MaterialBERT for Natural Language Processing of Materials Science Texts

Michiko Yoshitake[a]*, Fumitaka Sato[a,b], Hiroyuki Kawano[a,b] and Hiroshi Teraoka[a,b]

[a]*MaDIS, National Institute for Material Science, Tsukuba, Japan;* [b]*Ridgelinez, Tokyo, Japan*

1-1, Namiki, Tuskuba, Ibaraki, Japan 305-0044, yoshitake.michiko@nims.go.jp

MaterialBERT for natural language processing of materials science texts

A BERT (Bidirectional Encoder Representations from Transformers) model, which we named “MaterialBERT,” has been generated using scientific papers in wide area of material science as a corpus. A new vocabulary list for tokenizer was generated using material science corpus. Two BERT models with different vocabulary lists for the tokenizer, one with the original one made by Google and the other newly made by the authors, were generated. Word vectors embedded during the pre-training with the two MaterialBERT models reasonably reflect the meanings of materials names in material-class clustering and in the relationship between base materials and their compounds or derivatives for not only inorganic materials but also organic materials and organometallic compounds. Fine-tuning with CoLA (The Corpus of Linguistic Acceptability) using the pre-trained MaterialBERT showed a higher score than the original BERT.

Keywords: word embedding; pre-training; BERT; literal information

Subject classification codes: Databases, data structure, ontology

1. Introduction

Informatics techniques have been extensively utilized in the business and industrial fields [1-3]. In material science fields, machine learning of numerical data such as composition, electrical conductivity, reflective index, solubility, and friction coefficient, and that of processing data such as process temperature and pressure, have increasingly attracting attention [4-6]. In addition to numerical data, literature data, such as comments on SNS (Social Networking Service) and customer claims have been vigorously analysed with informatics techniques in business fields [7-10]. Informatics techniques on such literature data given in natural languages are called natural language processing (NLP) techniques; they have explosively developed and are applied in social business fields because of the huge data available from web sites and SNS. Here, to

apply machine learning techniques to natural language, characters or words are converted to numerical data, usually to high-dimensional vectors; this is called embedding. Among the many ways of conversion, Word2Vec [11] attracted sensational attention since it demonstrated that the embedding reflects the meaning of a word. Word2Vec is a simple 1-layer neural network, which does not require many computer resources. Many embeddings by Word2Vec method using corpora from different fields, such as Japanese language, materials science, and bioscience, were made. Embeddings using a corpus from materials science papers, especially focused on inorganic materials, have been made named Mat2Vec [12]. Among scientific abstracts in materials science taken from Elsevier's Scopus, Science Direct API, and the Springer Nature API, abstracts relevant to inorganic materials science were selected and used as a corpus in Mat2Vec. The successful embedding of meanings from materials science viewpoint was demonstrated [12].

Natural language is data with sequence, and the sequence of words is highly important. Therefore, NLP techniques basically use recurrent neural networks (RNNs) with embedded words. Word2Vec is a technique for embedding, which uses words surrounding a target word so that the context is taken into consideration to some extent, but the sequence of words is not considered. Advanced RNN techniques suitable for NLP, such as bidirectional LSTM (Long Short Term Memory) [13] have been developed, however, complicated RNN-based methods require excessive computational resources. Epoch-making methods to simplify the RNN network, transformer, attention, and BERT, have been developed [14]. BERT model is revolutionary because after pre-training (predicting a randomly masked word in two sequential sentences), fine-tuning for many tasks such as given in General Language Understanding Evaluation (GLUE) [15] can be trained with a small dataset. Examples of tasks in GLUE are Q&A,

paraphrasing, implicational relation between two sentences, grammatical correctness (CoLA), and sentiment judgment. Because of this feature of BERT, it can be used in various applications. The original BERT used a dictionary that contained 30M token vocabulary and the pre-training corpus consisted of the BooksCorpus (800M words) [16] and English Wikipedia (2,500M words). The corpus used contained general words that are not specified in a certain area. Therefore, many models using the BERT algorithm with a corpus from specific fields have been constructed such as BioBERT (bio-medical) [17], MedBERT [18], SciBERT (bio science 82% + computer science 18%) [19], Japanese BERT [20,21], FinBERT (financial) [22], LeagalBERT [23].

A BERT model specific to wide area of materials science (inorganic, organic, composite, metal-organic, etc.) was desired for our work to produce a kind of knowledge graph on material property relationships [24, 25]. Therefore, we started generating a BERT model specific to 'wide area of materials science' (MaterialBERT) and reported at a conference [26]. At the moment, we pre-trained using an original BERT except a corpus, which were scientific articles in materials science journals. However, despite huge technical terms specific to a materials science filed, the original vocabulary list released with the original BERT ("vocab.txt" file) contains only very general ones because it was made from the corpus used to pre-train the original BERT. Therefore, we built a vocabulary list specific to materials science from scientific articles in materials science journals and started generating another MaterialBERT using the newly made vocabulary list. Meanwhile, MatSciBERT [27], which is a kind of transfer learning of SciBERT using scientific papers in inorganic materials field (inorganic classes and ceramics, bulk metallic glasses, alloys, and cement and concrete) was posted. Then, MatBERT [28] was posted, which is a variant pre-training BERT in inorganic materials field (both solid state dataset and doping dataset were taken from

inorganic materials science and gold nanoparticle dataset). Both MatSciBERT and MatBERT are considered domain-specific to “inorganic materials science”.

It was reported [29] that there were no significant differences among BioBERT, SciBERT and MatSciBERT for their sentence classification task of polymer science texts, which is out of inorganic material science. Therefore, it would be useful to generate models specific to materials science in general, not limited to inorganic materials science. Moreover, recently, materials, which cannot be classified by traditional material classes such as inorganic or organic materials, have emerged (composite materials, perovskite solar cell materials, metal organic frameworks, etc.). Due to this situation, not only for our work on knowledge graph, a BERT model that is domain-specific to “wide materials science” could be useful for material-class-interdisciplinary works. If one focuses on phenomena such as fracture and refraction, the scientific principles of the phenomena is common among all classes of materials. In many materials R&D, researchers search materials that satisfy a specific functional characteristic which is based on the corresponding phenomena. Especially in the era of SDGs (Sustainable Development Goals), the replacement of current functional materials with those better fit SDGs is required. Such replacement often occurs beyond the traditional material classes. Furthermore, our MaterialBERT could be used as a starting point for generating a narrower domain-specific BERT model in materials science field by transfer learning.

2. Method

We downloaded and used the original BERT code to train MaterialBERT on our corpus with the same configuration and size as BERT-Base-uncased (12-layer, hidden layer dimension=768, Total Parameters = 110M) [14]. Sentence lengths up to 512

tokens were used for pre-training. In addition to the difference of a corpus from the original BERT, a variation in vocabulary list was made. One vocabulary list is the same as that the original BERT used ("vocab.txt file in the github [30], we refer to Original Vocab). The other vocabulary list was made in the following way: first, a vocabulary list was made in the same way as the authors of SciBERT [19] did except the vocabulary size, where the vocabulary list was made during the training of a tokenizer with SentencePiece [31] using our material science corpus. Then, this vocabulary list was added to the original BERT vocabulary list (vocab.txt) and used as a second vocabulary list (we refer to Sentence Vocab). Sentence Vocab contains material-specific words such as bond‐containing, radiation‐absorbed, isothermal, mesoporosity, chromatography, amide‐, acetate‐methanol, alkaline‐metal, α‐methyl‐α‐phenyl, etc. Two MaterialBERT were generated, one with Original Vocab and the other with Sentence Vocab, both with the architecture as the original BERT and with our materials science corpus. The Original Vocab contains about 30 K words and Sentence Vocab contains 140 K words. The embedded words vectors had 200 dimensions.

The corpus we used was taken from scientific articles our institute (NIMS) purchased in XML format from nine publishers (ACS, AIP, APS, ELSEVIER, IOP, JJAP, RSC, SPRINGER, WILEY), and most of them were published between 2005 to 2019. Our corpus contains scientific articles not only in inorganic materials but also in organic materials and composite materials. It also includes articles from journals that offer physical and/or chemical basis to phenomena in materials science (often cited in articles on a material papers). The list of the names of the journals, ISSNs and publication years used is provided in the appendix. Materials Science is a very board field and expanding further year by year. Therefore, the authors did not feel reasonable to use established criteria for choosing articles. Rather the authors rely on the decision

of each journal (manuscripts that are not the criteria of the journal are not accepted). We confirmed that the journals listed in the appendix are materials science related and used all published articles within the specified journal, since BERT need huge corpus. We exclude articles that contained only abstracts (without the main body). Approximately 750,000 articles were included in this study. Only abstract and body sections from article texts were extracted as a cleansing process because parts such as affiliation, acknowledgement, and references become noise in the NLP in our case. Chemical formulae and mathematical expressions (they are not natural language) in the articles were eliminated from the article texts for pre-training. The estimated number of words for approximately 750,000 articles was roughly 3000 M, which is comparable to the original BERT. Each model was trained on two NVIDIA Tesla V100 GPUs and took about three months to complete.

3. Results and Discussion

3.1. Pre-training

3.1.1. Learning curves

Figure 1 shows the learning curve during the pre-training. Learning using the original vocabulary list (Original Vocab) for the tokenizer is shown in (a), and that using the vocabulary list made from our corpus (Sentence Vocab) is shown in (b). Because the size of the Sentence Vocab (140M words) is more than four times larger than the Original Vocab (30M words), the time required for one iteration for (b) is much longer and the iteration end is taken for a much smaller iteration of 143,000 (b) instead of 410,000 (a). Because of the smaller number of iterations, the final loss was larger for (b). If the iterations continued until the numbers were similar to (a), the final loss for (b) would be similar to that of (a).

3.1.2. Embedding of meaning

The results of the evaluation of word embeddings are presented below. The 200-dimension word vectors of material names were subject of principal component analysis and projected onto a plane with two main components. The results of two sets of word vectors embedded using the two different dictionaries were compared.

3.1.2.1 Clustering of materials

Names of materials such as iron, aluminum, silicon, zinc selenide, zinc oxide, boron nitride, polystyrene, polyvinyl chloride were used for the analysis. Material names such as micelle, supramolecule, which are not classified in usual material classes, were also included as “others”. Words used are listed in Table1 with a class assigned by clustering. The clustering of word vectors of different types of material names is shown in Fig. 2. The word vectors make well-separated clusters according to well-established material classes, such as metals, semiconductors, and polymers [32, 33, 34]. The positions of the clusters themselves do not have a meaning and depend on the vocabulary list used for the tokenizer. This shows that words are well-embedded in both MaterialBERT models constructed using the Original Vocab (Fig. 2a) and Sentence Vocab (Fig.2b).

3.1.2.2. Inorganic materials

Word vectors for four typical elements, and their oxides, carbides, and chlorides were subject to principal component analysis, and the vectors were projected onto a plane with two main components. The results are shown in Fig. 3. For both models using different dictionaries, elements, oxides, carbides, and chlorides formed clusters. Accordingly, the vectors of oxide formation (oxide of), carbide formation (carbide of), and chloride formation (chloride of) are similar for all four elements. There is a slight

difference in the oxide formation vectors between (a) and (b). However, as the vectors are well separated, the difference is not meaningful.

To examine more elements, word vectors for aluminum, calcium, iron, lithium, magnesium, molybdenum, nickel, silicon, sodium, tantalum, titanium, zinc, and zirconium and their oxides, carbides, and chlorides were also analysed in the same way as described above and shown in Fig. 4. For both MaterialBERT models, elements, oxides, carbides, and chlorides formed clusters, as shown in Fig. 3.

3.1.2.3. Organic materials

Word vectors of names of organic compounds were analysed using the principal component analysis method. The vectors of organic compounds with different functional groups, alkanes, carboxylic acids, and amines are plotted in Fig. 5. The vectors of decane, ethane, heptane, hexane, octane, pentane, and propane, as well as their carboxylic acid derivatives and amine derivatives are plotted. Similar to inorganic compounds, different functional groups form a cluster with each other, and changes in the functional groups for the above seven alkanes can be represented as similar vectors, although the variance is larger than with inorganic materials, possibly because of a large number of similar names in organic compounds in various papers used as a corpus.

3.1.1..3 Organometallics

In Fig. 6, word vectors of organometallics are plotted after principal component analysis for R-metal-carbonyl (acetylcobalt tetracarbonyl, acetylmanganese pentacarbonyl, benzene chromium tricarbonyl, butadiene iron tricarbonyl, dicobalt octarbonyl, dimanganese decacarbonyl, ethyl cobalt tetracarbonyl, hexamethyl benzene chromium tricarbonyl, hexamethylborazine chromium tricarbonyl, methyl manganese

pentacarbonyl), alkyl-metal (diethylmagnesium, diethylzinc, dimethyl cadmium, dimethyl mercury, dimethyl zinc, methylcopper, tetramethyltin, trimethylgallium, triphenylgallium), and R-lithium (benzyl-lithium, butyl-lithium, ethyl-lithium, methyl-lithium, phenyl-lithium, vinyl-lithium), where R is an abbreviation for any group in which a hydrocarbon chain is attached to the rest of the molecule. Here, for alkyl-metal, "metal" is not lithium but magnesium, cadmium, mercury, zinc, copper, tin, and gallium. The scattering of vectors is similar to that of organic materials in Fig. 5, suggesting that the word embeddings with meanings as reasonable as in organic materials are achieved for inorganic-organic complex compounds. Despite a vast variety of materials in organometallics, various R and various metals are possible, listing the names of organometallics appearing in scientific papers (in the corpus) is difficult. Therefore, only a limited number of organometallic compounds were used for the evaluation.

3.2. Fine-tuning

Among GLUE, only CoLA [35] (grammatical correctness of sentences) can be used for the evaluation of MaterialBERT fine-tuning, because grammar does not depend on a specific field but others do depend on fields of texts used for the evaluation. Therefore, fine-tuning was preformed using CoLA. The score of the MaterialBERT model with the original vocabulary list (Original Vocab) was 62.5 %, and that with the newly made vocabulary list from our corpus (Sentence Vocab) was 66.2%, which is much higher than the score of the original $\text{BERT}_{\text{BASE}}$ (corresponding to our model) 52.1 % [14]. The score of the original $\text{BERT}_{\text{LARGE}}$ (deeper neural network used) was reported 60.5 % [14], which is still lower than both MaterialBERTs. It is unknown why MaterialBERTs showed higher score with CoLA, which is nothing to do with materials science. One

speculation is that the quality of the corpus used for the pre-training in our corpus, scientific articles were collected from selected scientific journals, which means that the articles are English-corrected and peer-reviewed so that the grammatical correctness of the sentences is high. However, there is no method to characterize a corpus and a evaluation dataset and to measure a kind of distance between them. It is difficult to specify the reason of the higher score.

Various different domain-specific BERTs have been generated since fine-tuning results are supposedly related to the overlap of the domain of corpus used for pre-training and that of the evaluation dataset. Results of fine-tuning using datasets and tasks of author's pick-up are often given as examples, but they do not logically indicate that users would obtain the similar score for their tasks with their datasets. Possibly due to this, FinBERT does not give the score of fine-tuning results of their tasks but offers web-based fine-tuning for sentiment predictions of uploaded users' text [38].

In materials science domain, MatSciBERT and MatBERT, both being pre-trained using corpuses that are domain-specific to materials (in close examination materials out of inorganic materials are not included), used inorganic materials datasets for evaluations [28, 36, 37]. MatSciBERT [27] reported approximately 8% better results on glass vs. non-glass topics classification task using in-house dataset (not disclosed) with their MatSciBERT than SciBERT. On the other hand, for sentence classification tasks of polymer science texts, no differences among BioBERT, SciBERT and MatSciBERT was reported [29], although MatSciBERT having material texts as a corpus is expected to have some advantages over BioBERT and SciBERT. With the development of tools such as HuggingFace Transformer [39], pre-training models begin to be used by users who want to do some text-mining tasks of their interests but are familiar to neither NLP nor machine learning. In such new circumstances, there are risks that high scores in

authors' fine-tuning examples give misleading information to users that high scores should be obtained by the model for users' tasks with users' datasets, which is not guaranteed.

With the above reasons, the authors intend to let users assess the fine-tuning effects for their specific tasks by making the present MaterialBERT models publicly available upon the publication of this article.

MaterialBERT should be useful for material science domains out of inorganic materials, and especially for NLP tasks that handle items regardless of material types such as inorganic, organic, or composite. Furthermore, MaterialBERT could be used as a starting point for transfer learning to generate a narrower domain-specific BERT model in materials science field such as "phase diagram", "fracture", "liquid crystal", "plasma", etc.

4. Conclusions

Pre-trained BERT models with wide range of materials science corpus have been successfully developed using the architecture of the original BERT. A new vocabulary list has been made from materials science corpus. Two MaterialBERT models were generated: one with the vocabulary list that the original BERT used and the other with the newly made vocabulary list. It was shown for both MaterialBERT models that word vectors embedded during the pre-training reasonably reflect the meanings of materials names in material-class clustering and in the relationship between base materials and their compounds or derivatives for not only inorganic materials but also organic materials and organometallic compounds. Fine-tuning using CoLA (sentence classification by grammatical correctness) marked a score much higher than the original

BERT, which would reflect the grammatical quality of the corpus used for MaterialBERT models.

The developed MaterialBERT models cover wide range of materials science, not only inorganic materials. Because of this wideness, an appropriate evaluation of fine-tuning from a viewpoint of material science is impossible due to the lack of suitable evaluation datasets. However, there is no comparable pre-trained BERT model for widely covered materials science. Furthermore, MaterialBERT models can be used as a starting point for transfer learning to generate a narrower domain-specific BERT model in materials science field such as “phase diagram”, “resin”, “liquid crystal”, etc. Because results on fine-tuning are strongly depend on the similarity between a corpus used for the pre-training and that for fine-tuning, the authors intend to let users assess the fine-tuning effects for their specific tasks by making the present MaterialBERT models publicly available upon the publication of this article. The models and the newly developed vocabulary list will be uploaded to the material data repository at NIMS [40] upon the publication of this article so that all users can use it freely.

Acknowledgements

The authors thank the NIMS TDM platform for supplying well-organized XML files from the publishers.

References:

[1] List of universities offering degrees in “business informatics”. Available from: https://en.everybodywiki.com/List_of_universities_offering_degrees_in_business_informatics

[2] BizNews. What is business informatics? Available from: https://biznewske.com/what-is-business-informatics/

[3] A journal with title “industrial informatics”. IEEE Transactions on Industrial Informatics. ISSN: 1551-3203.

[4] Ramprasad R, Batra R, Pilania G, et al. Machine learning in materials informatics: recent applications and prospects. npj Comput Mater. 2017;3:54 (1-13). https://doi.org/10.1038/s41524-017-0056-5

[5] Agrawal A, Choudhary A, Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science. APL Materials. 2016; 4: 053208 (1-10). https://doi.org/10.1063/1.4946894

[6] Tanaka F, Sato H, Yoshii N, et al. Materials Informatics for Process and Material Co-Optimization. IEEE Transactions on Semiconductor Manufacturing. 2019;32:444-449.

[7] Hassan AUl, Hussain J, Hussain M, et al. Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression. Proceedings of 2017 International Conference on Information and Communication Technology Convergence (ICTC); 2017 Oct 18-20; Jeju, South Korea. IEEE; 2017.

[8] Yoshida S, Kitazono J, Ozawa S, et al. Sentiment analysis for various SNS media using Naïve Bayes classifier and its application to flaming detection. Proceedings of 2014 IEEE Symposium on Computational Intelligence in Big Data (CIBD); 2014 Dec. 9-12; Orlando, FL, USA. IEEE; 2015.

[9] Ahn H, Lee S. An Analytic Study on Private SNS for Bonding Social Networking. In Meiselwitz, G. (eds) Social Computing and Social Media. SCSM 2015. Lecture Notes in Computer Science(), vol 9182. Springer, Cham. https://doi.org/10.1007/978-3-319-20367-6_12

[10] Khairi SSM, Ghani RAM. Analysis of social networking sites on academic performance among university students: A PLS-SEM approach. AIP Conference Proceedings. 2019; 2138, 050015. Available from: https://doi.org/10.1063/1.5121120

[11] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems 26 (NIPS 2013). Available from: https://papers.nips.cc/paper/2013.

[12] Tshitoyan V, Dagdelen J, Weston L, et al. Unsupervised word embeddings capture latent knowledge from materials science literature. Nature. 2019;571:95–98.

[13] Long short-term memory. Available from: https://en.wikipedia.org/wiki/Long_short-term_memory

[14] Devlin J, Chang MW, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. Available from: https://arxiv.org/pdf/1810.04805.pdf.

[15] Wang A, Singh A, Michael J, et al. GLUE: A multi-task benchmark and analysis platform for natural language understanding. Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, 2018, 353–355, Available from: https://aclanthology.org/W18-5446, https://gluebenchmark.com/

[16] Zhu Y, Kiros R, Zemel R, et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: Proceedings of the 2015 IEEE International Conference on Computer Vision(ICCV); 2015 Dec 7-13; Santiago, Chile. IEEE; 2015. p.19–27.

[17] Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining; 2019 Oct 31. arXiv:1901.08746. Available from: https://arxiv.org/abs/1901.08746

[18] Rasmy L, Xiang Y, Xie Z, Cui Tao, et al. Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction; 2020 May 22. arXiv:2005.12833v1. Available from: https://doi.org/10.48550/arXiv.2005.12833

[19] Beltagy I, Lo K, Cohan A. SCIBERT: A pretrained language model for scientific text. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019. p. 3615–3620, DOI:10.18653/v1/D19-1371. Available from: https://arxiv.org/abs/1903.10676

[20] BERT Japanese Pretrained Model. Available from: https://nlp.ist.i.kyoto-u.ac.jp/?ku_bert_japanese, https://laboro.ai/activity/column/engineer/laboro-bert/.

[21] Pretrained Japanese BERT models. Available from: https://github.com/cl-tohoku/bert-japanese

[22] Yang Y, Christopher M, Siy UYet al. FinBERT: A pretrained language model for financial communications. Available from: https://arxiv.org/abs/2006.08097

[23] Chalkidis I, Fergadiotis M, Malakasiotis P, et al. LEGAL-BERT: The Muppets straight out of Law School. arXiv:2010.02559v1, 2020 Oct 6. Available from: https://doi.org/10.48550/arXiv.2010.02559

[24] Yoshitake M, Kuwajima I, Yagyu S, et al. System for Searching Relationship among Physical Properties for Materials Curation™. Vac. Surf. Sci. 2018;61:200–205.

[25] Yoshitake, M. Tool for Designing Breakthrough Discovery in Materials Science. Materials 2021;14:6946(1-15). Available from: https://doi.org/10.3390/ma14226946

[26] Yoshitake M, Sato F, Kawano H, et al. MaterialBERT for Natural Language Processing of Materials Science Texts. Paper presented at: 68th JSAP Spring Meeting; 2021 Mar 16-19; On line.

[27] Gupta T, Zaki M. Krishnan ANM, et al. MatSciBERT: A Materials Domain Language Model for Text Mining and Information Extraction. arXiv:2109.15290v1, 2021 Sep 30. Available from: https://doi.org/10.48550/arXiv.2109.15290

[28] Walker N, Trewartha A, Huo H,aoyan, et al. The Impact of Domain-Specific Pre-Training on Named Entity Recognition Tasks in Materials Science. Available from SSRN: https://ssrn.com/abstract=3950755 or http://dx.doi.org/10.2139/ssrn.3950755

[29] Oka H, Ishii M, Sentence classification for polymer data extraction from scientific articles. Poster session presented at: 69th JSAP Spring Meeting; 2022 Mar 22-26; Sagamihara, Kanagawa.

[30] when one down load BERT-base from https://github.com/google-research/bert, vocab.txt file is included in the zip file

[31] Kudo T, Richardson J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; p. 66-71. 2018 Oct 31-Nov 4, Brussels, Belgium. Association for Computational Linguistics. Available from: https://github.com/google/sentencepiece

[32] Classes of Materials, University of Cambridge, https://www.doitpoms.ac.uk/tlplib/artefact/classes.php

[33] Introduction to Materials Science and Engineering, University of Washington USA, Prof. Christine Luscombe, http://courses.washington.edu/mse170/powerpoint/luscombe/Week1complete.pdf

[34] “semiconductor” is relatively new class of materials as mentioned in Materials science, https://en.wikipedia.org/wiki/Materials_science and in Materials science and engineering: https://en.wikiversity.org/wiki/Portal:Materials_science_and_engineering

[35] Warstadt A, Singh A, Bowman SR. Neural network acceptability judgments. Available from: https://arxiv.org/abs/1805.12471, https://nyu-mll.github.io/CoLA/

[36] Weston L, Tshitoyan V, Dagdelen J, et al. Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature. J. Chem. Inf. Model., 2019;59:3692–3702.

[37] Friedrich A, Adel H, Tomazic F, et al. The SOFC-Exp Corpus and Neural Approaches to Information Extraction in the Materials Science Domain. Proceedings of the 58th annual meeting of the association for computational linguistics, Association for Computational Linguistics, Online, 2020, pp. 1255–1268.

[38] ProsusAI / finBERT, https://github.com/ProsusAI/finBERT

[39] huggingface / transformers, https://github.com/huggingface/transformers

[40] https://doi.org/10.48505/nims.3705

Table 1. List of words used for material class clustering with a class assigned by clustering.

Figure 1. Learning curve for pre-training with the original dictionary (a) and the newly made dictionary from our corpus (b).

Figure 2. Material class captured by word embeddings: two-dimensional projection of the word vectors in the plane with the first and second principal components for 79 materials from different material classes using the original dictionary (a) and the newly made dictionary from our corpus (b). Others are materials such as metal-organic framework and composite material.

Figure 3. Word embeddings for magnesium, aluminum, silicon, iron, their principal oxides, carbides and chlorides projected onto two dimensions using principal component analysis and represented as points in space. (a) is obtained using the original dictionary and (b) using the newly made dictionary from our corpus. The projected space between (a) and (b) is slightly different but in both space the relative positioning of the words encodes materials science relationships, such that there exist consistent vector operations between words that represent concepts such as ‘oxide of’, ‘carbide of’ and ‘chloride of’.

Figure 4. Word embeddings for 13 elements (lithium, sodium, magnesium, aluminium, silicon, calcium, titanium, iron, nickel, zinc, zirconium, molybdenum, tantalum, and their principal oxides, carbides and chlorides projected onto two dimensions using principal component analysis and represented as points in space. (a) is obtained using the original dictionary and (b) using the newly made dictionary from our corpus.

Figure 5. Word embeddings for 7 alkanes, and their carboxylic acid, and amine derivatives projected onto two dimensions using principal component analysis and represented as points in space. (a) is obtained using the original dictionary and (b) using the newly made dictionary from our corpus.

Figure 6. Word embeddings for organometallics (R-metal-carbonyl, alkyl-metal, and R-lithium, where R means an abbreviation for any group in which a hydrocarbon chain is attached to the rest of the molecule) projected onto two dimensions using principal

component analysis and represented as points in space. (a) is obtained using the original dictionary and (b) using the newly made dictionary from our corpus.

Table 1. List of words used for material class clustering with a class assigned by clustering.

metals	ceramics	semiconductors	polymers	others
iron	aluminum oxide	silicon	polyethylene	metal complex
aluminum	silicon carbide	germanium	polypropylene	metal organic framework
copper	tungsten carbide	gallium arsenide	polystyrene	composite material
titanium	Yttria-stabilized zirconia	gallium phosphide	polyvinyl chloride	clathrate
gold	zinc oxide	indium phosphide	synthetic rubber	methane hydrate
platinum	zirconia	silicon carbide	phenol formaldehyde resin	supramolecule
chromium	boron nitride	zinc selenide	neoprene	crown ether
nickel	Sialon	cadmium sulfide	nylon	cyclodextrin
cobalt	silicon nitride	gallium nitride	polyacrylonitrile	liposome
tungsten	titanium carbide	gallium oxide	PVB	micelle
palladium	glass	diamond	cellulose	
steel	barium	black phosphorus	starch	
high-speed steel	titanate	fullerene	chitin	
superalloys	hydroxyapatite	carbon nanotube	protein	
inconel	ferrite		lignin	
duralumin	calcium fluoride		silicone	
bronze			celluloid	
amalgam				
alumel				
chromel				
intermetallics				
intermetallic compound				
metallic glass				

Figure 1. Learning curve for pre-training with the original dictionary (a) and the newly made dictionary from our corpus (b).

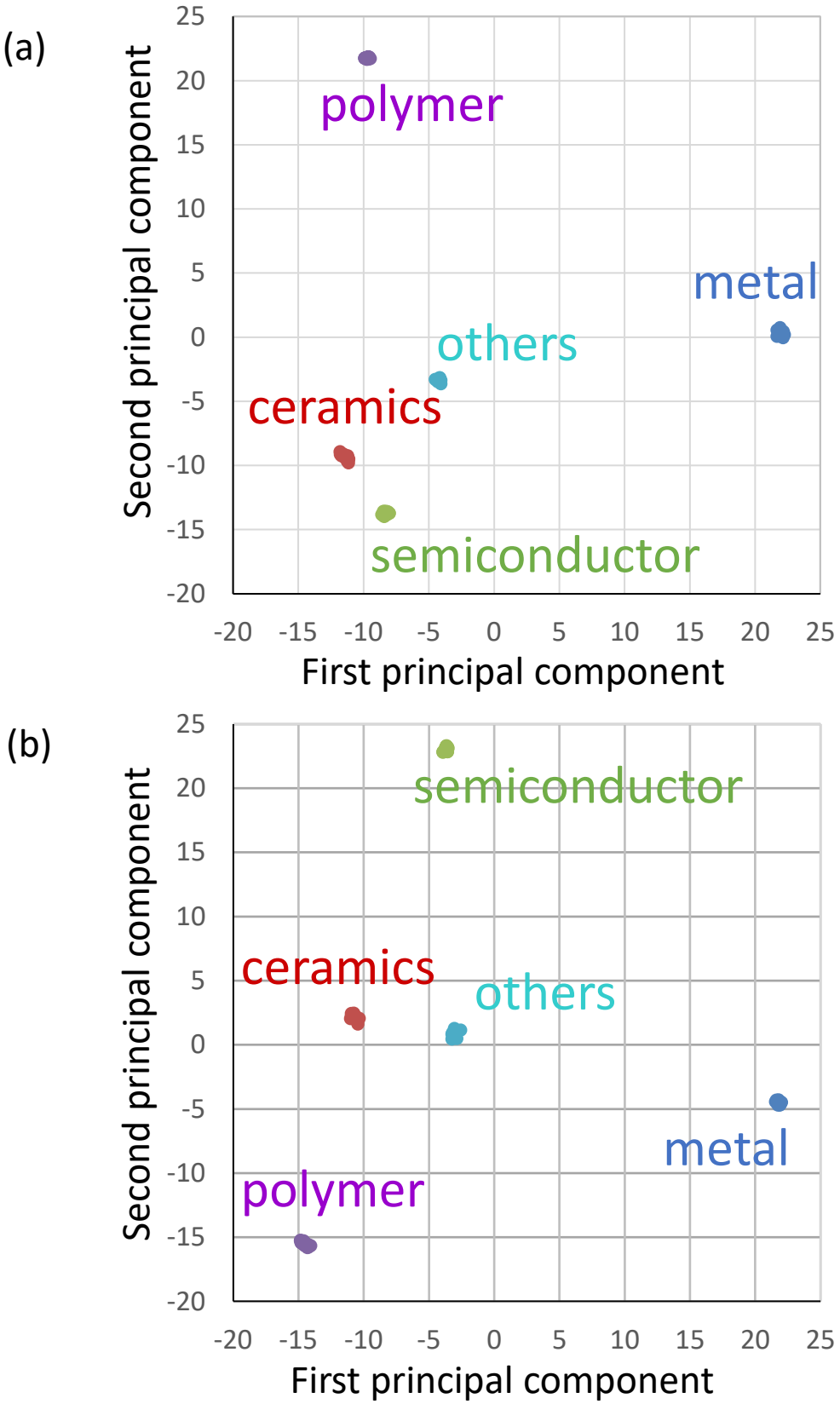

Figure 2. Material class captured by word embeddings: two-dimensional projection of the word vectors in the plane with the first and second principal components for 79 materials from different material classes using the original dictionary (a) and the newly made dictionary from our corpus (b). Others are materials such as metal-organic framework and composite material.

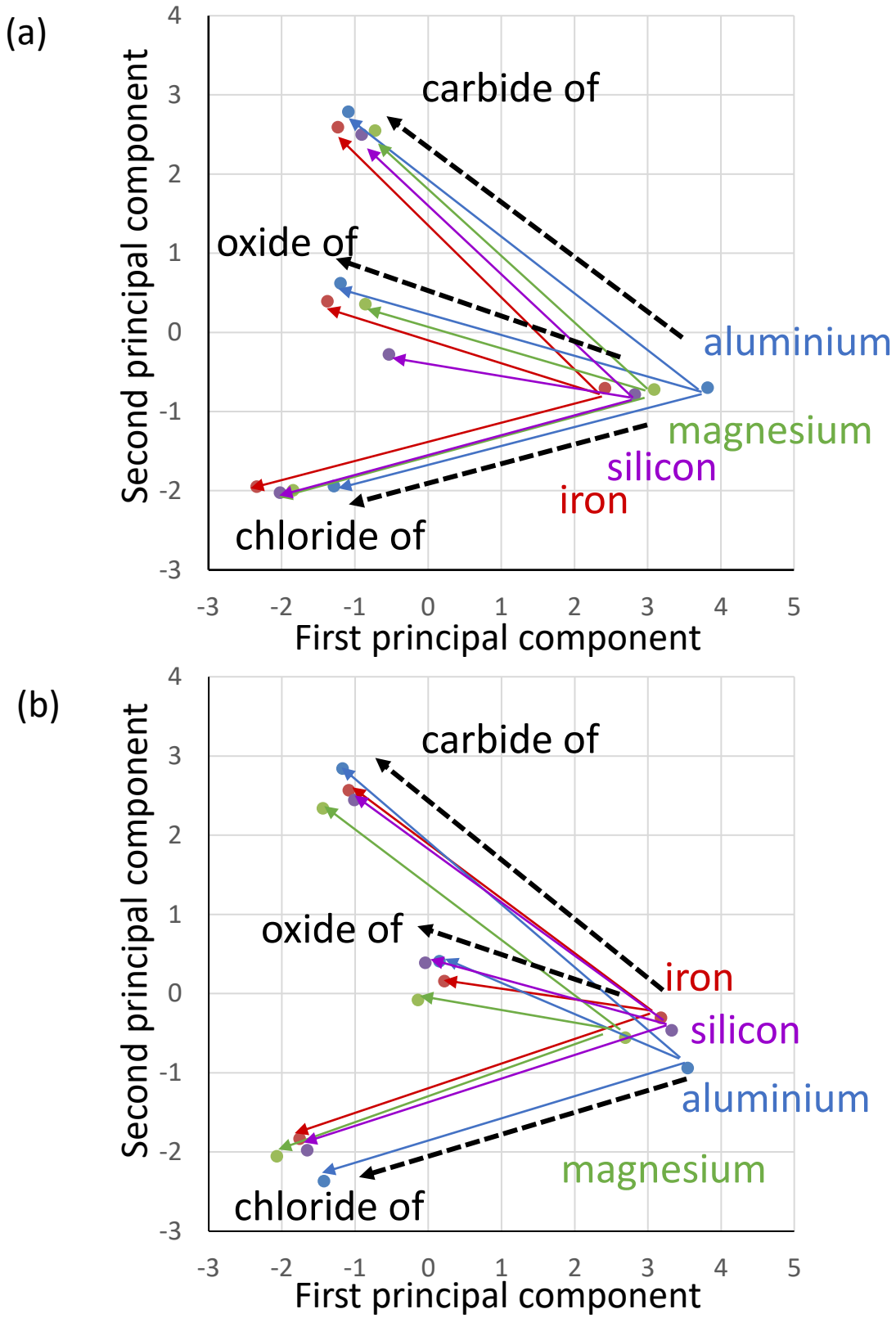

Figure 3. Word embeddings for magnesium, aluminum, silicon, iron, their principal oxides, carbides and chlorides projected onto two dimensions using principal component analysis and represented as points in space. (a) is obtained using the original dictionary and (b) using the newly made dictionary from our corpus. The projected space between (a) and (b) is slightly different but in both space the relative positioning of the words encodes materials science relationships, such that there exist consistent vector operations between words that represent concepts such as ‘oxide of’, ‘carbide of’ and ‘chloride of’.

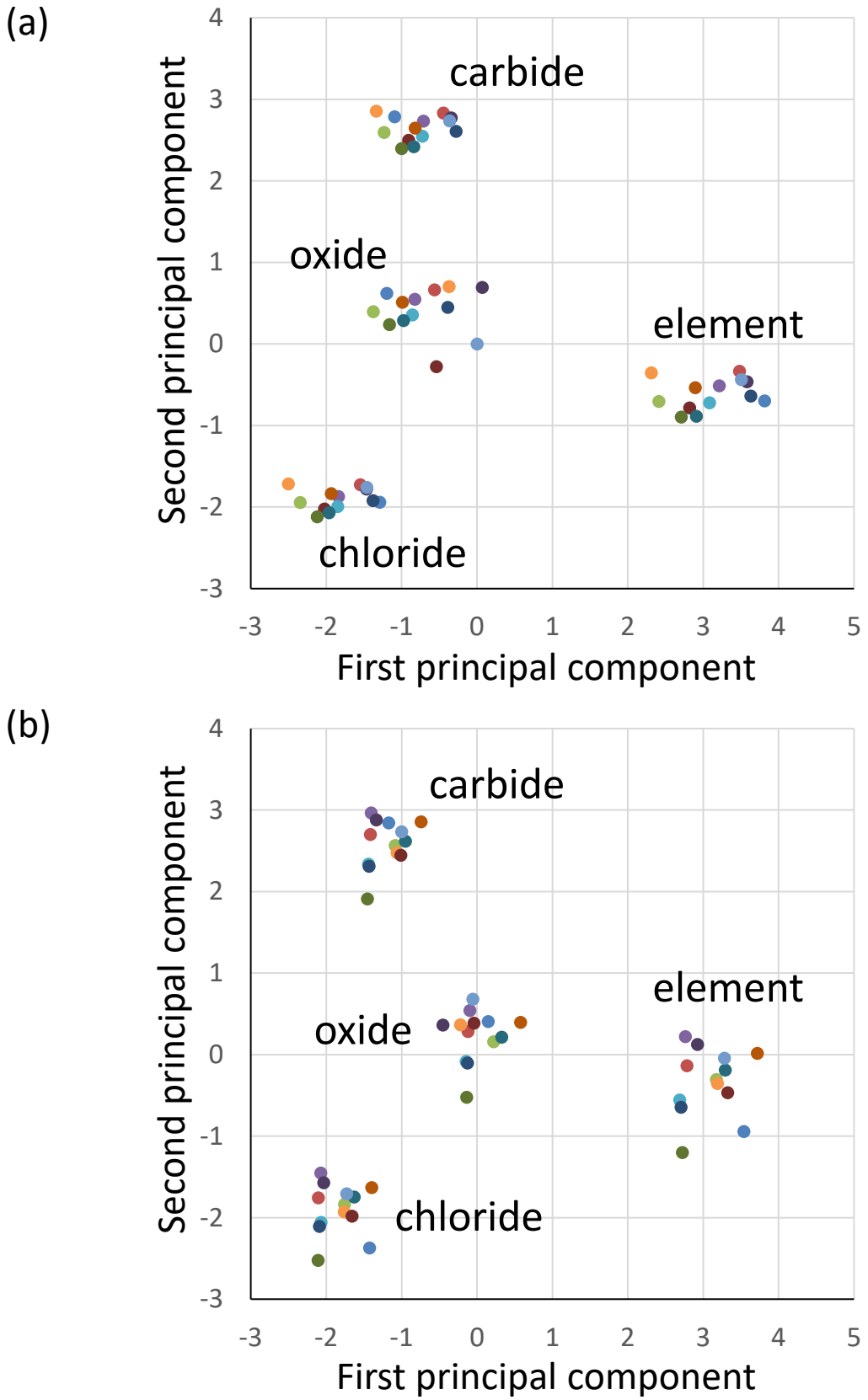

Figure 4. Word embeddings for 13 elements (lithium, sodium, magnesium, aluminium, silicon, calcium, titanium, iron, nickel, zinc, zirconium, molybdenum, tantalum, and their principal oxides, carbides and chlorides projected onto two dimensions using principal component analysis and represented as points in space. (a) is obtained using the original dictionary and (b) using the newly made dictionary from our corpus.

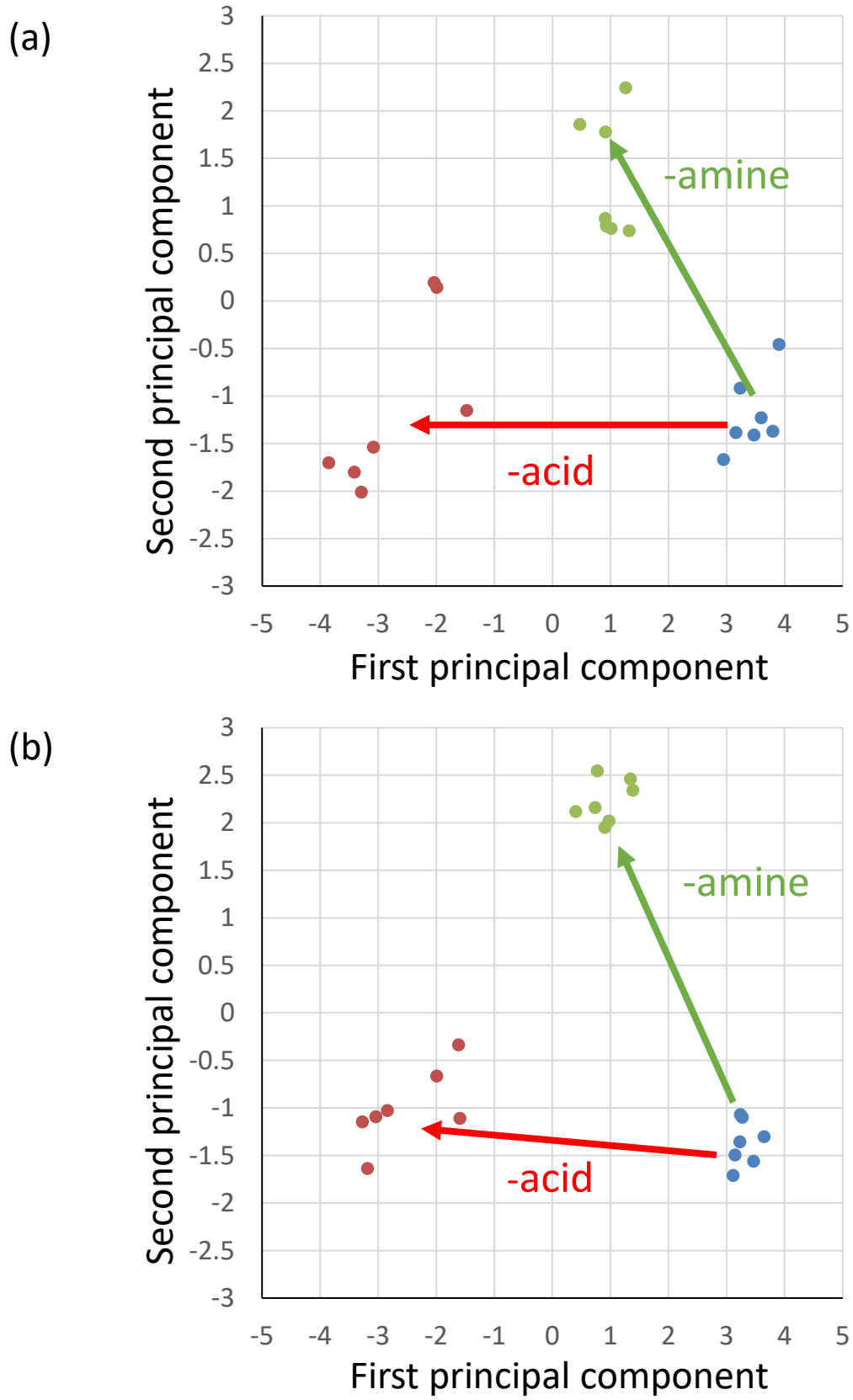

Figure 5. Word embeddings for 7 alkanes, and their carboxylic acid, and amine derivatives projected onto two dimensions using principal component analysis and represented as points in space. (a) is obtained using the original dictionary and (b) using the newly made dictionary from our corpus.

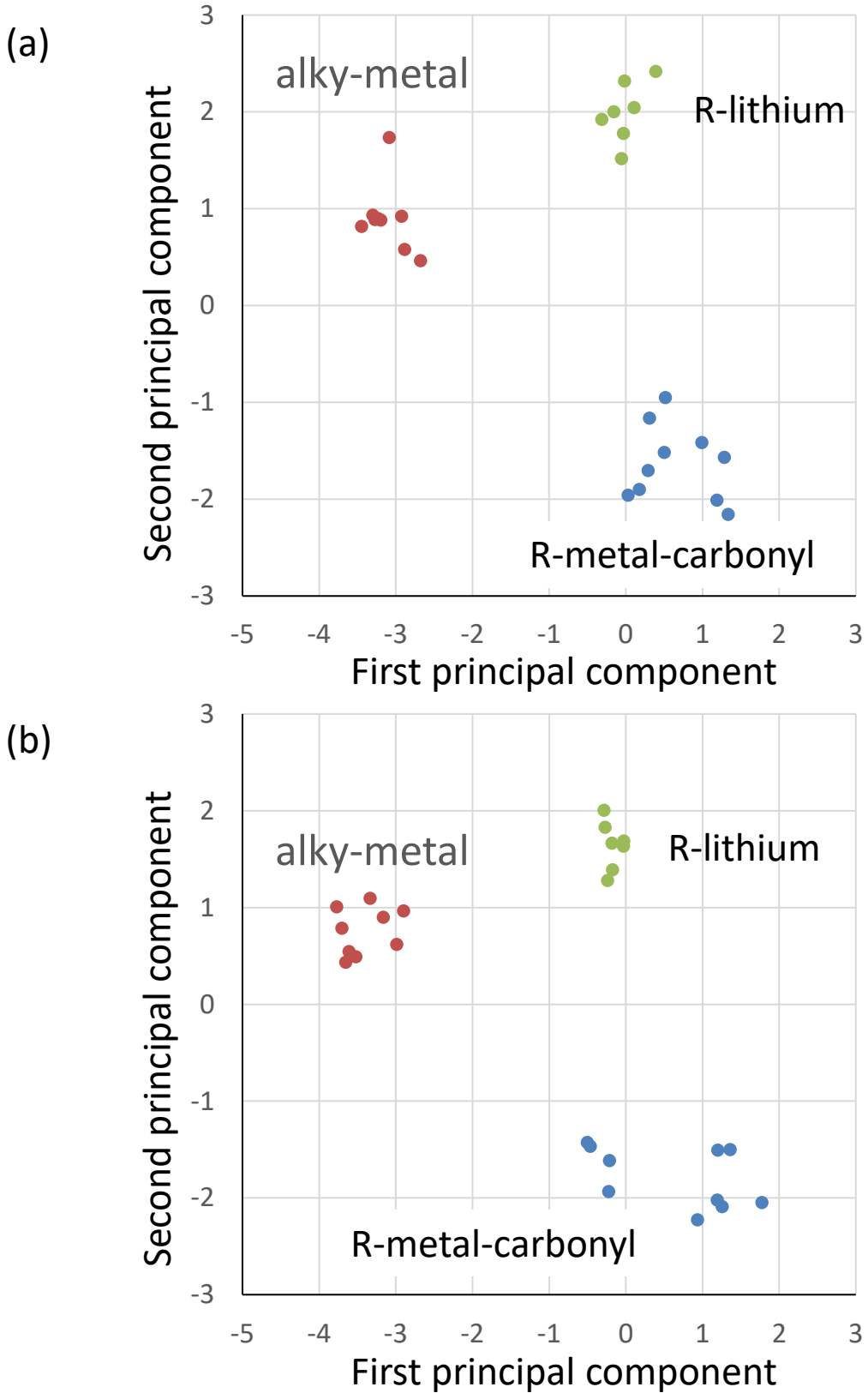

Figure 6. Word embeddings for organometallics (R-metal-carbonyl, alkyl-metal, and R-lithium, where R means an abbreviation for any group in which a hydrocarbon chain is attached to the rest of the molecule) projected onto two dimensions using principal component analysis and represented as points in space. (a) is obtained using the original dictionary and (b) using the newly made dictionary from our corpus.