NIPPON STEEL TECHNICAL REPORT No. 131 OCTOBER 2024

Technology

Natural Language Processing for Business Process Innovation

Kenichi IWATSUKI* Toshio AKAGI
Koji HIRANO

Abstract

Since business knowledge is stored in the form of documents, the application of natural language processing to them is mandatory for business process innovation. At Nippon Steel Corporation, the skip-gram model and BERT have been used on in-house documents in order to develop natural language processing technologies for the steel industry. In this report, we show that the word vector model helps to acquire synonyms of technical terms of the steel industry, and classification of technical documents improves using the BERT.

1. Introduction

Much of the knowledge related to business is stored in the form of documents. At Nippon Steel Corporation, for example, this knowledge includes the research and development documents accumulated over many years, the equipment maintenance reports recorded at the respective steelworks, and the specifications created in conjunction with the purchase of large quantities of materials and equipment.

The application of natural language processing technology is essential not only for storing these documents, but also for utilizing them to transform the business process. To this end, Nippon Steel has been working to build literature databases,[1] developing full-text search systems,[2] and extracting operation knowledge from operation problem reports.[3]

Natural language processing technology has advanced significantly in recent years. Thanks to the word2vec[4] and the BERT,[5] computers can process word and sentence meanings more easily. For example, word vectors (numerical representations of word meanings) can partly replace costly and laborious dictionary creation and maintenance. Sentence vectors (numerical representations of sentence meanings) enable sentence classification without summarizing synonyms or notation variations.

Some of the models trained by these methods are publicly available. For example, the "Japanese Wikipedia Entity Model"[6] and the chiVe[7] are skip-gram models. Kyoto University[8] and Tohoku University[9] have published their BERT models.

These models are versatile because they are trained on documents with diverse vocabularies, such as Wikipedia, but they are not suitable for specific domains like the steel industry. Therefore, the models are further trained on documents from relevant fields, such as financial statements[10] and medical papers.[11]

Nippon Steel has also created skip-gram models using documents from research and development, equipment maintenance, and intellectual property fields, and BERT models[12] using internal and external documents in the steel field.

The skip-gram model is a model trained by a neural network that learns to predict the surrounding words of a word in a sentence when a single word is input. The pretrained model can be obtained as pairs of words and their numerical representations.

The skip-gram model is used for document search and the BERT model is used for document classification. Both models contribute significantly to streamlining business processes at Nippon Steel.

In this paper, we describe the training of field-specific skip-gram and BERT models and the actual business process reform using these models within Nippon Steel.

2. Improvement in Efficiency of Document Search by Acquisition of Synonyms

Document search is vital for all business processes. A simple keyword search is often not enough. For example, when searching for research documents, technical terms may have different notations (including abbreviations), so a comprehensive search requires the listing of as many keywords as one can come up with. The same issue also occurs in patent searches. Moreover, in the procurement of materials and equipment, it is necessary to identify them to check past procurement prices, but synonyms and variations also pose problems.

Creating a dictionary manually is costly and difficult to maintain. Training a word vector model can replace this process. This way, the dictionary creation effort can be reduced, and dictionary

* Ph.D. in Computer Science, Intelligent Algorithm Research Center, Process Research Laboratories
20-1 Shintomi, Futtsu City, Chiba Pref. 293-8511
(Currently Senior Research Engineer, Mirai Translate, Inc.)

maintenance can be done by simply replacing training documents. However, the similarity accuracy of the word vector model is not as high as the manual dictionary, so the word vector model should be used for tasks where it is acceptable. In other words, the word vector model is more suitable for tasks where recall matters more than precision.

To apply the word vector model, we need to convert a document into a sequence of words. However, defining what a word is can be challenging. Morphological analysis is usually applied to split sentences into words. In this case, the sentences are split into morphemes, the smallest meaningful units of language. Many morphological analyzers have dictionaries. These dictionaries determine the minimum unit. If this level of granularity meets our goal, we can use an existing dictionary (or thesaurus) without creating a new one. A current issue is how to handle technical terms that are not in the dictionaries.

Therefore, we applied Byte Pair Encoding (BPE)[13] as a method that does not use a dictionary. We used Sentencepiece[14] for implementation. This method splits frequent character sequences in a document into single tokens, even if they are long. For example, the word "連続溶融亜鉛めっきライン" is split into "連続/溶融/亜鉛めっき/ライン" by MeCab[15] using mecab-ipadic-NEologd[16] as the dictionary. The BPE trained as described later splits "連続溶融亜鉛めっきライン" into "連続溶融亜鉛めっき/ライン." Similarly, "超音波探傷装置" is split into "超音波/探/傷/装置" and "超音波探傷装置" by the MeCab and BPE, respectively. This way, relatively long technical terms are represented by fewer tokens. Word vectors can be assigned to such technical terms. The relationships between the technical terms can be easily determined by computing the word vectors (for example, cosine similarity, which is the value obtained by dividing the inner product of two vectors by the product of their absolute values).

A skip-gram model was applied to assign word vectors to the words split in this way. There are three types of documents targeted for calculation, i.e., research-related documents accumulated at Nippon Steel (hereinafter referred to as research documents), reports related to equipment maintenance (hereinafter referred to as equipment documents), and patent documents for which steel companies are applicants (hereinafter referred to as patent documents). These documents are technical documents, but their purpose and content are slightly different.

Using the created word vectors, we investigated words with the highest cosine similarity (hereinafter referred to as similarity) with several words. Words with high similarity to "CGL, "ZAM", "モータ", "超音波探傷", "KR", and "圧延" are shown in **Tables**, **1**, **2**, **3**, **4**, **5**, and **6**, respectively.

"CGL" (Table 1) is an abbreviation for a continuous galvanizing line. In research documents, "CAPL" is the word with the highest similarity to "CGL". Since this word refers to a continuous annealing and processing line (C.A.P.L.), it can be said that the model is learned so that words belonging to the process genre are close to each other. In contrast, in patent documents, words that can be considered synonyms for CGL are close. Equipment names are listed in equipment documents. Such equipment documents are unique in that "2CGL," meaning a No.2 hot-dip galvanized steel sheet manufacturing line, is extracted as a single word.

"ZAM" (Table 2) is the trade name of ZAM™, a highly corrosion-resistant hot-dip coated steel sheet. "ZAM" was extracted as a word in research documents because information about product development is described in research documents, but "ZAM" was not extracted as a word in equipment documents or patent documents because it is rare for product names to appear in such documents.

"モータ" (Table 3) appears in each document. Words related to "モータ" are listed in research documents, and synonyms of "モー

Table 1 Top five synonyms of "CGL"

Research document	Equipment document	Patent document
CAPL	KAP	連続溶融亜鉛めっき
KAP	RCL	焼鈍炉
EGL	ETL	連続溶融めっき
CCL	EGL	直火
APL	2CGL	無酸化炉

Table 2 Top five synonyms of "ZAM"

Research document	Equipment document	Patent document
GI	-	-
めっき鋼板	-	-
溶融亜鉛めっき鋼板	-	-
めっき鋼板の	-	-
SGL	-	-

Table 3 Top five synonyms of "モータ"

Research document	Equipment document	Patent document
モーター	MOT	モータの
電動	モーター	電動機
インバータ	電動機	モーター
トルク	モーターの	モータは
アクチュエータ	Mot	電動機の

Table 4 Top five synonyms of "超音波探傷"

Research document	Equipment document	Patent document
探傷	-	探傷
非破壊検査	-	検査
検査	-	漏洩磁束探傷
欠陥検出	-	超音波探傷装置
探触子	-	渦流探傷

Table 5 Top five synonyms of "KR"

Research document	Equipment document	Patent document
脱硫	鍋	-
KIP	1REDA	-
脱珪	REDA	-
脱りん	RH	-
脱燐	3RH	-

Table 6 Top five synonyms of "圧延"

Research document	Equipment document	Patent document
圧延の	圧延中	の圧延
任延	次材	圧延機の
庄延	通材	圧延の
圧延で	圧延開始	圧延における
の圧延	パス目	熱間圧延

タ”, including spelling variants, are listed in equipment documents. This phenomenon is thought to occur because while research documents use somewhat standardized vocabularies, equipment documents are written on the site and contain abbreviations and idiomatic expressions that vary depending on the steelworks.

Concerning “超音波探傷” (Table 4), it is interesting to note that extremely niche words such as “漏洩磁束探傷,” “超音波探傷装置,” and “渦流探傷” are extracted from the patent documents. It is thought that these words appear quite a number of times in patent documents.

“KR” (Table 5) refers to hot metal pretreatment equipment based on the Kanbara Reactor method. Equipment documents list equipment names, while research documents list the purposes of various treatment methods in the steelmaking process, with a focus on desulfurization for which the KR method is used.

Concerning “圧延” (Table 6), it should be noted that the terms “任延” and “庄延” that do not exist are extracted from the research documents. These terms are thought to have resulted from reading errors when old documents were scanned and optically recognized.

As mentioned above, even in technical documents in the same steel field, the distribution of words included varies greatly depending on the type of document. It is considered meaningful to create a model for each type of document and to conduct document search by taking synonyms into account. This method is particularly useful in situations where comprehensive searches are required, such as searching for similar patents, investigating previous studies, searching equipment maintenance reports, and searching equipment purchase specifications.

3. Improvement in Efficiency of Document Classification by Steel BERT

If we can construct not only word vectors, but also sentence vectors and document vectors, the range of document utilization will expand impartially. In the manufacturing industry, proposals have been made to utilize equipment failure reports[17] and equipment maintenance reports[18] and to improve the efficiency of searching material databases.[19]

In these examples, the BERT, which has been pretrained on the Japanese version of Wikipedia and other documents, is often used. However, as mentioned in the previous section, the distribution of words differs greatly even in the literature in the steel field. It is expected that a model pretrained on texts from the steel field will perform better for each task than a model pretrained on general texts. A report on that effect is published in the financial field.[20]

Based on the above description, the BERT was pretrained on documents from the steel field. In recent years, many models larger than the BERT have been proposed, but the BERT was adopted because the amount of text in the steel field is small compared to all other fields and was expected to be around several GB. The texts used for pretraining were from open data such as patents and technical reports from various steel companies and research documents owned by Nippon Steel. We prepared two models. One was prepared by using only open data and the other was prepared using all data or open data and proprietary data. In addition, SentencePiece (BPE) was used as the tokenizer, and the models were pretrained on the respective texts.

To evaluate their performance, the models were compared with Tohoku University’s BERT (bert-base-japanese-whole-word-masking with the same tokenizer) by using the following two tasks. One task is patent classification.[12] It is concerned with a multi-label classification problem consisting of inputting a portion of a patent specification and of outputting the technology category to which the patent specification belongs. The other task is the classification of internal research documents. The problem setting is the same as for the first task. We compared two models for the former task and three models for the latter task. All models were fine-tuned and evaluated.

Table 7 Results of classification of patents

Model	Accuracy
NS-BERT (trained on open data)	0.592
General BERT (Tohoku Univ.)	0.563

Table 8 Results of classification of in-house research reports

Model	Accuracy
NS-BERT (trained on all data)	0.582
NS-BERT (trained on open data)	0.574
General BERT (Tohoku Univ.)	0.564

The results are shown in **Tables 7** and **8**. In the patent classification, the steel BERT (NS-BERT) outperformed the general-purpose BERT. Additionally, in the in-house research document classification, the NS-BERT pretrained on both in-house data and open data showed better performance than the NS-BERT pretrained only on open data.

As described above, the NS-BERT model pretrained on both in-house document data and open data can achieve more accurate document classification. Document classification is a burdensome task to perform manually. This NS-BERT model contributes to the improvement in the efficiency of document classification.

4. Conclusions

In this paper, we trained a tokenizer using documents from the steel field and reported the results of training skip-gram models and pretraining BERT models. The word vectors showed synonyms in different directions depending on the type of document. Regarding the BERT models, we showed that the NS-BERT model has outperformed the general BERT model in document classification.

From the perspective of business process innovation, word vectors can be applied to comprehensive document searches. In addition, the BERT can be applied not only to document classification, but also to information extraction through sequence labeling, which is an area of future application.

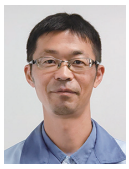
References

1) Ohmi, A., Akashi, Y.: Construction of In-house Databases in a Corporation. Journal of Information Processing and Management. 30 (11), 1097–1111 (1988)
2) Nakamoto, S.: Full Text Retrieval for Huge Volumes of Data in Patent System “PSEARCH/DB”. Nippon Steel Technical Report. (76), 43–46 (1998)
3) Aoyama, K., Sawada, T., Koga, T., Yaji, Y., Mori, J.: Text Mining Approach of Trouble Report to Manage Operation Knowledge for Continuous Casting Process: Discrete Model of Continuous Casting Process for Management of Operation Knowledge. The Proceedings of Mechanical Engineering Congress, Japan. 2012
4) Mikolov, T., Sutskever, I., Chen, K., Corrado, G., S., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems 26, 2013
5) Devlin, J., Chang, MW., Lee, K., Toutanova, K.: BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Asso-

ciation for Computational Linguistics: Human Language Technologies. 4171–4186 (2019)
6) Suzuki, M., Matsuda, K., Sekine, S., Okazaki, N., Inui, K.: A Joint Neural Model for Fine-Grained Named Entity Classification of Wikipedia Articles. IEICE Transactions on Information and Systems, Special Section on Semantic Weband Linked Data. E101-D (1), 73–81 (2018)
7) Manabe, H., Oka, T., Kaikawa, S., Takaoka, K., Uchida, Y. Asahara, M.: Japanese Word Embedding based on Multi-granular Tokenization Results (in Japanese). Proceedings of 25th Annual Meeting of the Association for Natural Language Processing. 1407–1410 (2019)
8) Shibata, T., Kawahara, D., Kurohashi, S.: Improvement in Accuracy of Japanese Syntactic Parsing using BERT (in Japanese). Proceedings of 25th Annual Meeting of the Association for Natural Language Processing. 205–208 (2019)
9) https://github.com/cl-tohoku/bert-japanese
10) Suzuki, M., Sakaji, H., Hirano, M., Izumi, K.: Construction and Validation of a Pre-Training and Additional Pre-Training Financial Language Model. JSAI Technical Report, Type 2 SIG. FIN-028, 132–137 (2022)
11) Sugimoto, K., Iki, T., Chida, Y., Kanazawa, T., Aizawa, A.: JMedRoBERTa: a Japanese Pre-trained Language Model on Academic Articles in Medical Sciences (in Japanese). Proceedings of 29th Annual Meeting of the Association for Natural Language Processing. 707–712 (2023)
12) Iwatsuki, K.: Possibility of Using BERT Pretrained on Relatively Small Amount of Domain-Specific Data: Case Study in Steel Industry (in Japanese). Proceedings of 28th Annual Meeting of the Association for Natural Language Processing. 741–745 (2022)
13) Sennrich, R., Haddow, B., Birch, A.: Neural Machine Translation of Rare Words with Subword Units. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 1715–1725 (2016)
14) 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. 66–71 (2018)
15) Kudo, T., Yamamoto, K., Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. 230–237 (2004)
16) Sato, T., Hashimoto, T., Okumura, M.: Implementation of a word segmentation dictionary called mecab-ipadic-NEologd and study on how to use it effectively for information retrieval (in Japanese). Proceedings of 23rd Annual Meeting of the Association for Natural Language Processing. 875–878 (2017)
17) Honma, H., Komachi, M., Manabe, A., Tanimoto, K.: Extraction of Important Points from Failure Reports using BERT Model (in Japanese). Proceedings of 27th Annual Meeting of the Association for Natural Language Processing. 189–193 (2021)
18) Takahashi, T., Taniguchi, M., Taniguchi, Y., Okuma, T.: Study on Text Structuring for Utilization of Equipment Maintenance Reports (in Japanese). Proceedings of 28th Annual Meeting of the Association for Natural Language Processing. 230–234 (2022)
19) Yomogida, A., Murase, F., Hirano, T., Mitani, A., Ban, K., Iida, T., Iwabori, K., Takeno. T.: Verification of Retriever-Reader Model for Utilization of Technical Knowledge (in Japanese). Proceedings of 29th Annual Meeting of the Association for Natural Language Processing. 2030–2033 (2023)
20) Suzuki, M., Sakaji, H., Hirano, M., Izumi. K.: Construction and Validation of a Pre-Trained Language Model Using Financial Documents. JSAI Technical Report, Type 2 SIG. FIN-027, 5–10 (2021)

Kenichi IWATSUKI
Ph.D. in Computer Science
Intelligent Algorithm Research Center
Process Research Laboratories
20-1 Shintomi, Futtsu City, Chiba Pref. 293-8511
(Currently Senior Research Engineer, Mirai Translate, Inc.)

Toshio AKAGI
Principal Researcher
Intelligent Algorithm Research Center
Process Research Laboratories
(Currently General Manager, Developing & Planning Division, Electrical Instrumentation Unit
Nippon Steel Texeng. Co., Ltd.)



Koji HIRANO
Ph.D., General Manager, Head of Dept.
Production Management Research Dept.
Intelligent Algorithm Research Center
Process Research Laboratories