Text Mining and Social Media Mining - Paper Review 2

Unsupervised word embeddings capture latent knowledge from materials science literature

Author: Michele Guderzo

Warsaw, Academic Year 2021/22

Contents

1 Introduction

We are going to review the article entitled *"Unsupervised word embeddings capture latent knowledge from materials science literature"*, conducted by Vahe Tshitoyan & John Dagdelen et al. in 2019.
The study addresses the difficulty of finding information in the field of materials science from scientific publications. It gives attention to the use of supervised natural language processing, demonstrating that information in publications can be efficiently encoded as information-dense word embeddings without human intervention. Thanks to these embeddings, it is possible to retrieve meaningful information about the periodic table and structure-property relationships in materials.

In this paper, our objective is to deepen and analyse the article. We are going to discuss the methods the authors used, giving more attention to the text mining techniques that have been applied, suggest, if required, improvements to the study and debate the results they found. But before doing so, we are going to provide the reader with some background information, in order to give a better understanding about the topic we are talking about.

2 Paper Review

2.1 Background

Scientific publications have been, for many years, one of the main places where people, especially scientists and researchers, can objectively document and inform themselves about a specific field. The roots of academic scientific publication can be traced back as far as 1665, when Henry Oldenburg of the British Royal Society founded the *Philosophical Transactions of the Royal Society.* This scientific journal created a sense of competition among scientists to be the first to publish a new scientific finding, an incentive that is continued in modern scientific journals.
Scientific researches not only provide the dissemination of studies and discoveries, but can also yield indirect rewards. For example, they affect a researcher's job prospects and ability to be promoted or gain tenure. Publishing a scientific paper can result in fruitful new scientific collaborations, including financially profitable arrangements for authors in academe, as a result of commercial overtures for collaboration or consultancy.
However, there could be risks for an author. Competitors might use results presented in a paper to advance their own research and "scoop" the original author in future publications. Also, other researchers might use information presented in a paper to invalidate or question the author's own findings, and publish conflicting results. At the same time, if an author choose to not publish a paper, there could be another researcher with the same idea and then publish a study with the same findings, thus receiving

the credit.

In the next section we will look at the main part of the study, including the dataset found, the methods used and the conclusions that have been drawn.

2.2 Techniques used

To begin with, the authors focused on some approaches of natural language processing (NLP). The *Natural Language Processing* is the application of computational techniques to the analysis and synthesis of natural language and speech. It consists of 3 main parts: text preprocessing, text analysis and visualization. Of these, the one we are most interested in is text preprocessing, since it represents the procedure of "coding" text information, from text to numbers.

The NLP is certainly a broad field. One technique that stands out in this area is to assign high-dimensional vectors to words in a text corpus, forming the so-called "word embeddings". *Word embeddings* are overall terms, for a set of modelling techniques in natural language processing, in which words or phrases of a vocabulary are mapped into vectors of real numbers. The aim is to keep the meaning and the information contained as intact as possible.

The machine learning algorithms generally used, according to the authors, to form word embeddings are "Word2vec", which is the first widely disseminated word embedding method, and "GloVe", which is another well-known model for distributed word representation. There are also other algorithms, such as "principal component analysis" (PCA) and "t-distributed stochastic neighbour embedding" (t-SNE), but these are mostly used to decrease the size of the space of word vectors and to allow their visualisation in two- or three-dimensional space.

Worc2vec, developed by Tomas Mikolov, is an algorithm that uses a neural network model to learn word associations from a large corpus of text. Once trained, such a model can detect synonymous words or suggest additional words for a partial sentence. This algorithm is based on two models of architecture, which are "Skip-Gram" and "continuous bag-of-words" (CBOW). While Skip-Gram predicts the surrounding words, basing on the current word, CBOW predicts the current word, basing on the context.

On the other hand, despite being very similar, *GloVe*'s training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.

2.3 Dataset collection and processing

In order to train the embeddings, the authors collected about 3.3 million scientific abstracts, mostly related to materials science, physics and chemistry, from more than 1000 journals. At the end of the data collection, a vocabulary of about 500000 words were formed. The data were found through a combination of Elsevier's Scopus and Science Direct application programming interfaces (APIs), the Springer Nature API, and web scraping.

At this point, the authors proceeded to heavily process the dataset.
First of all, they eliminated abstracts written in foreign words by using text search and regular expression. *Text search* (or *full-text search*) refers to those techniques for searching a document or a collection of them in a full text database. In a full-text search, the search engine examines all the words in each stored document and tries to find a match according to certain criteria (such as words provided by the user). On the other hand, a *regular expression* (or *regexp*) is a sequence of symbols that identifies a set of strings. It defines a function that takes a string as input, and returns a yes/no value as output, depending on whether the string follows a certain pattern or not. Thus, it is extremely useful in finding information in a text corpus.
Proceeding with data processing, all abstracts with words in the title that did not refer to scientific research were removed using the same methods as described above.
Then, a binary classifier was introduced to find only texts relating to inorganic materials, which can distinguish between relevant and irrelevant abstracts. A doubt arises here: it is not clear why they introduced such a classifier. The authors have always spoken of generic "materials", so we expected an analysis of the entire literature of materials science, thus including both organic and inorganic compounds. In principle, inorganic materials refer to those chemical compounds that do not contain carbon atoms, but there are some exceptions. This includes carbon dioxide and carbonic acid and their salts, bicarbonates, carbonates and carbon monoxide. Since the study places a strong focus on thermoelectric materials and the application of the term 'thermoelectric' as a keyword, we can deduce that the materials with this characteristic are precisely the inorganic materials. However, ideas aside, there is no specific explanation of this.
Once the relevant abstracts have been labelled and retrieved, they were used to train a logistic regression-based classifier, in which each document is described by a term frequency-inverse document frequency (TF-IDF) vector. This significantly increased the performance of the classifier, which achieved an f1-score of 89%. We remind that the TF-IDF vector is a function used in information retrieval to measure the importance of a term with respect to a document or collection of documents, while F1-score is a measure of test's accuracy: the closer to 1, the more accurate is the test.
Finally, the abstracts considered relevant were then subjected to the process of tokenization using Chem-

DataExtractor. The purpose of tokenization is to divide the text contained in a document into, precisely, tokens: depending on the application under consideration, these may be single words, whole sentences or parts of them. There are several ways to perform tokenization, the simplest and most intuitive but often also the most effective is to break up the text according to so-called delimiters. The delimiters usually considered are those belonging to the so-called "strong" punctuation, such as line (\n), sentence (.) or period (, ; :) terminators. Obviously, exceptions must be taken into account and managed: for example, the character "." used as a decimal digit separator is not to be considered in the same way as when delimiting the end of a sentence. In fact, numbers with units were not always tokenized correctly by ChemDataExtractor. The solution the authors proposed was to split the common units from numbers and converting all numbers to a special token ¡nUm¿, reducing the vocabulary size by approximately 20,000 words.

2.4 Procedures and results

In order to analyze the data, the authors opted for the Skip-Gram variation of Word2vec and we agree with this choice, since this architecture represents uncommon words more accurately compared to CBOW. The study shows that, although no specific chemical information is fed into the algorithm, the resulting word embeddings behave as expected when merged using vector operations such as addiction and subtraction. As an example, the authors represented the words $LiCoO_2$ and $LiMn_2O_4$ as "one-hot vectors". *One-hot vectors* are a 1 × N matrices used to distinguish each word in a vocabulary from every other word in the vocabulary. They consist of 0s in all cells with the exception of a single 1 in a cell used uniquely to identify the word, and are used to avoid higher numbers being considered more important. Using them as inputs for a neural network process with one hidden layer, they found that, for similar materials, the context words in the text are basically the same.

These embeddings not only support context words, but also analogies. The analogies can be discovered by the Word2vec model by finding the closest word to the result of vector operations between embeddings. Then, using the principal component analysis, the authors projected embeddings such as Zi, Cr and Ni, along with their oxides and structure, onto two dimensions. As we said in the previous section, principal component analysis is an example of dimensionality reduction, and aims to reduce the more or less large number of variables describing a data set to a smaller number of latent variables, limiting the loss of information as much as possible.

With this algorithm, it was discovered that the positions of the embeddings in space encode materials science knowledge, with an accuracy close to 50%, and that embeddings of chemical elements are representative of their positions in the periodic table.

This brought an interesting novelty: some materials, despite having relatively high similarities for a spe-

cific word ('thermoelectric'), never appeared together in the same abstract. In the light of the facts, the authors decided to investigate this issue further, which will be discussed in the next paragraph.
The study showed that from about 48000 total compounds, 7663 were found that were present both in our text and in another dataset containing reports of the thermoelectric power factors, but were never mentioned in relation to thermoelectric characteristics in the text, and are therefore useful for prediction purposes. To obtain the predictions, the authors ranked each compounds by the dot product of their normalized output embedding with the word embedding of 'thermoelectric'. Considering the top ten predictions, it turned out that the average maximum power factor is 3.6 times larger than the average of candidate materials and 2.4 times larger than the average of known thermoelectrics.
Subsequently, by comparing the newly formed ranking with the Spearman rank correlation, the authors obtained results that significantly outperformed the dataset of power factors used in the previous paragraph.
To conclude the analysis, attention was finally turned to forecasting ability. It was of interest to see if the model created could predict thermoelectric materials that would be used in future research literature. After generating 18 different corpora of "historical" texts, word embeddings were trained for each of the corpora and used to predict the thermoelectric materials that would most likely be used in future literatures. The results that came to light are very positive: thanks to word embedding-based predictions, top ranking materials obtained were, on average, eight times more likely to have been used as thermoelectrics as compared to a randomly chosen material from our corpus at that time (within the next five years).
Finally, the study showed that several predictions were found to exhibit promising properties despite not being in any well known thermoelectric material classes. Thus, it is possible to deduce that word embeddings do not stop at trivial compositional or structural similarity, but go beyond it. They may therefore be able to unlock latent knowledge that is not directly accessible to human scientists. As a last procedure, the generability of the approach followed by the authors was tested, and very similar results were obtained in all applications.

3 Conclusions

In this paper we reviewed the study by Vahe Tshitoyan & John Dagdelen et al. about retrieving latent knowledge from materials science literature using unsupervised word embeddings.
Despite some doubts, we can say that the authors did a good job, especially in the cleaning and processing part of the dataset. They were able to use different text processing techniques, showing each time the improvements that resulted from their application. In summary, the results obtained are quite significant and may be very promising for future implications in related studies.
However, something is missing. In the process of creating the word embeddings they did not mention

the 'FastText' approach, which is as renowned as the other two. It is a library created by the Facebook Research Team for efficient learning of word representations and sentence classification. A model trained in this way obtains a better morphological understanding of the language on which it's being trained and this form of modeling allows the algorithm to handle out of vocabulary words.

Furthermore, in the section on natural language processing, they only analysed the tokens when it would have been interesting, in our opinion, to use other natural language processing techniques, such as stemming, in order to find out which one would perform better.

Most importantly, however, is that stop words are not mentioned in this study. These are actually the most common words in any language (such as articles, prepositions, pronouns, conjunctions, etc.) and do not add much information to the text. Their removal is essential for saving space and time in processing of large data. On the contrary, by not removing stop words the authors analysed words that have no semantic value for the purpose of the analysis.

In order to improve the study, or with a view to future studies, we recommend treating stop words as a priority, then dealing with other text processing and word embedding techniques, such as those mentioned above.