International Journal of Smart Computing and Artificial Intelligence
International Institute of Applied Informatics
2024, Vol. 8, No. 1, IJSCAI780

Simile Identiýcation with Pseudo Data Acquisition and Re-labeling

Jintaro Jimi [*], Kazutaka Shimada [À]

Abstract

The simile is a kind of figurative language. It expresses the target of the figurative language by using some typical phrases such as “like”. It is important to distinguish whether the sentence is a simile or a literal for understanding a sentence. However, a large amount of data is required to generate a classifier by machine learning. Moreover, creating the dataset is costly. In this paper, we propose a pseudo dataset acquisition method for simile identification. We first construct a dataset of simile and literal sentences using machine translation. We utilize mBART as the machine translation system. This process automatically generates pseudo-simile and literal instances from three types of corpora. Then, we apply some machine learning approaches to the simile identification task. We compare Support Vector Machine, Naive Bayes, and BERT in the experiment. The experimental result shows the validity of the pseudo dataset as compared with a simple baseline (machine translation with rules). In addition, re-labeling with machine learning for the original pseudo data contributed to the improvement of the simile identification accuracy.

Keywords: simile identification, pseudo data, automatic training data acquisition, figurative language

1 Introduction

Understanding figurative language is one of the important tasks in natural language processing. This task is essential to understand sentences correctly since sentences with figurative expressions do not always contain literal meanings. However, it is a difficult task because extra-linguistic information, such as conceptual and commonsense knowledge, is often needed for understanding.

In this paper, we focus on simile expressions in Japanese text. Similes are a representative usage of figurative language. In a sentence with a simile, a comparator is often used to indicate the figurative target. In Japanese, the phrase “ような (like)” corresponds to the comparator. However, the phrase “ような” is not always used as a comparator. In the examples below, the phrase “ような” is used as a comparator in (1) and as a quotation in (2). (1e) and (2e) are English translations of (1) and (2).

[*] Department of Creative Informatics, Kyushu Institute of Technology, Fukuoka, Japan
[†] Department of Artificial Intelligence, Kyushu Institute of Technology, Fukuoka, Japan

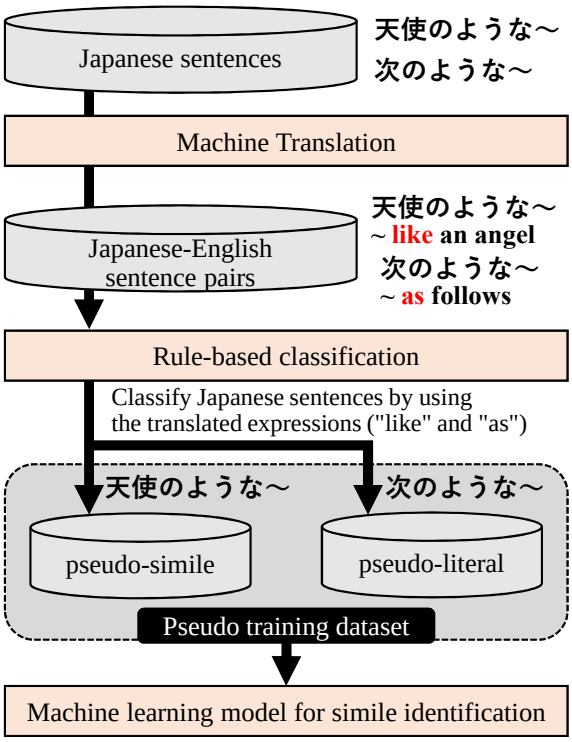

Figure 1: The outline of our method. The method consists of two steps. We obtain a pseudo dataset via machine translation as the first step. Then, we utilize the pseudo dataset for generating a simile identification model.

- Examples of sentences containing “ような”

  (1) 天使の**ような**かわいい小鳥よ。(comparator)

  (1e) A cute little bird **like** an angel. (simile)

  (2) 具体的には次の**よう**な問題である。(quoting)

  (2e) Specifically, it is **as** follows. (literal)

Therefore, we need to distinguish the meaning of this phrase for simile identification.

In this paper, we deal with a simile identification task as a classification problem using machine learning. Traditional machine learning methods require a large amount of training data: simile and literal sentences. However, manually creating such a training dataset is costly. Therefore, we need a method to acquire the data automatically.

We utilize a machine translation model for constructing pseudo-training data. Here we focus on the difference in translation outputs of the comparator. Depending on how the comparator is used, the expressions from machine translation are also different: e.g., “like” and “as” in (1e) and (2e). On the basis of this assumption, we acquire pseudo-training data automatically and then generate a model by using the pseudo-data. Figure 1 shows the outline of our method. First, our method translates Japanese sentences to English by a machine translation model. Then, it classifies the English sentences with original Japanese sentences into pseudo-simile and pseudo-literal by using two keywords: “like” and “as”. In this paper, we regard sentences with “like” as similes and ”as” as literal meanings. Since our target language is Japanese, we add only Japanese sentences to the pseudo dataset. If the

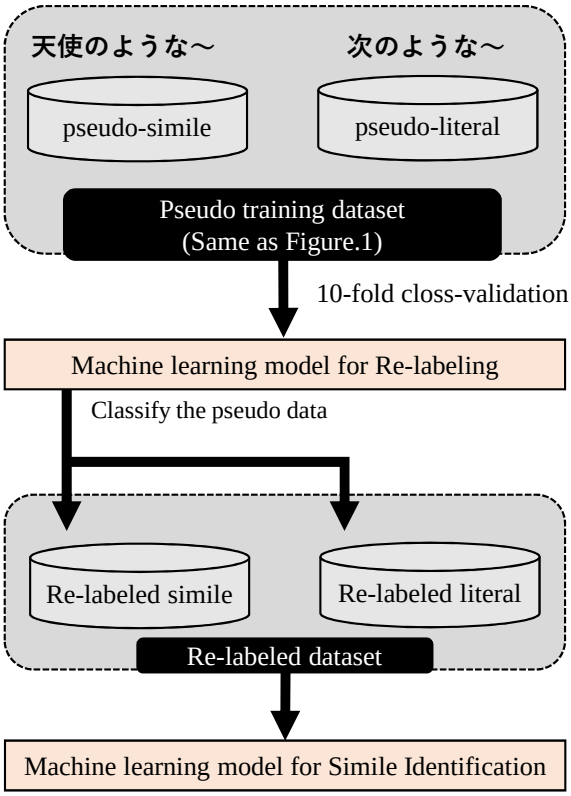

Figure 2: The outline of our relabeling method. We obtain a more accurate dataset via a machine learning model. By using the dataset, we generate a simile identification model in a similar way to Figure 1.

translated sentences do not contain the keywords, we delete the original Japanese sentences from the dataset. Finally, we generate a simile identification model using machine learning from the dataset. We compare several machine learning models for the task.

The pseudo dataset for training is based on a simple assumption from the result of machine translation. Therefore, the data contain incorrect labels: e.g., non-simile sentences with the pseudo-simile label. Noisy data tend to generate a wrong classifier. In this paper, we introduce a re-labeling approach to the simile identification task. We classify the pseudo data by using a machine learning model to obtain a more accurate dataset. Figure 2 shows the outline of our method. First, we perform a 10-fold cross-validation on the pseudo-training dataset. In this process, a machine learning model assigns a new label to each sentence. Then, we reconstruct a dataset with the re-labeled smile and re-labeled literal instances. Finally, we learn a machine learning model of the simile identification task with the new dataset in a similar way to Figure 1.

The contributions of this paper are:

- We propose a pseudo data acquisition method using machine translation for simile identification. The pseudo data work well in the experiment.

- We introduce a re-labeling approach using machine learning to the pseudo data acquisition method. The re-labeled dataset contributes to the improvement of the simile identification accuracy.

2 Related Work

Some rule-based approaches for simile identification have been proposed. They were usually based on relationships between words in metaphor. For example, (1e) in Section 1, "angel" and "little bird" are the metaphorical relation. Tazoe et al. [1] have tried to identify simile sentences by pattern classification based on semantic information of nouns. However, they handled only "A のような B (noun$_B$ like noun$_A$)" as the target pattern. Moreover, the size of their knowledge base was small. As a result, the method contains a problem in terms of versatility.

Recently, several approaches based on neural network models have been investigated for this task. Gao et al. [2] have proposed a metaphorical word detection model based on BiLSTM. In addition, Liu et al. [3] have reported a neural network framework for jointly optimizing three tasks: simile sentence classification, simile component extraction, and language modeling. The methods recorded higher accuracy in each task than the previous rule-based methods. However, neural network-based methods essentially require a large dataset of simile and literal sentences. Manually creating such a dataset is costly. Moreover, there is no large dataset in Japanese. We automatically construct a large dataset for simile identification.

Pseudo-data acquisition is one approach to solving the problem. The purpose is to improve the size of data and reduce manual costs in data construction. In image processing, pseudo-data acquisition can be easily performed by rotating or inverting the image. Kobayashi et al. [4] and Ji et al. [5] have reported the effectiveness of pseudo-data acquisition for machine learning. Rotating and inverting images analogize to replacing words and changing the word order in natural language processing. However, these processes often change the meaning of sentences. It denotes the difficulty of pseudo-data acquisition in NLP.

To create a pseudo dataset appropriately in NLP, Nishimoto et al. [6] have utilized a thesaurus. However, simple word replacement contains a problem with the diversity of the pseudo dataset. Sassano [7] has used a vector space model of words. However, the method needs initial training data for the augmentation process. Here we focus on machine translation. Recently, various neural network-based models have greatly improved the quality of machine translation models [8][9]. Using translation, we can obtain a wide variety of sentences because machine translation generates output sentences from any input sentences. Jimi and Shimada [10] have proposed an approach for pseudo dataset construction with machine translation and a simple rule based on keywords. They show the effectiveness of the pseudo dataset. However, the pseudo dataset contained many noise instances. This paper extends their approach. We introduce a re-labeling approach based on machine learning, while the previous study used only a simple rule.

3 Datasets

This section describes datasets in this task. We create two datasets: a pseudo-training dataset for machine learning and a dataset for evaluation. The first one is created automatically, and the second one is created manually.

3.1 Pseudo-training dataset

We automatically acquire a pseudo-training dataset by using machine translation. We translate Japanese into English. The comparators that we handle in this paper are “のような” and “のように”. Both comparators tend to be translated as “like” when the sentences with the comparators are similes.

In this paper, we use a neural machine translation model called mBART [11]. mBART is a neural translation model that is trained by applying BART of [12], to a large corpus of multiple languages. Several versions of mBART have been released. We use a model proposed by [13] that is available on Hugging Face[1] for this task.

We assume that the tendency and frequency of simile expressions depend on the source of information. For example, similes are more likely to occur in novels and less likely to occur in descriptive texts such as entities in Wikipedia. Therefore, in this paper, we use three different types of corpora for pseudo-data acquisition: Aozora Bunko (a Japanese novels dataset)[2], Wikipedia[3], and Mainichi newspaper 1995. In this paper, each corpus is called a domain. First, we extract sentences containing “のような” or “のように” from the three corpora. “の” is a case particle and limits the preceding phrase to a noun phrase. In addition, the conjunction with “ように” is an inflectional form of “ような”. Here, a sentence containing patterns “このような（に），そのような（に），あのような（に），どのような（に）” are excluded because the phrases are typical demonstrative expressions and not comparators in this study although the part of the comparators (のように/のような) is contained.

Next, we obtain the English translations of Japanese sentences with the comparator by mBART. Here we judge whether each input is a simile or literal sentence by using the presence of keywords in the translated sentence. The phrases “のような” and “のように” tend to be translated in English as “like” when they are used in the simile sense. Similarly, the phrases are often translated to “as” when they are used in a literal sense. In this process, we delete the sentence if it does not contain both ”like” and ”as” because we cannot judge whether it is a simile or a literal sentence. The following examples (3) and (4) are included in the dataset. (3e) and (4e) are obtained by machine translating (3) and (4).

- Pseudo-simile dataset: a sentence containing “like” in the English translation

  (3) 丁度鶏の脚**のような**骨と皮ばかりの腕である。

  (3e) It is just a bone and skin arm **like** a chicken’s leg.

- Pseudo-literal dataset: a sentence containing “as” in the English translation

  (4) このことは次**のように**明らかになるであろう。

  (4e) This will become clear **as** follows.

Table 1 shows the distributions of sentences in pseudo-training datasets extracted from three corpora.

[1] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

[2] It consists of about 17000 Japanese traditional nobels dataset. `https://github.com/aozorabunko/aozorabunko`

[3] It consists of about 1.3 million entities. `https://github.com/attardi/wikiextractor`

Table 1: The number of sentences extracted from three corpora.

	Pseudo-simile	Pseudo-literal
Aozora Bunko	39987	14320
Wikipedia	14449	36337
Mainichi Newspaper	3939	1788
All domains	58375	52445

Table 2: Fleiss κ for three annotators for each domain.

Domain	Fleiss κ
Aozora Bunko	0.437
Wikipedia	0.376
Mainichi Newspaper	0.341
Overall	0.426

3.2 Evaluation dataset

For the evaluation of simile identification, we need an evaluation dataset. However, our dataset created in Section 3.1 is an automatically acquired and classified dataset by translation and rules. In other words, the dataset essentially contains sentences with incorrect labels: e.g., a simile sentence with the literal label. Therefore, it is not suitable as the evaluation dataset. In this section, we manually create a dataset for the evaluation.

First, we randomly selected 300 sentences for each domain from the dataset in Section 3.1, namely 900 sentences. These sentences were deleted from the pseudo dataset. Then, we prepared nine annotators and assigned 100 sentences of a domain for each annotator, namely three annotators for one sentence. Each annotator judged whether a sentence was a simile, literal, or undecidable.

Table 2 shows the degree of agreement of each domain and entire annotation by calculating Fleiss κ[14]. Generally, it is said that when the value of κ is between 0.4 and 0.6, the annotation results match appropriately. As we can see from Table 2, the overall κ value exceeds 0.4. Hence, the annotation result contains a certain amount of confidence. However, the values of Wikipedia and Mainichi newspaper domains were less than 0.4. This result denotes the difficulty of judging simile or literal sentences, even for humans.

We adopt sentences that are completely matched within three annotators as the evaluation dataset. The distribution of the evaluation dataset is shown in Table 3. Although we prepared 900 sentences for the evaluation dataset, we just obtained approximately half of them, namely 500 sentences consisting of 217 similes and 283 literal sentences. This result also denotes the difficulty of simile identification. Moreover, the distributions of simile/literal in domains are imbalanced. In Aozora Bunko, simile sentences are the majority. On the other hand, the number of simile sentences is smaller than that of literal sentences in the Wikipedia and Newspaper domains. This result shows that the phrase "のような（に）" tends to be used as literal meanings in descriptive texts such as Wikipedia and newspapers, although it is often used as a simile in novels.

4 Simile Identification

The purpose of this study is to classify a sentence into "simile" or "literal". We apply machine learning models to this task. We compare three models: Support Vector Machine

Table 3: The distributions of simile and literal sentences in each domain.

	simile	literal
Aozora Bunko	124	55
Wikipedia	38	134
Mainichi Newspaper	55	94
Overall	217	283

(SVM), Naive Bayes, and BERT.

4.1 SVM

SVM [15] is a machine learning method that classifies two classes by supervised learning and was a widely-used model before the deep learning era. First, MeCab[4], a Japanese morphological analyzer, is used to divide each sentence into words. Then, each word is converted into a 300-dimensional vector by word2vec, which has been pre-learned with Japanese Wikipedia text. Here, we remove particles and auxiliary verbs from the sentence. Finally, the average vector of word embeddings in the sentence is used as sentence embedding for the input of SVM.

4.2 Naive Bayes

Naive Bayes [16] is also a supervised learning model based on the probability values calculated by Bayes' theorem. We use the surface information of words, while SVM uses word embeddings as the input. First, we also use MeCab, a Japanese morphological analyzer for dividing a sentence into words. Then, the sentence is vectorized by a Bag-of-Words model. Here, we exclude particles and auxiliary verbs as with SVM. The values of each vector are based on the total number of occurrences of words in the sentence.

4.3 BERT

BERT [17] is a general-purpose language model that was pre-trained using a large text corpus. The model can be adapted to various natural language processing tasks by fine-tuning the target task. We fine-tune the model using a pseudo-training dataset. We expect that the model utilized general knowledge and contextual information for the simile identification. We use the BERT model published by Tohoku University[5]. This model is pre-trained using a Japanese Wikipedia text corpus and uses MeCab and WordPiece by the tokenizer. The model adds the [*CLS*] token at the beginning of the token sequence and the [*SEP*] token at the end. The model determines whether the sentence is a simile or literal sentence on the basis of the value of the [*CLS*] token in the output layer.

5 Evaluation

We evaluate three models in Section 4 that are learned by the pseudo training dataset in Section 3.1 by the evaluation dataset in Section 3.2. First, we compare the recall, precision,

[4]https://taku910.github.io/mecab/

[5]https://github.com/cl-tohoku/bert-japanese

Table 4: Results of each model using the pseudo training data in Section 3.1.

	simile			literal		
	Precision	Recall	F-score	Precision	Recall	F-score
Baseline	0.627	0.613	0.620	0.708	*0.721*	0.714
SVM	0.641	0.834	0.725	0.834	0.643	0.726
Naive Bayes	0.664	0.811	0.730	0.826	0.686	0.749
BERT	*0.673*	*0.858*	*0.754*	*0.862*	0.680	*0.760*

and F-score of each model and a simple baseline. Then, we discuss a problem caused by imbalanced data. Finally, we introduce a re-labeling scheme to pseudo-data acquisition.

5.1 Experimental result on the original pseudo data

We introduce a simple baseline based on machine translation only for the evaluation. The baseline model classifies sentences that contain "like" in the English translation as simile sentences and the others as literal sentences. In other words, this is almost the same process as the pseudo training dataset acquisition[6].

As described in Section 3.1, we obtained 58,375 pseudo-simile sentences and 52,445 pseudo-literal sentences. Here we randomly selected 30,000 pseudo simile sentences and 30,000 pseudo literal sentences for the training data of three models. For the verification of BERT, we also randomly selected 300 similes and 300 literal sentences from the pseudo data. In this experiment, this random selection process was iterated five times for each model. We need to determine the final label from the five iterations. We apply the majority voting of the identification results of the five models. For example, if a sentence is identified as a simile three times or more, we assign the simile label to the sentence. The precisions, recalls and F-scores in the next section are computed based on the label.

For SVM, the cost parameter was 0.01, and the others were the default values of Python's scikit-learn. In a similar way, we used scikit-learn to implement the Naive Bayes model, and all parameters were also the default values. For BERT, the maximum number of tokens in each sentence is 50. If the number of tokens in a sentence exceeded 50, the excess was ignored. The optimization function was AdaDelta, and the loss-function was BCELoss. The learning rate and the batch size were 0.001 and 128, respectively. The maximum number of epochs was 30, and we used the model with the minimum loss during the 30 epochs in the validation as the optimal model.

Table 4 shows the experimental results of the baseline and three models. The italic values denote the maximum value of each criterion. On the F-score, the three machine learning-based models outperformed the baseline. This result shows the effectiveness of learning by the pseudo dataset even if the dataset contains incorrect labels.

As a tendency of all machine learning models, the recall rates of simile sentences were higher than those of literal sentences. The reason is that literal sentences cover a broad range of definitions of meaning while simile sentences are restricted to simile. The definition of literal sentences in this paper is just non-simile sentences. Therefore, literal sentences have much more variety than simile sentences. In other words, classification into the literal class is essentially a difficult task.

Among the three models, the BERT model produced the best average F-score. This

[6]The difference from the acquisition process is that the baseline does not use the keyword "as" for literal sentences, namely "like" or not.

Table 5: Recall of each domain

		SVM	NB	BERT
Aozora Bunko	simile	0.984	0.976	0.919
	literal	0.127	0.145	0.455
Wikipedia	simile	0.605	0.447	0.632
	literal	0.888	0.940	0.903
Mainichi Newspaper	simile	0.655	0.698	0.873
	literal	0.596	0.638	0.500
Overall	simile	0.834	0.811	0.858
	literal	0.643	0.686	0.680

Table 6: The number of sentences after down-sampling.

	pseudo-simile	pseudo-literal
Aozora Bunko	*14449*	14320
Wikipedia	14449	*14320*
Mainichi Newspaper	3939	1788
Overall	32837	30428

result indicates that it is important to pay attention to contextual information in the simile identification because SVM and Naive Bayes can not handle the contextual information.

Next, we discuss the results of each domain. Table 5 shows the recall rates of each domain. The tendency differs from domain to domain. For Aozora Bunko (novels), the models tended to classify each sentence into the simile class. On the other hand, they tended to classify each sentence to the literal class for Wikipedia. The tendency is similar to the distributions of each class in each domain (see Table 1). In other words, it indicates that the imbalanced distribution brought a negative impact on the evaluation[7]. As compared with SVM and Naive Bayes, the BERT model tends to identify literal instances in Aozora Bunko and simile instances in Wikipedia correctly, although these are minority labels in each domain. As a result, the overall result of the BERT became better than SVM and Naive Bayes in Table 4.

5.2 Experimental result on the balanced data

As mentioned in Section 5.1, the acquired pseudo dataset is imbalanced. Therefore, we generate a balanced dataset by down-sampling and evaluate the models by the dataset. We randomly down-sampled the instances of simile sentences of Aozora Bunko and literal sentences of Wikipedia. We did not carry out the down-sampling of Newspapers due to the size of the data. Table 6 shows the distributions of the balanced dataset. The italic values denote the changes.

We also compared the three models. In this experiment, we also randomly selected 30,000 similes and 30,000 literal sentences from 32,837 and 30,428 sentences, respectively. In a similar way to Section 5.1, the evaluation process was iterated five times. Table 7 shows the recall rates of each model in the balanced pseudo dataset. The bold values in Table 7 denote that they are better than the values in Table 5. Although the values decreased compared to Table 5, the difference between similes and literal sentences became smaller.

[7]The reason that the difference was not large in the newspaper domain was the size of the dataset. The number of sentences in Mainichi Newspaper was one-tenth the numbers in Aozora Bunko and Wikipedia.

Table 7: Recall of each domain by the balanced pseudo dataset.

		SVM	NB	BERT
Aozora Bunko	simile	0.766	0.782	0.669
	literal	**0.618**	**0.564**	**0.582**
Wikipedia	simile	**0.632**	**0.710**	**0.789**
	literal	0.821	0.821	0.754
Mainichi Newspaper	simile	0.691	**0.818**	0.609
	literal	**0.723**	0.596	**0.540**
Overall	simile	0.724	0.779	0.748
	literal	**0.749**	**0.696**	0.599

Table 8: Results of each model using the balanced pseudo dataset from Table 6

	simile			literal		
	Precision	Recall	F-score	Precision	Recall	F-score
Baseline	0.627	0.613	0.620	0.708	0.721	0.714
SVM	**0.689**	0.724	0.706	0.779	**0.749**	**0.764**
Naive Bayes	0.663	0.779	0.716	0.804	**0.696**	0.746
BERT	0.588	0.748	0.659	0.756	0.599	0.668

This result indicates the negative influence of the imbalanced dataset.

Table 8 shows the experimental results on balanced data of the baseline and three models. The values of the baseline are the same as Table 4. The bold values denote that the values are better than the imbalanced dataset (Table 4). In Table 8, the F-scores of Naive Bayes and SVM for literals increased: e.g., for SVM, 0.764 in Table 8 vs. 0.726 in Table 4. On the other hand, the F-scores for similes decreased: e.g., for SVM, 0.706 in Table 8 vs. 0.725 in Table 4. The balanced data do not always contribute to the improvement. Moreover, the F-scores of BERT for both labels dramatically decreased. Generally, machine learning models tend to classify each instance to the majority class because they are based on statistics. The difference between results from imbalanced and balanced datasets was caused by the characteristic of statistical machine learning models. In other words, although the balanced data provided a positive influence on the recall rates, especially Aozora Bunko and Wikipeida domains, it did not contribute to the improvement as a whole. Therefore, we need to discuss other approaches in the balanced data.

5.3 Experimental result on the re-labeled data

From the experimental results from Section 5.1 and Section 5.2, we confirmed that the pseudo-training dataset is effective in simile identification. However, the pseudo data acquired automatically in Section 3.1 contain sentences with incorrect labels. In general, noisy training data generate a weak classifier in machine learning. Therefore, it is necessary to polish the quality of the pseudo-training data. It leads to the improvement of the simile identification accuracy.

The original pseudo data were selected by a simple rule based on keywords and results of machine translation. This was the baseline in the experiment. On the other hand, machine learning methods outperformed the baseline in Section 5.1 and Section 5.2. In other words, the machine learning methods are more suitable pseudo data acquisition models, as compared with the rule and machine translation approach. Therefore, we create a better-

Table 9: The number of sentences included in balanced dataset and each re-labeled dataset.

	pseudo-simile	pseudo-literal
Balanced dataset	32837	30428
Re-labeling by SVM	41005	23360
Re-labeling by NB	35830	28535
Re-labeling by BERT	34645	29720

Table 10: Results of each simile identification model that was learnt by each re-labeled dataset from Table 9

Model	Dataset	Simile			Literal		
		Precision	Recall	F-score	Precision	Recall	F-score
SVM	Balanced	0.689	0.724	0.706	0.779	0.749	0.764
	Re-labeling by SVM	0.670	**0.853**	**0.751**	**0.857**	0.678	0.757
	Re-labeling by NB	**0.694**	**0.793**	**0.740**	**0.821**	0.731	**0.774**
	Re-labeling by BERT	**0.706**	**0.751**	**0.728**	**0.799**	**0.760**	**0.779**
NB	Balanced	0.663	0.779	0.716	0.804	0.696	0.746
	Re-labeling by SVM	0.656	**0.862**	**0.745**	**0.860**	0.654	0.743
	Re-labeling by NB	**0.664**	**0.802**	**0.727**	**0.819**	0.689	**0.749**
	Re-labeling by BERT	**0.658**	**0.806**	**0.725**	**0.821**	0.678	0.743
BERT	Balanced	0.588	0.748	0.659	0.756	0.599	0.668
	Re-labeling by SVM	**0.632**	**0.876**	**0.735**	**0.865**	**0.609**	**0.715**
	Re-labeling by NB	**0.648**	**0.794**	**0.713**	**0.809**	**0.669**	**0.732**
	Re-labeling by BERT	**0.594**	**0.757**	**0.665**	**0.763**	**0.602**	**0.673**

quality dataset by using each machine learning model. This is a re-classification task of original pseudo-simile and literal labels by machine learning.

The outline of the method has already been shown in Figure 2 of Section 1. We regard the pseudo data in Section 3.1 as input data of this process. We classify the data by a machine learning model with 10-fold cross-validation. The result from this process becomes a new pseudo-simile and pseudo-literal dataset, namely re-labeled data. On the basis of the new pseudo data, we learn a machine learning model (SVM, Naive Bayes, or BERT) for the simile identification task, in a similar way to the previous experiment.

In this experiment, we use the balanced dataset in Table 6 for the re-labeling process. For the classifier of re-labeling, we compare three machine learning models that were used in the previous experiment, namely SVM, Naive Bayes, and BERT. Hence, we obtained three new datasets from three re-labeling models. The distributions of the re-labeled datasets are shown in Table 9. As compared with the original balanced dataset, the number of pseudo-similes tended to increase. It is because models trained on pseudo data tended to classify instances into similes. In other words, the model tended to classify simile instances into non-similes incorrectly. As a result, the number of pseudo-similes increased in the re-labeled dataset.

Table 10 shows the experimental results on the original balanced dataset and re-labeled datasets about the three models. The values of the balanced dataset are the same as Table 8. The bold values denote that the values are better than the balanced dataset. There are many improvements, namely bold values, for similes on the F-score. Although the difference between BERT with balanced data and BERT with BERT re-labeling is marginal (0.659 vs.

0.665), some combinations improved 2 to 5 points[8] on the F1-score, e.g., 0.706 vs. 0.728 for SVM with balanced data vs. SVM with BERT re-labeling and 0.659 vs. 0.713 for BERT with balanced data vs. BERT with NB re-labeling. In particular, the F values of the similes based on the SVM-based relabeling dataset were dramatically improved. From the results, introducing machine learning-based re-labeling models to pseudo training data acquisition is effective for the simile identification task.

On the other hand, an overfitting problem must be considered in machine learning. In general, repeated training processes tend to overfit the model. As mentioned above, in the relabeling process, the number of pseudo-similes tends to increase. In other words, it became an imbalanced dataset by re-labeling process. If the number of pseudo-similes in the training data increases, the identification model tends to predict each instance as "simile" because machine learning approaches are based on statistics of the data. The decrease in the F-scores for SVM and Naive Bayes for literals was due to the overfitting problems. As discussed through Section 5.1 and Section 5.2, we need to consider the relation between the distribution of data and the accuracy of the model carefully. This is one important future work of our pseudo-training data acquisition.

6 Conclusion

We reported simile identification by machine learning for sentences containing comparators "のような" and "のように". One of the biggest problems is the construction of training data and its size. Manual construction is costly. Therefore, we proposed a method for acquiring a training dataset using machine translation. Our method automatically acquired pseudo data of simile and literal instances by using the translated outputs of target sentences. It used an assumption that comparators are translated to suitable words for simile identification in the translation process. We obtained approximately 110,000 instances from three domains: novels, Wikipedia, and newspapers. The method can easily be adapted to other corpora.

We applied the pseudo-training dataset to three machine learning models: SVM, Naive Bayes, and BERT. We compared the three models with a simple baseline based on machine translation only. We obtained higher average F-scores than the baseline. The result shows the effectiveness of machine learning with the pseudo-training dataset. Among the models, the BERT model produced the best F-score.

The acquired dataset was imbalanced. As a result, the recall rates of each domain were also biased. Therefore, we created a balanced dataset by down-sampling. Although the F-score decreased, the biased recall rates were improved. The improvement of the F-score with stable recall rates is important future work. One approach to solve this problem is scaling up the size of the dataset while down-sampling. Since our method is based on pseudo-data acquisition, we can generate additional pseudo-data if there is a larger pool of unlabeled data.

Since the rule for the pseudo data acquisition was very simple (just machine translation results and keywords), the dataset contains incorrect labels. To solve this problem, we introduced a re-labeling approach to the acquisition process. The approach was also based on machine learning models As compared with the original balanced dataset, the re-labeled dataset contributed to the improvement of the simile identification accuracy. The result shows the effectiveness of the re-labeling approach.

[8]Expressed as percentages for intuitive understanding.

Future work includes refining the pseudo-training data. In this paper, we just focused on two words: “like” and “as”. There are other suitable comparators for the task. Moreover, the word “as” is often used in simile sentences. Adding more appropriate words for the acquisition process is one important future work.

We handled “ような/ように” as the target. However, there are other expressions that relate to similes. Experiments with other datasets based on other simile expressions are also important for future work.

Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Number 23K11368.

References

[1] Takehiro Tazoe, Tsutomu Shiino, Fumito Masui, and Atsuo Kawai. The metaphorical judgment model for ”noun b like noun a” expressions. *Journal of natural language processing (in Japanese)*, 10(2):43–58, 2003.

[2] Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. Neural metaphor detection in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 607–613. Association for Computational Linguistics, 2018.

[3] Lizhen Liu, Xiao Hu, Wei Song, Ruiji Fu, Ting Liu, and Guoping Hu. Neural multitask learning for simile recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1543–1553, 2018.

[4] Kenichi Kobayashi, Junpei Tsuji, and Masato Noto. Application of data augmentation to image-based plant disease detection using deep learning. *The 79th nathional convention of IPSJ (in Japanese)*, 79(2):289–290, 2017.

[5] Ji Dang, Toya Matsuyama, Pang-Jo Chun, Jiyuan Shi, and Shogo Matsunaga. Deep convolutional neural networks for bridge deterioration detection by UAV inspection. *Intelligence, Informatics and Infrastructure*, 1(J1):596–605, 2020.

[6] Shinnosuke Nishimoto, Hiroshi Noji, and Yuji Matsumoto. Detecting aspect in sentiment analysis by data augmentation. *2017 The Association for Natural Language Processing (in Japanese)*, pages 581–584, 2017.

[7] Manabu Sassano. Using vitual examples for text classification with support vector machines. *Journal of natural language processing (in Japanese)*, 13(3):21–35, 2006.

[8] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv:1609.08144*, 2016.

[9] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

[10] Jintaro Jimi and Kazutaka Shimada. Pseudo data acquisition using machine translation and simile identification. In *2022 12th International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 391–396, 2022.

[11] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

[12] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

[13] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv: 2008.00401*, 2020.

[14] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[15] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[16] Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.