually. Using NLP, this same researcher can download the several thousand relevant abstracts, create a Skip-gram model, and start searching keywords similar to the reagent were searching for in under an hour.

Data and code access

The data abstracted came from Elsevier and Springer Nature's API. The abstracts can be similarly extracted from our code and a unique public key. The public key can be input into the text file provided. The API keys for Elsevier and Springer Nature can be created from the publisher's API on the developer portal, respectively (Elsevier, 2020; Springer Nature API Portal, 2020). After those keys are created, they can be used in the code base, which also contains the training models used for the natural language processing (Ho, 2014).

Declaration of Competing Interest

None.

CRediT authorship contribution statement

David Ho: Investigation, Methodology, Software, Visualization, Validation, Writing - original draft. **Albert S. Shkolnik:** Software, Visualization, Writing - review & editing. **Neil J. Ferraro:** Data curation. **Benjamin A. Rizkin:** Conceptualization, Writing - review & editing. **Ryan L. Hartman:** Conceptualization, Investigation, Supervision, Writing - review & editing.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant Number CBET-1701393. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.compchemeng.2020.107026.

References

Berger, M.J., 2015. Large scale multi-label text classification with semantic word vectors. Technical report, Stanford University.
Butler, K.T., Davies, D.W., Cartwright, H., Isayev, O., Walsh, A., 2018. Machine learning for molecular and materials science. Nature 559 (7715), 547–555.
Chen, Y., Perozzi, B., Al-Rfou, R., Skiena, S., 2013. The expressive power of word embeddings. arXiv preprint arXiv:1301.3226.
Chowdhury, G.G., 2003. Natural language processing. Annual review of information science and technology 37 (1), 51–89.
Corbett, P., Boyle, J.Chemlistem, 2018. chemical named entity recognition using recurrent neural networks. J Cheminform 10, 59. https://doi.org/10.1186/s13321-018-0313-8.
Elsevier, 2020. Elsevier Developers. Elsevier Developer Portal. https://dev.elsevier.com/.
Elsevier, 2020. Elsevier Developers. Elsevier Developer Portal. https://dev.elsevier.com/apikey/manage.
Goldberg, Y., Levy, O., 2014. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722.
Han, J., Kamber, M., Pei, J., 2000. Data Mining: Concepts and Techniques. Morgan Kaufmann.
Ho, D. Aritificial Intelligent Bibliometric Analyzer. 2014 https://github.com/davidhodev/Artificial-Intelligent-Bibliometric-Analyzer.
Huang, L., Ling, C., 2019. Representing Multiword Chemical Terms through Phrase-Level Preprocessing and Word Embedding. ACS Omega 4 (20), 18510–18519. doi:10.1021/acsomega.9b02060.
Kitchin, J.R., 2018. Machine learning in catalysis. Nat Catal 1, 230–232. https://doi.org/10.1038/s41929-018-0056-y.
Lever, J., Krzywinski, M., Altman, N, 2017. Principal component analysis. Nat Methods 14, 641–642. https://doi.org/10.1038/nmeth.4346.
Lung, P.Y., et al., 2019. Extracting chemical–protein interactions from literature using sentence structure analysis and feature engineering. Database 2019.
Maaten, L.V.D., Hinton, G., 2008. Visualizing data using t-SNE. Journal of machine learning research 9 (Nov), 2579–2605.
Mikolov, T., Sutskever, I., et al., 2013. Distributed Representations of Words and Phrases and their Compositionality. arXiv:1310.4546.
Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
Nam, J., Kim, J., 2016. Linking the neural machine translation and the prediction of organic chemistry reactions. arXiv preprint arXiv:1612.09529.
Rehurek, R., 2019. n.d.. Gensim: Topic modelling for humans. Retrieved June, from https://radimrehurek.com/gensim/models/keyedvectors.html.
Roberts, K., 2016. Assessing the corpus size vs. similarity trade-off for word embeddings in clinical NLP. In: Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP), pp. 54–63.
Rothenberg, G., 2008. Data mining in catalysis: Separating knowledge from garbage. Catalysis Today 137 (1), 2–10. doi:10.1016/j.cattod.2008.02.014.
Salton, G., Buckley, C., 2002. Term-weighting approaches in automatic text retrieval. Information Processing & Management 24 (5), 513–523. doi:10.1016/0306-4573(88)90021-0.
Schnabel, T., Labutov, I., Mimno, D., Joachims, T., 2015. Evaluation methods for unsupervised word embeddings. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp. 298–307.
Springer Nature API Portal, 2020. Springer API. https://dev.springernature.com/.
Springer Nature API Portal, 2020. Springer API. https://dev.springernature.com/signup.
Toyao, Takashi, et al., 2020. Machine Learning for Catalysis Informatics: Recent Applications and Prospects. ACS Catalysis 10 (3), 2260–2297. doi:10.1021/acscatal.9b04186.
Tshitoyan, V., Dagdelen, et al., 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. Nature 571 (7763), 95–98.
Ware, M., Mabe, M., March 2015. The STM Report: An overview of scientific and scholarly journal publishing. Fourth Edition 6, 6–7.
Weston, L., shitoyan, T, et al., 2019. Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature. Journal of chemical information and modeling 59 (9), 3692–3702.
Weston, L., Tshitoyan, V., et al., 2019. Named Entity Recognition and Normalization. Applied to Large-Scale Information Extraction from the Materials Science Literature 2019.
Weston, L., Tshitoyan, V., Dagdelen, J., Kononova, O., Trewartha, A., Persson, KA, et al., 2019. Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature. Journal of chemical information and modeling 59 (9), 3692–3702. http://dx.doi.org/10.1021/acs.jcim.9b00470. Retrieved from https://escholarship.org/uc/item/7r45h4mf.
Yang, M.Y., Rosenhahn, B., Murino, V., 2019. Multimodal scene understanding: Algorithms, applications and deep learning. Academic Press, Elsevier, London.