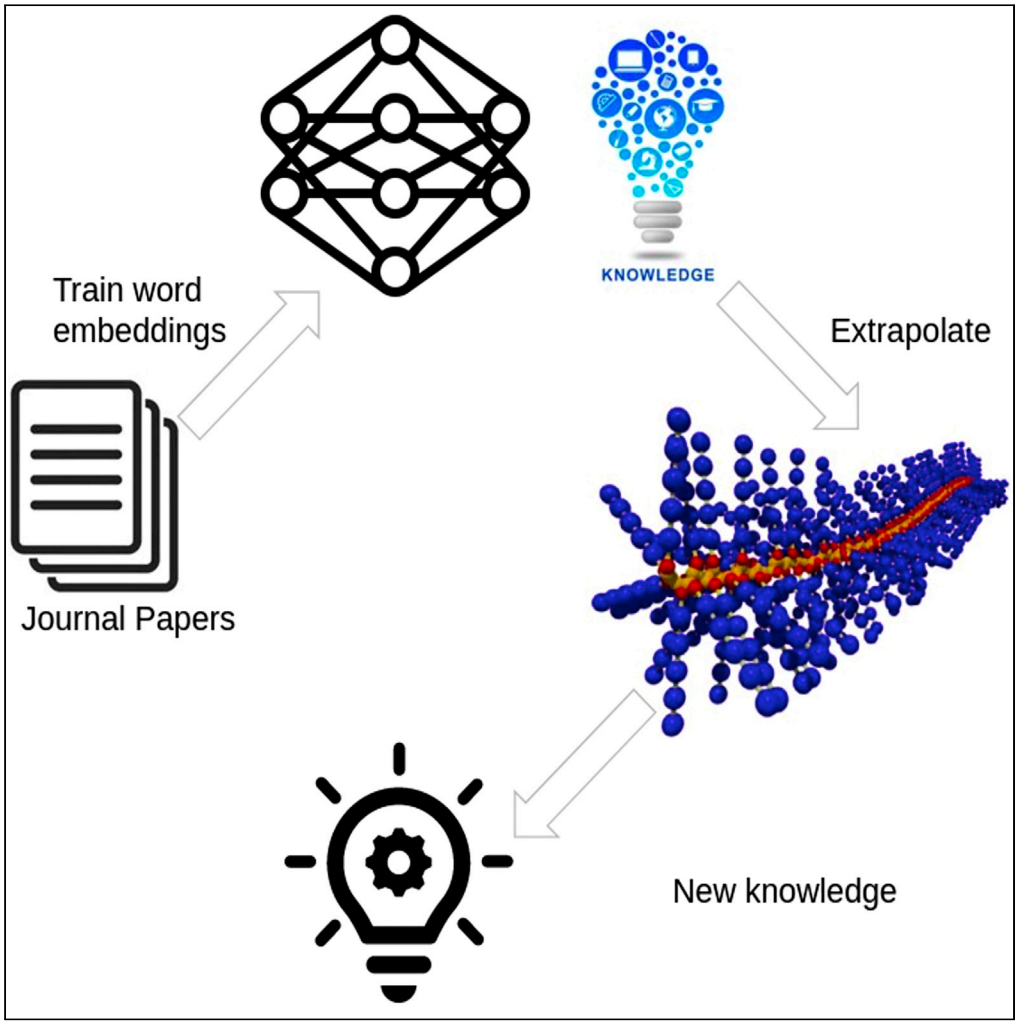
iScience

CellPress
OPEN ACCESS

Article

Automated knowledge extraction from polymer literature using natural language processing



Pranav Shetty,
Rampi Ramprasad

rampi.ramprasad@mse.gatech.edu

Highlights

Word embeddings trained on a corpus of polymer papers

Common polymers and their corresponding applications analyzed over time

Polymer domain knowledge encoded in word vectors

Novel polymers for certain applications predicted and validated using word embeddings

Shetty & Ramprasad, iScience *24*, 101922
January 22, 2021 © 2020
https://doi.org/10.1016/j.isci.2020.101922

iScience

CellPress
OPEN ACCESS

Article

Automated knowledge extraction from polymer literature using natural language processing

Pranav Shetty[1] and Rampi Ramprasad[2,3,*]

Summary

Materials science literature has grown exponentially in recent years making it difficult for individuals to master all of this information. This constrains the formulation of new hypotheses that scientists can come up with. In this work, we explore whether materials science knowledge can be automatically inferred from textual information contained in journal papers. Using a data set of 0.5 million polymer papers, we show, using natural language processing methods that vector representations trained for every word in our corpus can indeed capture this knowledge in a completely unsupervised manner. We perform time-based studies through which we track popularity of various polymers for different applications and predict new polymers for novel applications based solely on the domain knowledge contained in our data set. Using co-relations detected automatically from literature in this manner thus, opens up a new paradigm for materials discovery.

Introduction

The number of papers published annually in all sub-domains of materials science and engineering including in polymer science and engineering has increased exponentially in recent years. It is therefore difficult for any single individual or group of individuals to master the information in the ever evolving landscape of polymer science. This consequently limits inductive reasoning that fully exploits past knowledge and the evolution of new hypotheses. Advances in the field of Natural Language Processing (NLP) (Mikolov et al. (2013)) present a route through which we can represent this domain knowledge in a numerical form, algebraically manipulate it and even formulate new hypotheses and predictions. NLP is a discipline that aims to transform written text to a form that is easy for computers to manipulate (Collobert and Weston (2008)). A basic building block of NLP is the notion of a word vector or word embedding.

A word vector is a dense numerical representation of a word in a high dimensional vector space. Word vectors were popularized by the work of Mikolov (Mikolov et al. (2013)) and have since become ubiquitous in the field of NLP. The idea is to embed in a vector space, the semantic and syntactic information implicit in a word, given its context. These word vectors can then be used as features for building downstream models for tasks relevant in NLP, such as document classification (Lilleberg et al. (2015)). Large corpora of text are used as inputs to models such as Word2Vec (Mikolov et al. (2013)) or Glove (Pennington et al., 2014) and vector representations for words in the corpus is the output. Word2Vec uses the idea that words which occur before and after a word in a sentence, referred to as a context window, capture its meaning. This is illustrated in Figure 1. The phrases "polystyrene undergoes" and "transformation at" form a context window of size 2 around "phase". Word2Vec is trained by predicting the context words given the center word by sliding a context window centered at every word over the entire corpus. This is known as the skip-gram variant of Word2Vec. When trained on a large corpus of text, the Word2Vec model "sees" several contexts in which every word occurs and thus "learns" its meaning in the form of a vector representation.

In order to train word vectors, the text in our corpus has to be tokenized. Tokenization is the process of chopping up a sentence into units which are used for downstream processing (Figure 2). In the simplest case, a sentence may be separated into words, i.e., space separated units and word vectors

[1]School of Computational Science and Engineering, Georgia Institute of Technology, 771 Ferst Drive NW, Atlanta, GA 30332, USA

[2]School of Materials Science and Engineering, Georgia Institute of Technology, 771 Ferst Drive NW, Atlanta, GA 30332, USA

[3]Lead contact

*Correspondence: rampi.ramprasad@mse.gatech.edu

https://doi.org/10.1016/j.isci.2020.101922

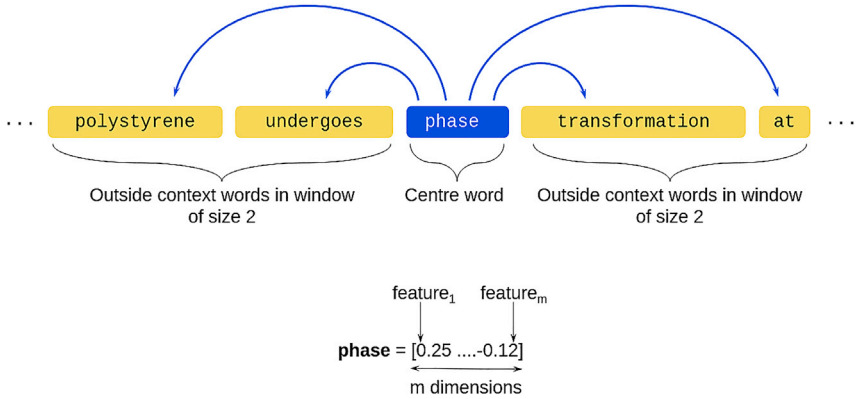


Figure 1. Illustrating context window around a word

In the phrase "polystyrene undergoes transformation at", "phase" is the center word and "polystyrene", "undergoes", "transformation", "at" are its context words. A vector representation for "phase", shown here as an m-dimensional vector, is trained based on all the contexts it is seen in, in the data set, of which this is one example.

can be trained for those. In a polymer corpus, however there will be several chemical named entities (CNEs) (e.g.: poly(lactic acid), poly(butylene succinate)), which should be treated as one unit and other materials science specific terms such as characterization methods (e.g.: scanning electron microscopy) and properties (e.g.: glass transition temperature), which are best treated as phrases for word vector training.

Early work in the Chemistry and Bioinformatics literature to use NLP methods, focused primarily on named entity recognition (NER). NER is a common task in NLP that seeks to identify words in text that correspond to certain pre-defined categories such as locations, organizations, currencies, etc. In a materials science/chemistry context, NER taggers focused on identifying spans of text corresponding to chemical species and more recently have expanded to include categories like synthesis conditions and characterization methods (Weston et al. (2019)). ChemSpot (Rocktäschel et al. (2012)) and ChemDataExtractor (Swain and Cole (2016)) were early works to identify chemical species. The formulaic style of writing found in chemistry papers also gave rise to rule-based tools to parse synthesis information from chemistry papers through the tool ChemTagger (Hawizy et al., 2011).

The application of NLP techniques in materials science to generate insights from text data is more recent. One of the earliest efforts in this direction used NLP techniques to extract synthesis insights for oxide and zeolite systems from literature (Kim et al. (2017); Jensen et al., 2019). More recently, it was shown that word vectors trained on a corpus of materials science abstracts could be used for materials discovery of inorganic thermoelectrics Tshitoyan et al. (2019).

To our knowledge, this is the first work to apply NLP methods to the polymer domain to generate previously unknown materials science insights and inferences. As illustrated in Figure 2, we develop text-mining tools to collect a large corpus of ~ 0.5 million published polymer papers. A treasure trove of insights about polymer science that would otherwise be inaccessible, was inferred from the textual information contained in our corpus. This includes trends such as which polymers and which applications are most commonly mentioned in polymer literature and tracking how the popularity of various applications for polymers has waned and waxed over the years. In addition, we clean-up and parse the data contained in our corpus of papers to train word vector models (Word2Vec and fastText). We show that these word vector representations which were trained in a completely unsupervised manner, can be used to reconstruct known facts about polymer science. This is illustrated vividly by polymer analogies which capture materials science relationships such as monomer-polymer and functional group name and corresponding chemical species. We also show that word vectors for polymers cluster according to their application. This demonstrates that word vectors can capture existing domain knowledge. Finally, by training word vector models on a subset of our data, i.e., up to a certain year, we show that we can predict polymers that were discovered in subsequent years for certain applications. This indicates that the domain knowledge captured in the word vector space can be extrapolated to discover co-relations that were previously unknown.

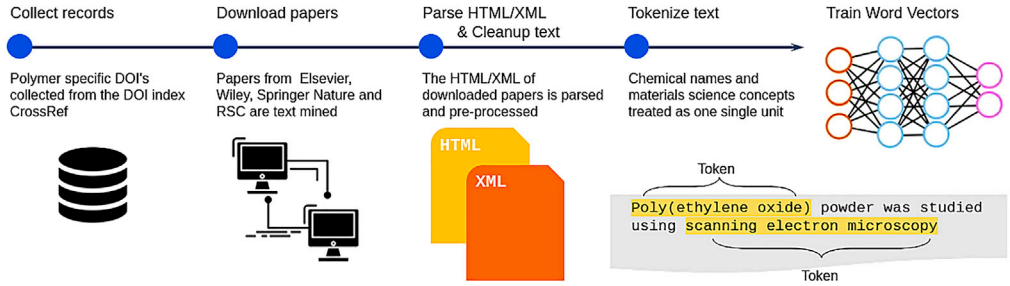


Figure 2. Our workflow begins with collecting records for DOI's from polymer specific journals

The papers corresponding to these DOI's are then downloaded. Plain text is extracted from the HTML/XML text followed by which the text is tokenized using Spacy and ChemDataExtractor. This is the input which is used to train a word vector model and word vectors for every token in our corpus is the output.

Results and discussion

Trends in data

We analyze certain aggregate trends in our corpus of papers, looking at only the frequency with which materials science relevant tokens occur in the corpus. Figure 3A plots the relative frequency of the 10 most common polymer named entities (PNEs) occurring in our data set. Polystyrene is the most frequently occurring polymer followed by polyethylene. It is interesting to note that polystyrene is almost 1.5 times as common in polymer literature as the next most common polymer, i.e., polyethylene. There was no systematic bias toward any particular polymer or application while collecting our data set. Moreover, if we examine the trend of occurrence of the 5 most common polymers over the last 20 years (Figure 3B), this trend has persisted. The relative number of occurrences for each of these polymers has remained stable over the last 20 years.

Figure 3C also looks at the most common application tokens in our data set. The candidate list of applications is the materials science applications dataset by Lawrence Berkeley National Lab mentioned in the Data Processing section (Weston et al. (2019)). The most common token is "membrane" which is more than twice as frequent as the next token which is "fiber". This indicates that the most common application by far for polymers in our data set is as a membrane.

Additionally, Figure 4 analyzes the historical trend of polymers and their corresponding applications. An application is marked as "corresponding" to a polymer CNE if it occurs in a context window of size 4 around it. Similar approaches have been followed in literature to infer relationships between entities from co-occurrence (Crasto et al., 2011; Li et al. (2009)). We validate our approach by randomly selecting polymer-application pairs that appear at least 5 times in our corpus and verify that the corresponding sentence in the paper where the polymer and application co-occur is indeed a true relationship and not a random co-occurrence. For 60 such polymer-application pairs, we found that 50 (83.3%) co-occurrences were true relationships. For each polymer, we consider multiple variations in how a polymer might occur in literature. For the polymers shown in Figure 4, we manually construct a list of variations in polymer name for each of the polymers considered and consider context windows corresponding to each PNE in that list. For instance, for polyethylene we also use the PNE's poly(ethylene), poly ethylene, and poly-ethylene. The rank on the Y axis corresponds to how frequently a particular application co-occurs with a given polymer up to a given year. Thus, the rank 1 application up to any given year would be the most frequently co-occurring application for that polymer. The rank on the Y axis corresponds to how frequently a particular application co-occurs with a particular polymer. Thus, rank 1 in any given year would be the most frequently co-occurring application. Each year on the plot indicates that all papers up to and including that year are considered for computing the rank of an application. The X axis for each plot consists of all applications which were in the top 7 in any of the years in the plot. The applications are ordered according to their rank observed up to 2006 in order to set a baseline for comparison. We consider the 5 most common polymers as per Figure 3A.

Upon inspection several interesting trends are clear from Figure 4. In the case of polypyrrole and polyaniline, the rank of "transistors" for both these polymers has increased from 2006 to the present day. This indicates that both these polymers increasingly co-occur more frequently with applications other than "transistors" over the time period in consideration. In the case of polystyrene, we note that it appears increasingly in the context of coatings

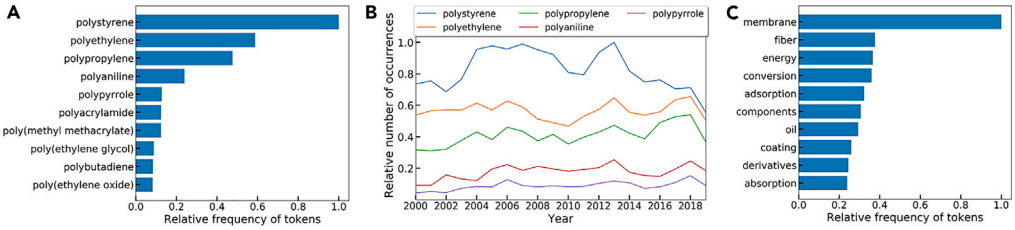


Figure 3. Relative frequencies of polymer and application tokens

(A–C) (A) The relative frequency of different polymers in our data set, (B)The relative frequency of the top 5 polymers in our dataset over a period of time, (C) The relative frequency of different application tokens.

over the same time period. Upon inspection, several interesting trends are clear from Figure 4. In the case of polypyrrole and polyaniline, the rank of the polymer against "transistors" has increased from 2006 to the present day reflecting the waning number of papers in this area. In the case of polystyrene, we note that it appears increasingly in the context of coatings over the same time period. Membranes have consistently been the rank 1 application across all years considered for polyethylene and polypropylene. This is to be expected considering how ubiquitous both these polymers are as separators for lithium ion batteries (Lee et al. (2014)) and other selective separation applications (Tan and Rodrigue (2019)). In a similar vein, we observe that "electrochemical" is the highest ranked application for polyaniline in all years under consideration except 2006. This is to be expected considering the conducting properties of polyaniline (Bijwe et al. (2019)). The other top application listed for polyaniline such as "electrical" and "electroactive" also allude to its conducting nature. Solvents are a dominant co-occurring application for polystyrene and in this case, this clearly refers to solvents which are used for dissolving polystyrene and its variants and not the application of polystyrene as a solvent. This highlights a limitation of context based approaches and human interpretation is still required to disambiguate cases like this.

Validation of word vector model

Polymer analogies

In addition to serving as feature vectors, it is known that word vectors capture intrinsically useful information about the space of words. A prominent example of this would be analogies, shown schematically in Figure 5. Given a relationship of the type

polystyrene: styrene:: polyethylene: ?, word vectors are able to predict the correct answer. Equation 1 is used to compute the result of this analogy.

$$\underset{w \in vocab}{argmax}\ vec(w) \odot \left(vec(\text{styrene}) - vec(\text{polystyrene}) + vec(\text{polyethylene})\right) = 1 - \text{octene} \quad \text{(Equation 1)}$$

vec(w) in Equation 1 denotes the word vector associated with a particular word, w, and $\odot$ indicates the vector dot product. The vector sum *vec*(*polystyrene*) − *vec*(*styrene*) can be thought of as defining a vector subspace from the vector space of all word vectors which encodes polymer-monomer relationships. Observe that conditioned on styrene:polystyrene, using the above equation, the token with the highest dot product is 1-octene which is known to be a commonly used monomer for producing polyethylene Soga et al. (1996). Ethylene is also one of the top 4 predictions. This indicates that a non-trivial correlation is being learned while training the word vectors. This process can be repeated for polymer name-abbreviation pairs. The result is illustrated in Figure 5. This plot was obtained by projecting the Word2Vec word vectors of all CNEs onto 2 dimensions using principal component analysis (PCA). This structure observed in Figure 5 in a lower dimensional linear subspace of the word vector space is indicative of the structure that must exist in higher dimensions.

The intrinsic quality of word vectors can be evaluated using analogies to assess whether known patterns in materials science are being accurately reproduced. We develop a data set of analogies for polymers and consider 5 classes of polymer analogies shown in Table 1. To evaluate analogies, all possible pairs of analogy relationships in the data set are considered. The data set size shown in Table 1 corresponds to the data set with all possible pairs of relationships. Best of 4 accuracy is reported, i.e., if the correct answer is found among the top 4 words which have the highest dot product (Equation 1), then that data point is marked as correctly predicted. Polymer-monomer pairs are poorly predicted. In all the other analogy classes, the 2

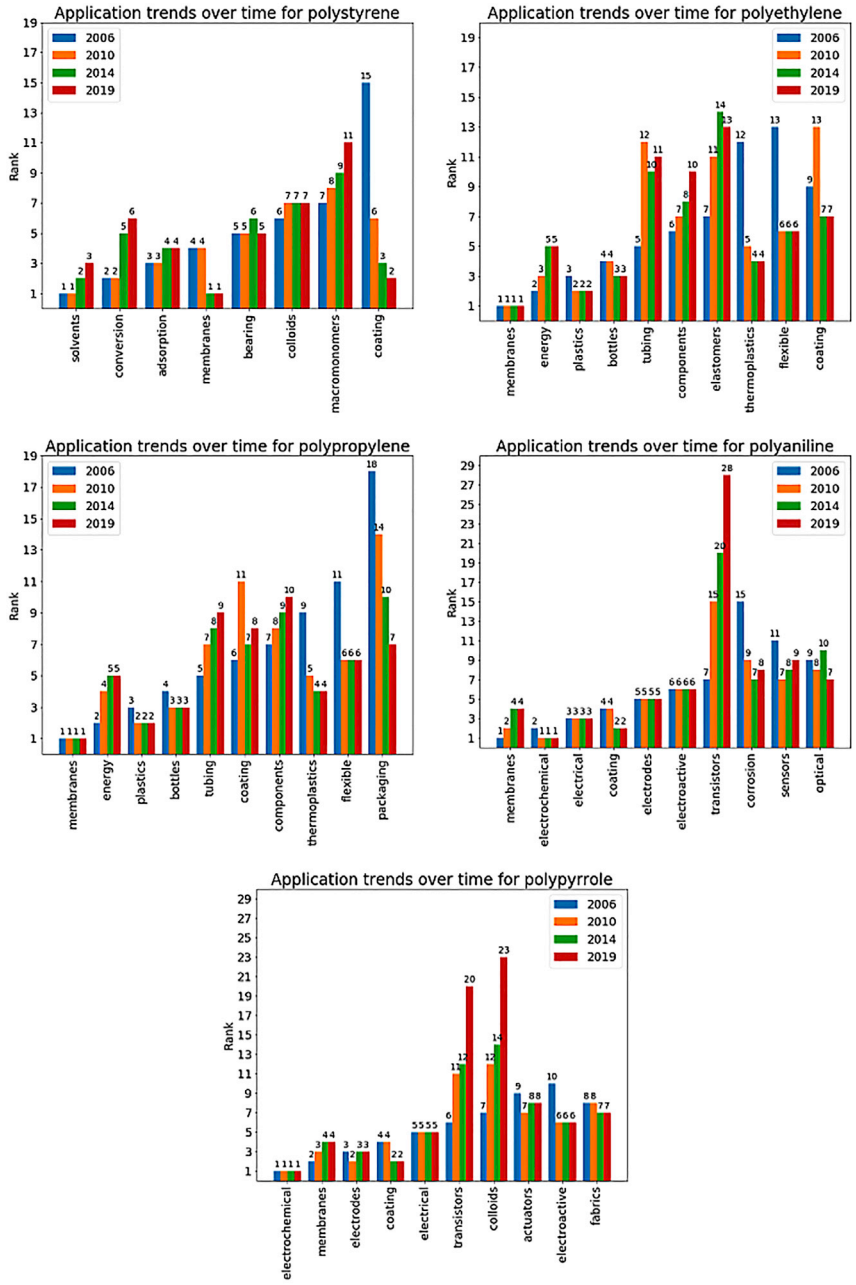


Figure 4. Tracking application trends for polymers

The plots above show the most frequently co-occurring applications for the top 5 polymers in Figure 3. Each application shown was one of the top 7 applications in at least one of the years considered with applications ordered by 2006 ranks. Rank 1 in a particular year corresponds to the top application that year.

entities in the analogy are likely to co-occur in text ex: PE is likely to be written in brackets next to polyethylene making this relationship easier to learn. For monomer-polymer pairs this may not always be the case. Thus, this is a second order relationship which has to be inferred during the training process based on the typical contexts that monomers and polymers occur in. An accuracy level of 42.3% is statistically non-trivial and as is typical with neural network based methods, a larger data set of papers would likely produce much better accuracy results. It should be noted that the morphological structure of the word is not taken into account while training Word2Vec vectors and so polypropylene and propylene have independently trained word vectors. Predicting "polypropylene" is thus not just a matter of figuring out the prefix "poly" but the

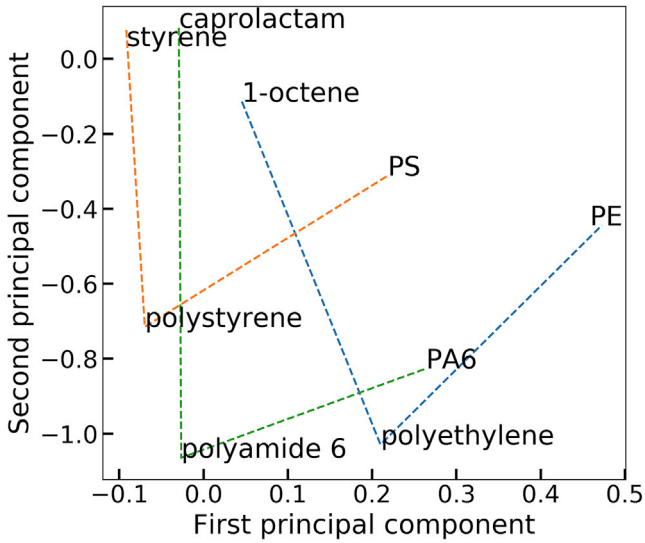


Figure 5. Polymer analogies

Relationships of the form monomer: polymer: abbreviation are captured in this plot. The word vector for each entity in this plot is reduced to 2 dimensions using PCA. The structure observed in 2 dimensions in a linear subspace of the original word vector space is indicative of the structure that must be present in the higher dimensional word vector space.

word in its entirety must be predicted. For the other categories, analogy accuracy is close to 80% indicating that those semantic materials science relationships are being learned by our word vectors. Training a Word2Vec model with phrases allows us to learn representations for distinct MSE keywords like "scanning electron microscopy" which in turn allows the model to learn its relationship with the corresponding abbreviation "SEM". Our model is bench marked against other Word vector models in literature in Table S1.

Polymer CNE embedding

Three hundred dimensional Word2Vec word vectors are reduced to 2 dimensions using t-distributed stochastic neighbor embedding (t-SNE) (Maaten and Hinton (2008)). Of all the tokens in our vocabulary, we identify CNEs which occur in our corpus using ChemDataExtractor. Of these, we look closely only at PNEs. ~ 25,000 PNE's are recovered from our corpus in this manner. The PNE's in our corpus are reduced to 2 dimensions using t-SNE. We find the polymers with the highest dot product with each of the following 4 application categories, i.e., biodegradable, adhesive, conducting polymers and semiconducting polymers and plot the first 35 polymers for each category in Figure 6A. We automatically label a polymer with its application this way. All 4 categories appear as distinct clusters with polymers of similar application clustered together. Moreover, the clusters for conducting and semiconducting polymers which share chemical similarities such as conjugated double bonds appear to be close together. This indicates that our word vectors are learning meaningful chemical trends.

A closeup of the t-SNE embeddings for the category of conducting polymers is shown in Figure 6. For the purpose of showing a clear plot, we show only a subset of the polymers in this category. The polymers in the Figure 6B are indeed conducting polymers. This is likely due to the fact that polymers and their applications often co-occur in text. The reader will note that some CNEs appear multiple times in Figure 6B. An example would be polypyrrole and polypyrole, the latter of which is a misspelling but is commonly found in literature (Ishtiaq et al. (2020); Bello et al. (2016); Khan et al. (2017)). This is because each of these tokens is distinct and has a separate word vector associated with it. In order to have a single word vector for both representations, we would have to carry out named entity normalization, i.e., identify all separate tokens used to refer to the same real world entity. We intend to take up this problem in the future.

Extrapolating to new knowledge and predictions

The performance of word vectors on materials science word analogy tasks, i.e., monomer-polymer relationships and the clustering by application suggests that chemical information is encoded in these word vectors. We test if a model trained on a subset of our data up to a particular year can predict polymers that

Table 1. Performance of the Word2Vec model on material science-specific analogy tasks

Analogy class	Sample analogies	Data set Size	Accuracy (best of 4)
Polymer abbreviations	PE : polyethylene PP : polypropylene PS : polystyrene	3192	85.8%
Monomer-polymer	styrene: polystyrene propylene: polypropylene lactic acid: poly(lactic acid)	1122	42.3%
Functional groups	OH: hydroxyl SH: thiol CH3: methyl	506	82.2%
MSE abbreviations	DFT: density functional theory LC: liquid crystal XRD: X-ray diffraction	992	77.7%
Chemical formulas	H2SO4: sulfuric acid KBr: potassium bromide CO2: carbon dioxide	2352	88.8%

occur in the context of a new application in subsequent years. Predictions are made by taking the dot product of the word vector associated with a particular application with a candidate list of polymers. The top 250 polymers are taken as the predictions of our model (1% of the candidate list). Three applications are selected for investigation denoted by the keywords "membranes", "electrodes" and "magnetic". The candidate list of polymers is not normalized and we manually remove duplicate polymers which occur in our validated set of predicted polymers. A caveat to note, is that our predictions point to polymers that are likely to occur in the context of an application and can be used for that particular application in a number of different ways. For instance a polymer that occurs in the context of the term magnetic need not itself be magnetic but might be used as say a coating material for magnetic particles. Similarly a polymer occurring in the context of "membranes" may be used directly as a membrane or may be used in a formulation that is used to make a membrane. Context-based prediction can be thought of as a coarse filter that narrows the set of materials we should focus on. Our method can only predict polymers that may be appropriate for a new application that have already occurred in literature in the time window considered but within the context of other applications.

Our results are shown in Figure 7. The number of polymers that were "discovered"by our model, is plotted as a function of the year up to which a word vector model was trained. "Discovered" here, refers to polymers that were predicted by the model trained up to a particular year despite never co-occurring with that application, and co-occurred in subsequent years. We consider predictions using fastText and Word2Vec embeddings for these 3 applications. These two embedding methods are considered as they represent different schools of thought in how word embeddings are constructed. Word2Vec has a vector representation for each polymer token while fastText constructs its vector representation from sub-word information. The results in Figure 7 show that there is no major difference between these 2 methods across the application categories considered with the exception of membranes. Incorporating sub-word information appears to be beneficial in predicting new membranes. In general, the set of polymers predicted by both these methods are different (Refer Tables S2–S4). Observe that there is a downward trend in the number of polymers predicted with time. This is because the set of polymers yet to be discovered becomes smaller over time and despite the increased size of the model, there are not that many polymers that are left to be predicted. The polymers predicted in earlier years are in general supersets of polymers predicted in later years, falling out of the predicted list once they are "discovered".

In order to validate our model, we use randomly generated embeddings of unit norm of the same dimension as the word vector as a baseline. A different set of random embeddings is generated for each year considered in Figure 7 for the polymers and applications. We make predictions for the random embeddings in the same way as Word2Vec and fastText. From Figure 7, it is clear that the word vector models outperform randomly generated word embeddings. This indicates that the model is learning meaningful correlations.

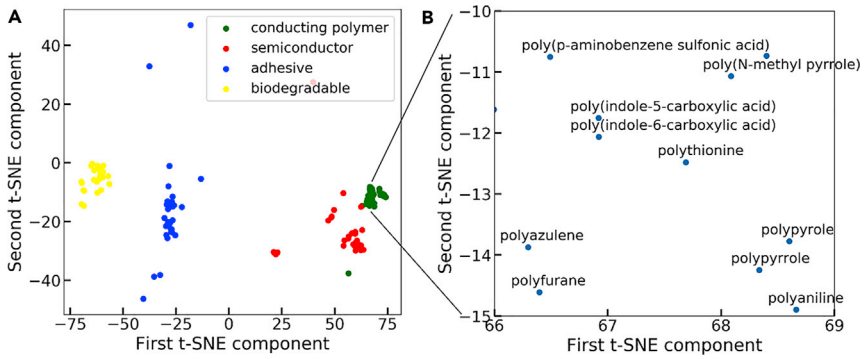


Figure 6. Clustering of polymer word vectors by application

(A and B) (A) Word vectors for polymers associated with four common applications projected in 2 dimensions using t-SNE, (B) Close-up of region corresponding to conducting polymers.

In the next section, we look at some specific case studies and explore the reason why the model is able to make these predictions.

Illustrative examples

A complete list of predicted polymers can be found in Tables S2–S4. To take one particular example, the polymer polyrhodanine is a predicted polymer in the context of "magnetic" by the fastText model trained up to 2010 data. Before 2010, polyrhodanine is mentioned as an anti-microbial agent (Kong et al. (2009)). As an anti-microbial agent, metal ions adsorbed in the polymer matrix are released when a bacterium makes contact with the polyrhodanine-metal polymer matrix leading to cell death of the bacterium. Before 2006, there are several polymers such as poly(HEMA) (Punyani and Singh (2008); Horák et al., 2000), polyurethanes (Francolini et al. (2006); Schmidt (2006)) and poly(MMA) (Sayar et al. (2006); Patel et al. (2004)) which co-occur in the context of both "antimicrobial" and "magnetic". Polyrhodanine is, in fact used as an encapsulating polymer for Fe_2O_3 magnetic nanoparticles which is first reported in our data set in 2015 (El-Sonbati et al. (2016)). In this case too, polyrhodanine is adsorbed on the surface of the metal. Thus, these two seemingly unrelated applications have the polymer's ability to adsorb metal in common, which the model "figures" out.

To take another example, polysulfobetaine is predicted by the fastText model trained up to 2006 data in the context of "membranes". This polymer does not occur in the context of membranes before 2006. It is used to modify the surface of cellulose membranes starting from 2013 in our data set (Yuan et al. (2013); Wang et al. (2015)). In the cited references, polysulfobetaine was used to modify the surface of cellulose membranes in order to confer it with antibiofouling properties. Sulfobetaine was polymerized on the surface of the cellulose membrane using reversible addition-fragmentation chain transfer (RAFT) polymerization. Before 2006, polysulfobetaine is mentioned as a polymer that undergoes RAFT polymerization to form copolymers (Donovan et al. (2003)) which form two phase solutions with water (Ali et al. (2003)) and are used as stimuli-responsive polymers for water remediation and fluid modification. There are other references to water-insoluble polymers like polystyrene forming graft polymers on cellulose via RAFT (Hernández-Guerrero et al. (2005)) before 2006. The model thus "joined the dots" and predicted polysulfobetaine as a relevant polymer in the context of "membranes".

Summary and outlook

This is the first work to use NLP methods in the polymer domain, establishing the use of NLP methods to augment the field of polymer informatics (Kim et al. (2018)). A corpus of ~ 0.5 million polymer journal papers has been collected, which has allowed us to examine trends in polymer literature by simple frequency and context window based approaches. We show that word vectors trained using this corpus in an unsupervised manner are able to encode materials science knowledge. This was demonstrated through polymer analogies which showed the different kinds of materials science knowledge embedded in our word vector space and through embedding polymer word vectors in 2 dimensions which also showed that meaningful chemical trends are encoded in this vector space. The polymer domain word vector space so generated can be used to predict new polymers for particular applications. This illuminates a new pathway through which materials discovery may take place. In addition to the traditional methods of experimental and computationally driven materials

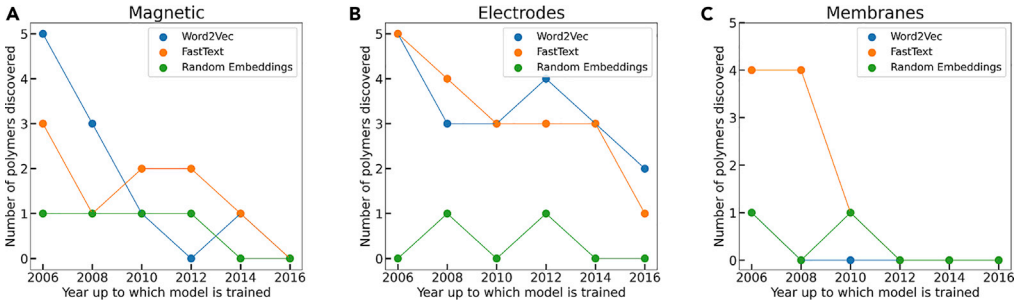


Figure 7. Polymers predicted by the word vector model

Polymers are predicted by taking dot product of a polymer's word vector and that of an application and considering the top polymers. Of the predicted polymers, each validated polymer plotted above, did not co-occur in the years up to which the word vector model was trained but was "discovered" in subsequent years.

(A–C) correspond to the keywords "magnetic", "electrodes" and "membranes", respectively. Randomly generated embeddings are used as a baseline.

discovery, we see that new applications for known materials can be found automatically by extrapolating the domain knowledge contained in literature using NLP methods.

In addition to building vector spaces of the domain knowledge contained in our corpus, the textual information contained therein can be used to parse polymer properties such as glass transition temperature, dielectric constants etc. Building databases of these properties, enables building better models that can predict polymer properties (Kim et al. (2018)) and in turn be used to design new polymers (Sharma et al., 2014). These are the future directions we intend to investigate.

Limitations of the study

Application of NLP methods to polymers present certain unique challenges. Although IUPAC naming conventions do exist for polymers, common names of polymers are frequently used. These names are not standardized and may vary such as "polyethylene", "poly ethylene" and "poly-ethylene". A bottleneck we faced in this work faced while "predicting" new polymers was that polymer names had to be normalized manually, i.e. duplicate names after taking dot product had to be discarded. Normalization of PNE's to create a list of known unique polymers will be addressed in future work.

Resource availability

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Rampi Ramprasad (rampi.ramprasad@mse.gatech.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The Word2Vec model is uploaded at Figshare: https://doi.org/10.6084/m9.figshare.13211015 and code to access the same has been uploaded to https://github.com/Ramprasad-Group/polymer_knowledge_extraction. In addition, these can also be accessed from khazana.gatech.edu.

Methods

All methods can be found in the accompanying Transparent Methods supplemental file.

Supplemental information

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.101922.

Acknowledgment

This work was supported by the Office of Naval Research through grants N00014-19-1-2103 and N00014-20-1-2175. The authors also acknowledge useful discussions and feedback from Dr. Lihua Chen and Dr. Chiho Kim. We are also grateful for the suggestions provided by the reviewers.

Authors contribution

Conceptualization, R.R. and P.S.; Methodology, P.S.; Investigation, P.S.; Writing – Original Draft, P.S.; Writing – Review & Editing, P.S. and R.R.; Funding Acquisition, R.R.; Resources, R.R.; Supervision, R.R.

Declaration of interests

The authors declare no competing interests.

Received: September 28, 2020
Revised: November 12, 2020
Accepted: December 5, 2020
Published: January 22, 2021

References

Ali, S.A., Al-Muallem, H.A., and Mazumder, M.A. (2003). Synthesis and solution properties of a new sulfobetaine/sulfur dioxide copolymer and its use in aqueous two-phase polymer systems. Polymer *44*, 1671–1679.

Bello, A., Barzegar, F., Madito, M., Momodu, D.Y., Khaleed, A.A., Masikhwa, T., Dangbegnon, J.K., and Manyala, N. (2016). Stability studies of polypyrole- derived carbon based symmetric supercapacitor via potentiostatic floating test. Electrochim. Acta *213*, 107–114.

Bijwe, D., Yawale, S., Kumbharkhane, A., Peng, H., Yawale, D., and Yawale, S. (2019). Complex dielectric behavior of doped polyaniline conducting polymer at microwave frequencies using time domain reflectometry. Rev. Mex. Fís. *65*, 590–600.

Collobert, R.; Weston, J. 2008 A unified architecture for natural language processing: Deep neural networks with multitask learning. Proceedings of the 25th international conference on Machine learning. 2008; pp 160–167.

Crasto, C., Luo, D., Yu, F., Forero, A., and Chen, D. (2011). GenDrux: a biomedical literature search system to identify gene expression-based drug sensitivity in breast cancer. BMC Med. Inform. Decis. Making *11*, 28.

Donovan, M.S., Lowe, A.B., Sanford, T.A., and McCormick, C.L. (2003). Sulfobetaine-containing diblock and triblock copolymers via reversible addition-fragmentation chain transfer polymerization in aqueous media. J. Polym. Sci. A Polym. Chem. *41*, 1262–1281.

El-Sonbati, A., Diab, M., El-Bindary, A., and Mossalam, A. (2016). Polymer complex LXIV: Coordination chemistry of some rhodanine polymer complexes. J. Mol. Liquids *216*, 797–807.

Francolini, I., Ruggeri, V., Martinelli, A., D'Ilario, L., and Piozzi, A. (2006). Novel Metal-Polyurethane Complexes with Enhanced Antimicrobial Activity. Macromolecular Rapid Commun. *27*, 233–237.

Hawizy, L., Jessop, D.M., Adams, N., and Murray-Rust, P. (2011). ChemicalTagger: A tool for semantic text-mining in chemistry. J. Cheminformatics *3*, 17.

Hernández-Guerrero, M., Davis, T.P., Barner-Kowollik, C., and Stenzel, M.H. (2005). Polystyrene comb polymers built on cellulose or poly(styrene-co-2-hydroxyethylmethacrylate) backbones as substrates for the preparation of structured honeycomb films. Eur. Polym. J. *41*, 2264–2277.

Horák, D., Boháček, J., and Šubrt, M. (2000). Magnetic poly (2-hydroxyethyl methacrylate-co-ethylene dimethacrylate) microspheres by dispersion polymerization. J. Polym. Sci. A Polym. Chem. *38*, 1161–1171.

Ishtiaq, F., Bhatti, H.N., Khan, A., Iqbal, M., and Kausar, A. (2020). Polypyrole, polyaniline and sodium alginate biocomposites and adsorption-desorption efficiency for imidacloprid insecticide. Int. J. Biol. Macromolecules *147*, 217–232.

Jensen, Z., Kim, E., Kwon, S., Gani, T.Z., Romaì n-Leshkov, Y., Moliner, M., Corma, A., and Olivetti, E. (2019). Machine Learning Approach to Zeolite Synthesis Enabled by Automatic Literature Data Extraction. ACS Cent. Sci. *5*, 892–899.

Khan, A.A.P., Khan, A., Rahman, M.M., Asiri, A.M., and Oves, M. (2017). Sensor development of 1,2 Dichlorobenzene based on polypyrole/Cu-doped ZnO (PPY/CZO) nanocomposite embedded silver electrode and their antimicrobial studies. Int. J. Biol. Macromolecules *98*, 256–267.

Kim, E., Huang, K., Saunders, A., McCallum, A., Ceder, G., and Olivetti, E. (2017). Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. Chem. Mater. *29*, 9436–9444.

Kim, C., Chandrasekaran, A., Huan, T.D., Das, D., and Ramprasad, R. (2018). Polymer Genome: A Data-Powered Polymer Informatics Platform for Property Predictions. J. Phys. Chem. C *122*, 17575–17585.

Kong, H., Song, J., and Jang, J. (2009). One-Step Preparation of Antimicrobial Polyrhodanine Nanotubes with Silver Nanoparticles. Macromolecular Rapid Commun. *30*, 1350–1355.

Lee, H., Yanilmaz, M., Toprakci, O., Fu, K., and Zhang, X. (2014). A review of recent developments in membrane separators for rechargeable lithium-ion batteries. Energy Environ. Sci. *7*, 3857–3886.

Li, J., Zhu, X., and Chen, J.Y. (2009). Building Disease-Specific Drug-Protein Connectivity Maps from Molecular Interaction Networks and PubMed Abstracts. PLoS Comput. Biol. *5*, e1000450.

Lilleberg, J.; Zhu, Y.; Zhang, Y. 2015 Support vector machines and word2vec for text classification with semantic features. 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC). 2015; pp 136–140.

Maaten, L.v. d., and Hinton, G. (2008). Visualizing data using t-SNE. J. Machine Learn. Res. *9*, 2579–2605.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process. Syst. 3111–3119.

Patel, M.V., Patel, S.A., Ray, A., and Patel, R.M. (2004). J. Polym. Sci. A Polym. Chem. *42*, 5227–5234.

Pennington, J., Socher, R., and Manning, C.G. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543.

Punyani, S., and Singh, H. (2008). Synthesis, characterization, and antimicrobial properties of novel quaternary amine methacrylate copolymers. J. Appl. Polym. Sci. *107*, 2861–2870.

Rocktäschel, T., Weidlich, M., and Leser, U. (2012). ChemSpot: a hybrid system for chemical named entity recognition. Bioinformatics *28*, 1633–1640.

Sayar, F., Güven, G., and Pişkin, E. (2006). Magnetically loaded poly(methyl methacrylate-co-acrylic acid) nano-particles. Colloid Polym. Sci. *284*, 965.

Schmidt, A.M. (2006). Electromagnetic Activation of Shape Memory Polymer Networks Containing Magnetic Nanoparticles. Macromolecular Rapid Commun. *27*, 1168–1172.

Sharma, V., Wang, C., Lorenzini, R., Ma, R., Zhu, Q., Sinkovits, D., Pilania, G., Oganov, A., Kumar, S., Sotzing, G., et al. (2014). Rational design of all organic polymer dielectrics. Nature Communications *5*, 1–8.

Soga, K., Uozumi, T., Nakamura, S., Toneri, T., Teranishi, T., Sano, T., Arai, T., and Shiono, T. (1996). Structures of polyethylene and copolymers of ethylene with 1-octene and oligoethylene produced with the Cp2ZrCl2 and [(C5Me4)SiMe2N(t-Bu)]TiCl2 catalysts. Macromolecular Chem. Phys. *197*, 4237–4251.

Swain, M.C., and Cole, J.M. (2016). ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. J. Chem. Inf. Model. *56*, 1894–1904.

Tan, X., and Rodrigue, D. (2019). A Review on Porous Polymeric Membrane Preparation. Part II: Production Techniques with Polyethylene, Polydimethylsiloxane, Polypropylene, Polyimide, and Polytetrafluoroethylene. Polymers *11*, 1310.

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K.A., Ceder, G., and Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. Nature *571*, 95.

Wang, P., Meng, J., Xu, M., Yuan, T., Yang, N., Sun, T., Zhang, Y., Feng, X., and Cheng, B. (2015). A simple but efficient zwitterionization method towards cellulose membrane with superior antifouling property and biocompatibility. J. Membr. Sci. *492*, 547–558.

Weston, L., Tshitoyan, V., Dagdelen, J., Kononova, O., Trewartha, A., Persson, K.A., Ceder, G., and Jain, A. (2019). Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature. J. Chem. Inf. Model. *59*, 3692–3702.

Yuan, J., Huang, X., Li, P., Li, L., and Shen, J. (2013). Surface-initiated RAFT polymerization of sulfobetaine from cellulose membranes to improve hemocompatibility and antibiofouling property. Polym. Chem. *4*, 5074–5085.

iScience, Volume *24*

Supplemental Information

Automated knowledge extraction from polymer literature using natural language processing

Pranav Shetty and Rampi Ramprasad

1. BENCHMARKING MODEL PERFORMANCE

In Table S1 we compare the performance of our model against other Word2Vec models reported in literature. The metric used is the performance on a set of grammatical analogies consisting of $\sim$ 15,000 analogy pairs. This consists of analogies such as plurals i.e., jacket: jackets:: person: people ; opposites i.e., tall: short :: hot: cold and so on. We report the best of 1 performance here to be consistent with reported benchmarks elsewhere. All three models are clearly of similar performance by this metric.

Table S1. Benchmarking analogy performance, Related to Table 1

Work	Dataset used	Grammar analogies accuracy
This work	Corpus of polymer papers ($\sim$ 0.5 million documents)	60.9 %
Tshitoyan et al. (2019)	Inorganic papers abstract ($\sim$ 3 million documents)	61.6 %
Mikolov et al. (2013)	Large Newspaper dataset (1 billion words)	61.0 %

2. VALIDATED POLYMERS

Table S2. Validated polymers corresponding to 'magnetic', Related to Figure 7

Year	Word2Vec	fastText
2006	polydiphenylamine, poly(styrene-co-divinylbenzene) poly(N-vinylimidazole), poly(divinylbenzene) poly(NIPAAm-co-AA)	polyoxotungstates, polypyridyl polyoxomolybdate
2008	poly(divinylbenzene), polydiphenylamine poly(NIPAAm-co-AA)	polyoxotungstate
2010	poly(NIPAAm-co-AA)	polyrhodanine, polyoxotungstate
2012	-	polyrhodanine, polyoxotungstate
2014	polyrhodanine	polyrhodanine
2016	-	-

Table S3. Validated polymers corresponding to 'electrodes', Related to Figure 7

Year	Word2Vec	fastText
2006	poly(2-methoxyaniline), poly(vinylferrocene) poly(9-vinylcarbazole), poly(5-amino-1-naphthol) polyacetate	polyazomethine, poly(3-methoxythiophene) polybenzimidazole, polypyridine polyyne
2008	poly(vinylferrocene), poly(5-amino-1-naphthol) polyacetate	polypyridine, polybenzimidazoles polyviologen, polyyne
2010	poly(vinylferrocene), poly(5-amino-1-naphthol) polyacetate	polypyridine, polyviologen polyyne
2012	poly(thionine), polyviologen poly(5-amino-1-naphthol) polyacetate	polypyridine, polyviologen polyyne
2014	poly(thionine), poly(5-amino-1-naphthol) polyacetate	polyporphine, polyviologen polyyne
2016	poly(thionine), polyacetate	polyyne

Table S4. Validated polymers corresponding to 'membranes', Related to Figure 7

Year	Word2Vec	fastText
2006	polyvinylimidazole	polybetaine, polyetheramines polyprenol, polysulfobetaine
2008	-	polybetaine, polyetheramines, polytriazoles, polysulfobetaine
2010	-	polysulfobetaine
2012	-	-
2014	-	-
2016	-	-

3. TRANSPARENT METHODS

A. Data acquisition

The workflow followed in this work, starting from data acquisition to training word vectors is illustrated in Fig. 2 in the main paper. Polymer relevant DOI's are searched for from the DOI index Crossref (Lammey (2015)) by using polymer specific keywords and by retrieving all records for papers from polymer relevant journals. The papers corresponding to those records from publishers Elsevier, Wiley, Springer Nature and Royal Society of Chemistry are automatically downloaded. The first three publishers have Text and data mining policies which allow text mining at subscribing institutions. Permission was obtained from Royal Society of Chemistry for text mining. In case of Elsevier, the ScienceDirect API (https://dev.elsevier.com/) was used. We restrict ourselves to HTML/XML version of documents as these are easier to parse compared to pdf files. As HTML/XML files for scientific papers are mostly available only after 2000, we also restrict ourselves to papers published after 2000. Our corpus consists of $\sim$ 0.5 million documents.

B. Data Processing

Plain text is extracted from the HTML/XML documents in our corpus and pre-processed to remove references and numeric data as this would have increased the vocabulary size for word vector training making it more computationally expensive. Numeric information is replaced with the dummy token _NUM_. The abstract as well as the body of the paper is used for training the model.

We use the python library Spacy (https://github.com/explosion/spaCy) for English language tokenization. ChemDataExtractor (Swain & Cole (2016)) is used to identify and tokenize Chemical Named entities (CNE). In particular, polymers are identified out of the set of chemical named entities by the presence of the string 'poly' , referred to as polymer named entities (PNE). This is imperfect and misses out some polymers like cellulose and will include CNE's like 'polysulfide' but this heuristic appears to work well in practice. For other materials science entities, a dataset of materials science named entities released by Lawrence Berkeley National Lab (Weston et al. (2019)) is used. Consecutive space separated words which match entities present in this dataset are merged into a single token. This dataset was obtained based on models that were trained on inorganic materials but contains named entities which are common to inorganic materials as well as polymers. We use the named entities in this dataset corresponding to properties, applications and characterization methods. This amounts to 34,200 materials science named entities.

C. Word Vector Training

We use the python library Gensim (https://radimrehurek.com/gensim/) for training word vectors. 300 dimensional Skip-Gram (a variant of Word2Vec) and fastText word vectors are trained using a minimum count of 5 occurrences in our corpus, i.e., a word vector is generated only if a particular token occurs at least 5 times in the corpus. We use a context window of size 8 and train the model for 10 epochs. In case of Word2Vec, we use negative sampling with n=5. Word2Vec generates a vector representation for every token in the corpus. In contrast, fastText (Bojanowski et al. (2017)), uses subword information in every token and adds up the vector representation for each subword to obtain the vector representation for a word. For instance, the vector representation for 'polyethylene' would be the sum of the subword vector representations associated with 'poly', 'eth' and 'ylene'.

D. Dimensionality reduction

We use two different methods of dimensionality reduction in this work. We use standard scikit-learn (https://scikit-learn.org/) functions for both.

Principal Component Analysis (PCA): PCA (Ringnér (2008)) is used for determining the directions within a high dimensional vector space (containing some data points) which explain the greatest variance among the data points in that vector space. This is done by computing the co-variance matrix of the data points and finding the eigenvectors corresponding to the largest eigenvalues. Thus PCA enables us to see certain patterns in the data.

T-distributed Stochastic Neighbor Embedding (t-SNE): t-SNE (Maaten & Hinton (2008)) preserves distance with high probability when projecting from a high dimensional space to a low dimensional space. t-SNE is better able to capture non-linearity in high dimensional spaces and preserves clustering information with high probability. t-SNE is more computationally intensive than PCA. Thus the 300 dimensional vectors are reduced first to 50 dimensions using PCA and then t-SNE is used to reduce that to 2 components. We use a perplexity of 30, early exaggeration of 12 and 10,000 iterations to generate the t-SNE embeddings.

REFERENCES

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2017), 'Enriching word vectors with subword information', *Transactions of the Association for Computational Linguistics* **5**, 135–146.

Lammey, R. (2015), 'Crossref text and data mining services', *Science Editing* **2**(1), 22–27.

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., Chen, K., Corrado, G. S. & Dean, J. (2013), Distributed representations of words and phrases and their compositionality, *in* 'Advances in neural information processing systems', pp. 3111–3119.

Ringnér, M. (2008), 'What is principal component analysis?', *Nature biotechnology* **26**(3), 303–304.

Swain, M. C. & Cole, J. M. (2016), 'Chemdataextractor: a toolkit for automated extraction of chemical information from the scientific literature', *Journal of chemical information and modeling* **56**(10), 1894–1904.

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G. & Jain, A. (2019), 'Unsupervised word embeddings capture latent knowledge from materials science literature', *Nature* **571**(7763), 95.

Weston, L., Tshitoyan, V., Dagdelen, J., Kononova, O., Trewartha, A., Persson, K. A., Ceder, G. & Jain, A. (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.