Extraction of Inorganic Material Synthesis Procedure from Literature

Shanshan Liu[1] Fusataka Kuniyoshi[2,3] Jun Ozawa[2,3]
Masaki Kiyono[2] Yuji Matsumoto[1,4]
[1]Nara Institute of Science and Technology (NAIST)
[2]Panasonic Corporation
[3]National Institute of Advanced Industrial Science and Technology (AIST)
[4]RIKEN Center for Advanced Intelligence Project (AIP)
{kuniyoshi.fusataka, kiyono.masaki}@jp.panasonic.com
{ozawa.jun}@aist.go.jp
{liu.shanshan.lm2, matsu}@is.naist.jp

1 Introduction

Structured information extraction from the literature has always been an indispensable technique to support data-driven methods in various fields. However, due to the lack of machine-readable datasets and widely used structured definitions of information, the extraction technology of inorganic material synthesis information is just beginning, related researchers have not enjoyed the convenience of data-driven methods yet.

Previous researches most focused on named entity recognition (NER) and action graphs extraction. Kim et al.[1] annotated some paragraphs with entity labels that were selected from the literature including the synthetic procedure of inorganic substances. Mysore et al.[2] provided a more complete labeled dataset containing over 100 procedures. Making use of expert-annotated datasets, Mysore et al.[3] defined a structured procedure as a set of linked action events and extracted action graphs by heuristic model. Kim et al.[4] extracted the action series in the procedure and developed a system to provide possible precursors and action graph given the target material. Tamari et al.[5] converted the procedure description into instructions in a text-based interactive game, construct the action graph of the procedure from the text.

On this basis, we try to challenge a more difficult problem, extracting the complete material synthesis procedure including condition factors such as "pressure" using NER and Relation Extraction (RE) techniques. After the manual analysis of the descriptions in the literature, we clearly defined the procedure, provide a pipeline extracting method. The result on Kuniyoshi's dataset[6] demonstrates that our method is feasible and full of potential.

2 Methodology

2.1 Task Formulation

Multiple operations are performed sequentially in a synthesis procedure (see Figure 1 and 2). An operation is carried out on certain precursor materials, generating some products under specific environmental conditions. We can thus represent a procedure by three types of entity and six types of relation. A material m consists of a mention d and property descriptions while n_p is the number of property entities:

$$m = \{d, p_1, p_2, \ldots, p_{n_p}\}$$

The set of precursor materials M_{in} can be represented as a set of materials:

$$M_{in} = \{m^i_1, m^i_2, \ldots, m^i_{n_{in}}\}$$

The set of generated materials M_{out} is:

$$M_{out} = \{m^o_1, m^o_2, \ldots, m^o_{n_{out}}\}$$

The set of conditions C of the operation is defined as:

$$C = \{c_1, c_2, \ldots, c_{n_c}\}$$

Based on the above definitions, an operation O with a mention o can be represented as:

$$O = \{o, C, M_{in}, M_{out}\}$$

A full material synthesis procedure S is a series of operations while n_o is the number of operations:

$$S = \{O_1, O_2, \ldots, O_{n_o}\}$$

2.2 Procedure Extraction Model

Our procedure extraction system (see Figure 3) consists of one NER block and three RE blocks. Considering the portion of cross-sentence relations in training data, we use Graph State

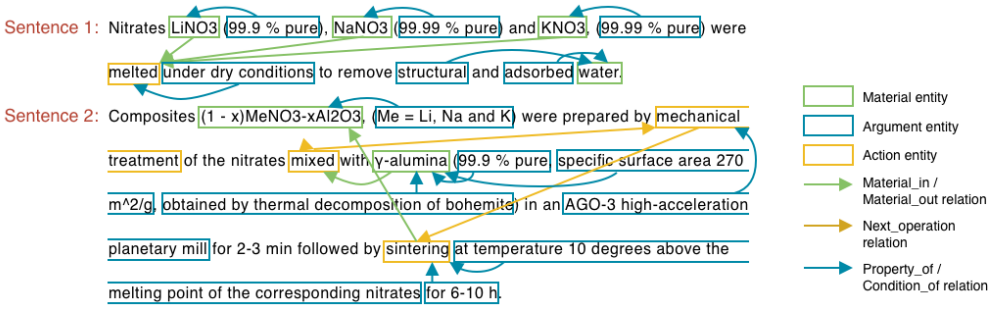

Figure 1: A example of a paragraph describing synthesis procedure. "LiNO3", "NaNO3", and "KNO3" are precursors of the operation "melted". After "melted", "mixed", "mechanical treatment", "sintering", the target material "(1 - x)MeNO3-xAl2O3" is obtained.

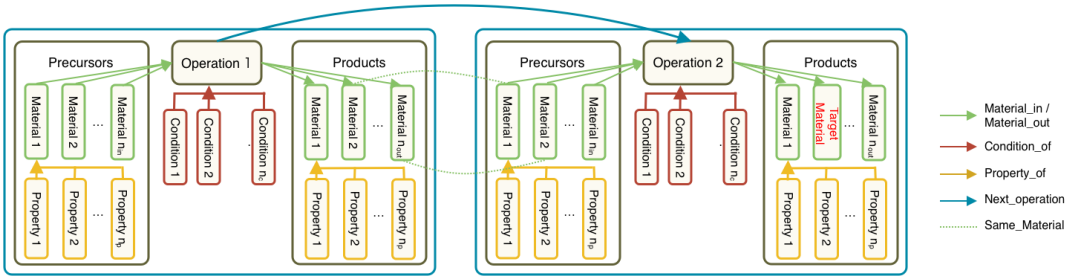

Figure 2: The structure of a synthesis procedure including 2 operations. The arrow represents the relationship between 2 entities. The dot line means two entities are the same material. The second operation takes the output of the first operation as its input to generate the target material.

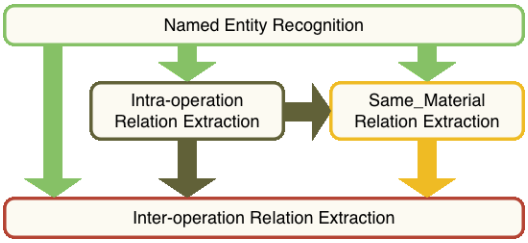

Figure 3: The workflow of our procedure extraction system. The result of each block is passed to others as arrows point.

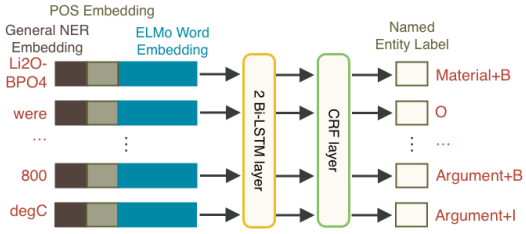

Figure 4: The architecture of the NER model.

Long Short-Term Memory (GS-LSTM) model presented by (Song et al., 2018)[7] which can extract both the intra-sentence and cross-sentence relation. We implemented our model except Same_Material RE part so far.

NER We combine an ELMo embedding model released by [4], bidirectional Long Short-Term Memory (Bi-LSTM) model, and Conditional Random Fields (CRF) model as our NER framework that is widely used (Figure 4). A rule-based method is also used for Operation entity. We constructed a vocabulary for operation expressions common in material synthesis after analyzed Kuniyoshi's data, divided them into 17 categories: *add, cool, crush, deposit, disperse, dry, finish, ground, heat, keep mill, mix, open, pelletize, press, wash, weigh.* We directly label the matching words in the text with their category labels. For example, "dissolve" is classified as "mix" while "sinter" is to "heat". In capturing the relation between "dissolve" and "sinter", we also consider the possible connection between their categories "heat" and "mix". The construction of the action graph may be more accurate additionally take these category labels into consideration.

Intra-operation RE The extraction of intra-operation relations is to get *"Material_in"*, *"Material_out"*, *"Property_of"*, and *"Condition_of"* relations. The positive instances predicted by the RE model are recorded to support the Inter-operation RE.

Same_material RE This part aims to make a compliment on other two RE tasks. Sometimes, different mentions mean the same material while they may play roles in different operations, providing information to capture the relation between operation.

Inter-operation RE This is for the relationships we defined as *"Next_operation"* that of

Table 1: The statistics of Procedure dataset

Type	#Train	#Dev	#Test	#All
Entity	6559	938	926	8423
Relation	6150	888	846	7884
Operation	1309	202	169	1680
Paragraph	193	25	25	243

Dataset statistics about the number of named entities, relations and operations for training (#Train), development (#Dev), testing (#Test) and of all paragraphs (#All).

Table 2: Precision(P), Recall(R) and F1 score(F1) of NER task

Task	P	R	F1
Material (extended)	88.6	96.01	92.15
Material	82.83	82.1	82.47
Argument	66.83	65.93	66.38
Operation	79.73	76.87	78.28
Operation (rule)	62.6	66.17	64.33

Operation (rule) identifies the entity by exact-matching. **Material (extended)** takes both the dataset of [4] and [6] for training. Other tasks are trained on Kuniyoshi's dataset using the model like Figure 4.

operation entity-pairs, while the subject is the current operation, the object is the next one. Considering the information we get in other tasks, we may take information of the key attributes in operation into account for classification.

3 Experiment

3.1 Dataset

We utilized a procedure dataset constructed by **Kuniyoshi et al.,[6]** to evaluate our extraction approach (statistics as Table 1).This dataset is a set of human-annotated paragraphs of material synthesis procedures selected from the literature. Each paragraph contains at least one full synthesis procedure annotated with the entity label and relationship label by chemistry experts. Besides, we use **Kim et al., [4]** to improve the capability of identification of "Material" with its 6,744 material entities labeled from a dataset contains 235 synthesis procedures.

3.2 Implementation

Pre-processing Given a plain-text, after the parsing by Stanford NLP Toolkit[8], each word is annotated with its part-of-speech(POS) tag, and the general named entity tag (i.e., NUMBER), and align it to its human-annotated named entity label (i.e., Material). When we combine Kim's data with Kuniyoshi's data, we find that the label standards of the two datasets are not consistent. Kim subdivides the material mentions into "property_misc, precursor, material, target", but Kuniyoshi marks the material uniformly as "Material". They also differ in word segmentation. Inorganic substances in Kim, whether they include brackets or not, are treated as one token, but the words including brackets in Kuniyoshi are cut into tokens by the tokenizer. Therefore, all of the words in the two datasets that are in the form as [V2O5] are processed to [V(2)O(5)], and split into several tokens by brackets. We built training instances according to the form of (Peng et al., 2017)[9] taking the golden truth of NER.

Statistical analysis of training data shows most of the relationships, two entities are within 5 adjacent sentences (when subject/object entity is in the first sentence of the paragraph, 97% of the object/subject entity is in the second to the fifth sentence). When constructing the training sample, we thus set the window size as 5 sweeping through each paragraph. If paragraph contains less than five sentences, window_size will be the length of paragraph. For each entity appearing in the first sentence, we combine it to other entities in the window under the constraint by the entity types. Combined entity-pairs are divided into 6 relation types we defined and "No_relation".

NER This study aims to process the chemistry literature text, thus the ELMo pre-trained models open-sourced by the [4] were selected. This embedding model was trained on 2.5 million materials science journal articles. We trained on only Kuniyoshi's dataset for three entity types.We trained a "Material" extraction model on extended dataset that including all data of [4] and training set of Kuniyoshi's. For "Operation", we divided text expressions of the action that appeared in the training data into 17 categories and identify entities by exact-matching.

Intra-operation RE We found the GS-LSTM model can not learn well when imbalance skews problem occurs. *#Neg:#Pos* reaches $13:1$ in Intra-operation relation if $window\_size = 5$. To cope with this, we used sampled data selected by the K-Means cluster while $k = 8$ for training. *#Neg:#Pos* in each sampled dataset is $1:1$ including all positive instances and sampled negative instances of original training set.

Inter-operation RE We tried four different strategies. 1) Given an operation entity, use the rule-based method that sets the nearest operation that occurs after the given entity as the next operation. 2) Use GS-LSTM as the encoder and directly use the encoded entity representation for the prediction of entity-pairs. We do not sample the training data in this task so far. 3) Encode the sentences by Bi-LSTM,

Table 3: RE performance by GS-LSTM model

Task	Accu	P	R	F1
Intra-operation	37.93	9.98	93.62	18.04
Inter-operation	37.5	23.8	64.5	34.7

The accuracy (Accu) and other evaluation indicators on testing dataset.

Table 4: Performance of Inter-operation RE

Method	Accu	P	R	F1
Rule-based	**84.8**	**84.8**	**92.6**	**88.5**
GS-LSTM	37.5	23.8	64.5	34.7
Bi-LSTM	85.2	69.5	77.2	73.1
Ensemble	62.5	29.0	34.4	31.5

concatenate the encoded entity representation, the embeddings of entity type and the distance between the two entities to predict. 4) Ensemble the output of Bi-LSTM and the output of rule-based method. Compared with the output of the rule-based method, if they are the same, it is directly used as the final prediction result. Otherwise, a linear layer with inputting the encoded entity representation and the embedded rule-based result does the final prediction.

3.3 Results and Analysis

The NER result (Table 2) shows our model can get reliable entity labels for further RE tasks. The operation vocabulary is not complete to achieve a higher score than NN-based method. We have to enrich expressions to make our rule-based method more accurate.

The GS-LSTM model performs not promising in Table 3. The recall is high in Intra-operation RE, which means we can add one more classifier to remove irrelevant relations after we identify relations. A better way to solve the imbalanced skews problem is needed. Comparing the results of four strategies for Inter-operation RE, the rule-based method is simplest but best (Table 4). Encoding sentences by Bi-LSTM but considering entity types and distance between entities reach an F1 score of 73.1. We can infer that rule-based features can help to better performance of Inter-operation RE, however, we still need to research on the way of ensembling. We still not refer to the Inter-operation relation information when connecting two operations, which means there is still an improvement zone for our approach. The RE result on the ground truth of entity labels is not good, the situation that taking the result of NER as the input of RE can be very challenging. To consider the cascade errors in our pipeline model is particularly under low precision on Intra-operation RE, changing pipeline learning to multi-task learning may solve this problem.

4 Conclusion

In this paper, we provide a pipeline method to extract the procedure of inorganic material synthesis from literature. We define the procedure as various types of entities and connections between entities. The procedure extraction consists of two subtasks: Named Entity Recognition (NER) and Relation Extraction(RE). We evaluate the capability of our approach on a procedure dataset involving 243 paragraphs selected from the literature. The result shows the feasibility that our approach can extract named entity and relationships in procedure well in such cases, get a more clear vision of the difficulty to apply NLP techniques in practice. We will keep researching on catching the relationships of attributes in the procedure, dealing with the imbalance skews problem and reducing the potential cascade errors caused by the pipeline framework.

References

[1] Edward Kim, Kevin Huang, Adam Saunders, Andrew McCallum, Gerbrand Ceder, and Elsa Olivetti. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chemistry of Materials*, 29(21):9436–9444, 2017.

[2] Sheshera Mysore, Zach Jensen, Edward Kim, Kevin Huang, Haw-Shiuan Chang, Emma Strubell, Jeffrey Flanigan, Andrew McCallum, and Elsa Olivetti. The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures. *arXiv preprint arXiv:1905.06939*, 2019.

[3] Sheshera Mysore, Edward Kim, Emma Strubell, Ao Liu, Haw-Shiuan Chang, Srikrishna Kompella, Kevin Huang, Andrew McCallum, and Elsa Olivetti. Automatically extracting action graphs from materials science synthesis procedures. *arXiv preprint arXiv:1711.06872*, 2017.

[4] Edward Kim, Zach Jensen, Alexander van Grootel, Kevin Huang, Matthew Staib, Sheshera Mysore, Haw-Shiuan Chang, Emma Strubell, Andrew McCallum, Stefanie Jegelka, and Elsa Olivetti. Inorganic materials synthesis planning with literature-trained neural networks, 2018.

[5] Ronen Tamari, Hiroyuki Shindo, Dafna Shahaf, and Yuji Matsumoto. Playing by the book: An interactive game approach for action graph extraction from text. In *Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications*, pages 62–71, 2019.

[6] Fusataka Kuniyoshi, Jun Ozawa, Makoto Miwa, et al. Graph representation for synthesis process extraction from inorganic material literature. In *15th Text Analytics Symposium in IEICE-NLC*, 2019.

[7] Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. N-ary relation extraction using graph state lstm. *arXiv preprint arXiv:1808.09101*, 2018.

[8] Richard Socher, John Bauer, Christopher D Manning, and Andrew Y Ng. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 455–465, 2013.

[9] Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. Cross-sentence n-ary relation extraction with graph lstms. *Transactions of the Association for Computational Linguistics*, 5:101–115, 2017.