Classification of Causality between Two Sentences Based on Automatically Constructed Causal Knowledge

1910224 Yamada Ryota

A causal relation is a relation of cause and effect between two events, such as "it rains" and "the ground gets wet." In other words, it is a relation in which an event written in one sentence likely cause an event written in another sentence. A collection of pairs of sentences under the causal relation is called a causality database, which is regarded as one of the common sense knowledge. A causality database is useful knowledge for natural language processing, and used for reasoning in text understanding and for showing evidence of users' evaluation in opinion mining. Studies in natural language processing on the causal relation include retrieval of sentence pairs under the causal relation from a large amount of text, and classification whether two given sentences have the causal relation or not. In this study, a model to perform the latter task is called the causality classification model. Most of the previous studies are based on supervised machine learning that requires labeled training data. However, in general, it is necessary to manually annotate sentence pairs with gold labels indicating whether the causal relation is held or not. It is very costly and time consuming to prepare a large amount of a manually labeled dataset. Therefore, it is preferable to construct training data for the causality classification automatically. The goal of this study is to obtain a causality classification model by supervised machine learning without manually labeled training data.

The proposed method consists of the following steps: "Initial data construction", "Unlabeled data construction", "Training of model", "Evaluation of model", "Causality classification", "Data selection and addition", and "Decision on stop".

First, in "Initial data construction" step, we use the corpus of Mainichi Shimbun news article data to automatically collect sentence pairs under the causal relation using some heuristics. We define a conjunction indicating the causality between two clauses as "causal keyword". Complex sentences including a causal keyword are supposed that two clauses in them have the causal relation. Then, such complex sentences are retrieved from news articles. Next, pairs of clauses (single sentences) connected by the causal keyword are extracted from these sentences. In this study, two conjunctions "から (*kara*)" and "ので (*node*)" are used as the causal keywords. In addition, pairs of sentences that are not under the causal relation are made by randomly shuffling those collected sentence pairs. The constructed initial data is divided into training data, development data, and validation data. In "Unlabeled data construction" step, we collect sentence pairs using the causal

keyword "ため (*tame*)" in the same way, where some pairs of sentences have the causal relation and some do not.

The entire procedure of training of the causality classification model is performed by the bootstrapping method. In "Training of model" step, a model is trained using Bidirectional Encoder Representations from Transformers (BERT) with the initial training data and the development data. The development data is used to optimize the parameters of the model. In "Evaluation of model" step, we apply the trained causality classification model to the validation data and measure its accuracy on the causality classification. In "Causality classification" step, we apply the trained causality classification model to each pair of sentences in the unlabeled data and determine whether they have the causal relation. In "Data selection and addition" step, we get the value of the output node in the BERT model as the reliability of the decision, select instances with the high reliability score, and add them to the training data. By repeating the above procedures, the number of the training data is increased incrementally. In "Decision on stop" step, the iterative procedure is terminated when the accuracy of the current model becomes worse than that of the previous model. Here the accuracy is measured on the validation data in the step of "Evaluation of model". Through the above iterative learning, a large amount of training data is constructed without manual annotation, and a highly precise causality classification model is obtained.

We conducted an experiment to evaluate the proposed method. First, the initial data consisting of 2,236 instances were constructed by our method. Next, the correlation between the reliability of the classification of the BERT model and the accuracy were investigated. It was confirmed that the accuracy became high as the reliability increased, and reached 0.906 at maximum. Next, the causality classification model was applied to the unlabeled data and the most reliable 2,000 instances were chosen and added to the training data at each iteration step. After three iteration steps, the number of the training data was increased from 2,236 to 8,236. We prepared 200 sentence pairs as the evaluation data. Two workers judged whether each pair had the causal relation or not. The inter-annotator agreement was 0.72, and the kappa coefficient was 0.44. When the judgments of two workers did not agree, they discussed and determined the final label. The causality classification model trained by the proposed method was applied to the evaluation data and the accuracy of the classification was measured. The accuracy of the model trained from the initial data only was 0.475, while the accuracy of the model trained from the training data after two iterations was improved to 0.520. However, after the third iteration, the accuracy decreased to 0.495.

Next, the precision, recall, and F-measure on retrieval of positive sam-

ples (sentence pairs under the causal relation) as well as negative samples (sentence pairs not under the causal relation) were measured. Through the iterative learning, the F-measure for the positive samples was declined, while that for the negative samples was improved. The initial model failed to classify the causality for the negative samples, but the errors were reduced by incremental enlargement of the training data by the proposed method. From these results, it was confirmed that the training data automatically acquired by the bootstrapping method contributed to improve the quality of the causality classification model. However, the accuracy of the causality classification was not high, 0.520. It should be improved.

One of the future work is to improve the automatic construction of the initial data. Although we supposed that sentences including the causality keyword always represented the causality, we found that it was not always true and some instances were wrongly created as the positive samples. Therefore, it is necessary to develop rules that precisely select pairs of sentences in which the causal relation is truly held. In addition, the method of creating unlabeled data to extend the training data needs to be improved. Although we collected data using "ため (*tame*)" as the causality keyword, it is insufficient to retrieve the cause-effect sentence pairs exhaustively. Some sentence pairs under the causal relation may include another causality keyword, and some may not include any keywords.

It is necessary to develop more patterns that can extract the sentence pairs under the causal relation.