ISIKUN at the FinCausal 2020: Linguistically informed Machine-learning Approach for Causality Identification in Financial Documents

Hüseyin Gökberk Özenir
Işık University
219MCS8082@isik.edu.tr

İlknur Karadeniz
Işık University
ilknur.karadeniz@isikun.edu.tr

Abstract

This paper presents our participation to the FinCausal-2020 Shared Task whose ultimate aim is to extract cause-effect relations from a given financial text. Our participation includes two systems for the two sub-tasks of the FinCausal-2020 Shared Task. The first sub-task (Task-1) consists of the binary classification of the given sentences as causal meaningful (1) or causal meaningless (0). Our approach for the Task-1 includes applying linear support vector machines after transforming the input sentences into vector representations using term frequency-inverse document frequency scheme with 3-grams. The second sub-task (Task-2) consists of the identification of the cause-effect relations in the sentences, which are detected as causal meaningful. Our approach for the Task-2 is a CRF-based model which uses linguistically informed features. For the Task-1, the obtained results show that there is a small difference between the proposed approach based on linear support vector machines (F-score 94%) , which requires less time compared to the BERT-based baseline (F-score 95%). For the Task-2, although a minor modifications such as the learning algorithm type and the feature representations are made in the conditional random fields based baseline (F-score 52%), we have obtained better results (F-score 60%). The source codes for the both tasks are available online (https://github.com/ozenirgokberk/FinCausal2020.git/).

1 Introduction

The causality detection, which is a well-known task in text mining, has a goal of the extraction of the cause-effect relations from a given text. Although the causality detection task has been applied to various types of texts in various domains such as biomedical (Khoo et al., 2000; Mihăilă et al., 2013; Mihăilă and Ananiadou, 2014), clinical (Casillas et al., 2016), and financial (Blanco et al., 2008; Asghar, 2016; Kumar and Ravi, 2016) until now, there is still room for the improvement of these approaches.

The FinCausal-2020 Shared Task is a community-wide effort, which presents an opportunity for the researchers to test their financial causality detection systems on a common platform (Mariko et al., 2020). The FinCausal-2020 Shared Task consists of two sub-tasks, which are called Task-1 (sentence classification) and Task-2 (relation extraction). Task-1 is a binary classification task in which the participants are intended to classify the given financial sentences as 1 if the text section is considered containing a causal relation, and 0 otherwise. Task-2 is a relation extraction task in which the participants are intended to extract the text sections as a cause or an effect from the given text in which it is known that there is at least one causality relation.

In this paper, the two systems that are developed for the two sub-tasks of the FinCausal-2020 Shared Task are explained. For the Task-1, the problem is handled as a binary classification problem. We have used two different linguistically informed machine learning classifiers which are naive bayes (McCallum et al., 1998) and linear support vector machines. For the Task-2, which aims the automatic extraction of cause-effect relations from a given financial text, Conditional Random Fields (Lafferty et al., 2001) are used.

Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation, pages 85–89
Barcelona, Spain (Online), December 12, 2020.

2 Data Set

The data set, which is provided by the organizers, is created from 2019 financial news. The training data set for the Task-1 has entries which are the index of the sentence, the sentence, and the class label which the corresponding sentence belongs to, while the test data set entries consist of the index of the sentence and the sentence only without the class labels.

The test data set that is related with the Task-2 consists of the index of the sentence and the sentence entries, while the training data set has entries which are the index of the sentence, the sentence, as well as the cause mention, and the effect mention of the corresponding sentence.

3 Task-1: Sentence Classification

The workflow of our system for the Task-1 begins with the preprocessing phase of the data, which is followed by the feature representation of the sentences. After obtaining the sentences as feature vectors, different supervised machine learning classifiers are used to predict the given text as containing a causal relation or not-containing a causal relation. In the following subsections, a detailed explanation of the system that we have developed for the Task-1 is provided.

3.1 Preprocessing

For the preprocessing phase, the stop words, the punctuations and the abbreviations are removed to clear the uninformative words. Our assumption is that the uninformative words are the words which do not convey any meaning for the classification of the causality of the given sentences. With this assumption, as uninformative words, any numbers, currency signs and time specifications such as *"am, pm"* are removed. The intuition behind this pre-processing phase is that the sentences which include similar numerical specifications such as dates and the money amount may be classified as different classes. For instance, considering the two sentences in the training data set, one sample sentence begins with *"Third Democratic presidential debate September 12, 2019 at 9:54 PM EDT* - and an another sentence *"It found that total U.S. health care spending would be about $3.9 trillion under Medicare for All in 2019, compared with about $3.8 trillion under the status quo. Part of the reason is that Medicare for All would offer generous benefits with no copays and deductibles, except limited cost-sharing for certain medications."*. Although both sentences include the date mention *"2019"*, the first sentence has a causality relation (class 1), while the second sentence does not carry a causality meaning (class 0).

After the removal of the unnecassary words, the stems of the remaining words in the given text are obtained by utilizing nltk library's tokenizer (Loper and Bird, 2002). Since there are no distinct data sets such as training and development, the training data set is splitted by using 5-fold cross validation.

3.2 Feature Representations

Term frequency-inverse document frequency (TF-IDF) representation is used to represent the text sections which are treated as separate sentences. TF-IDF scoring computes the relative frequency that a word appears in a sentence compared to its frequency across all training data. In TF-IDF scoring, the IDF measure makes the words that are rare in the data set become more informative than the words which are more frequent in the data set. For the realization of the transformation of the given sentences into real-valued vectors, we used the TF-IDFTransformer as a vectorizer, which is provided by Python's scikit-learn library (Pedregosa et al., 2011).

Maximum document frequency value (*max df*) is used for removing terms that appear too frequently, which are also known as "corpus-specific stop words". The value *max df = 0.2* is heuristically used in the submission, where *max df = 0.2* means "ignore terms that appear in more than 20 percentage of the documents". On the other hand, minimum document frequency value (*min df*) is used for removing terms that appear too abundant. The value *min df = 3* is heuristically used in the submission, where *min df = 3* means "ignore terms that appear in less than 3 documents of the document collection".

3.3 Training the model

Two different classifiers such as support vector machines (SVMs) and Multinominal Naive Bayes (MNB) are used to classify the financial sentences in the training data set, whose results are shown at the Table 1. MNB produces an estimation using the frequency of usage of words, which are existed in the dictionary that is created with all sentences in the data set, therefore we need to use Count Vectorizer and TF-IDF that is explained in previous section. SVMs examine inputs to determine hyper-plane which is separate binary class samples.This hyper-plane separation is expected to be as larger as possible to label input correctly.In addition to MNB model, vectorizers and term frequency approaches are also used for training SVMs. For the reason that the data set is limited in size, k-Fold cross validation, where k is heuristically chosen as 5, is used for the validation of the data set. The best results on the training data set are obtained with the system based on SVMs that we submitted the system as the first run for the Task-1 in the shared task.

4 Task-2: Cause-Effect Detection

4.1 Preprocessing & Feature Representations

Firstly, the punctuations and the general stop words in English are removed. The remaining words are converted to lower case.After that, the word tokenizer in NLTK library is used to tokenize the sentences.

For the feature representation, for each word the following features, which are the Part-of-Speech (POS) tags, surface form of the words, the reduced form of the words (the removal of the last three letters and the last two letters), digit or not, title or not,whether each word is at the beginning of the sentence or not, and tuples, which are explained in detail below, are extracted.

A tuple is automatically created for each sentence in the training data set depending on the constituent word which is a part of the followings: a cause mention, an effect mention or none of them. For example, considering the following sentence, *'Transat loss more than doubles as it works to complete Air Canada deal.'* , the cause mention is *'it works to complete Air Canada deal.'* , while the effect mention is *'Transat loss more than doubles '* . If a word is located in the sentence as a cause mention, the word is assigned to ***"C"*** tuple, while a word is located in the sentence as an effect mention, the word is assigned to ***"E"*** tuple. Otherwise, If a word is not either a part of a cause mention or an effect mention, the word is assigned to ***"_"*** tuple. Our sample tuples for the corresponding sentence will be such as *[('Transat','E'), ('work','C'), ('as',"_") , . . .]* .

4.2 Method

For the Task-2, Conditional random fields (CRFs) ,which are the one of the statistical modeling methods that has been successfully applied to the natural language processing problems previously, are used. The open-source Python library scikit-learn (Pedregosa et al., 2011) is utilzied to train our model on CRF approach for the utilities such as cross-validation, and hyperparameter optimization. We have focused on CRF models by combining with Averaged Perceptron (AP) as learning algorithms of the CRF approach. As far as we know, for the baseline which is provided by the organizers, the stochastic gradient descent algorithm is used.

5 Results

The results for the Task-1 of the FinCausal 2020 Shared task, which are obtained on the training data set with 5-fold cross-validation, are shown at Table 1.

The results of the CRF-based basedline model with the base features and the L-BFGS learning algorithm for the Task-2 are shown at Table 2.

On the other hand, the results of the CRF-based system with the linguistically informed features and the average perceptron learning algorithm for the Task-2 are shown at Table 3 which is above the baseline.

The official evaluation results for the Task-1 of the FinCausal 2020 Shared task, which are obtained on the test data set, are shown in Table 4. The obtained results consist of performance metrics by using linear support vector machines. On the other hand, the post evaluation results for the Task-2 of the FinCausal

Algorithm	Classes	Precision	Recall	F-1 Score
Baseline (BERT)	0	0.96	0.97	0.97
	1	0.64	0.64	0.64
	Macro Avg	0.81	0.81	0.81
	Weighted Avg	0.95	0.95	0.95
Multinomial Naive Bayes	0	0.94	**1.00**	0.97
	1	0.87	0.17	0.28
	Macro Avg	0.91	0.58	0.63
	Weighted Avg	0.94	0.93	0.93
Linear Support Vector Machine	0	**0.97**	0.99	0.98
	1	0.79	0.32	0.57
	Macro Avg	0.89	0.68	0.75
	Weighted Avg	0.95	0.95	**0.96**

Table 1: *Task-1 results.* The results are also obtained in the training data set with 5-fold cross validation.

	Precision	Recall	F-1 Score
Cause	0.53	0.70	0.60
Effect	0.53	0.57	0.53
-	0.65	0.17	0.27
Accuracy			0.52
Macro Avg	0.56	0.48	0.47
Weighted Avg	0.55	0.52	0.49

Table 2: *The results of CRF based model with base feautures and L-BFGS learning algorithm (Baseline) for the Task-2.*

	Precision	Recall	F-1 Score
Cause	0.57	0.72	0.68
Effect	0.53	0.66	0.63
-	0.65	0.13	0.22
Accuracy			0.60
Macro Avg	0.62	0.54	0.52
Weighted Avg	0.60	0.58	0.55

Table 3: *The results of CRF based model with linguistically informed features and the Averaged Perceptron learning algorithm (Our system) for the Task-2.* The results are obtained in the training data set with 5-fold cross validation.

Task (System)	Precision	Recall	F-1 Score	Exact Match
Task-1 (Our system)	0.93	0.94	0.94	1.000
Task-1 (Baseline)	0.95	0.95	0.95	1.000
Task-2 (Our system)	0.62	0.59	0.60	0.006
Task-2 (Baseline)	0.51	0.52	0.51	0.111

Table 4: *The test data test results* The results are obtained in the test data set at the official evaluation (for Task-1) and the post-evaluation (for Task-2) phases.

2020 Shared task, which are obtained on the test data set, are shown in Table 4. The obtained results consist of performance metrics by using averaged perceptron algorithm in CRF. The results show that for both tasks, the obtained results on the training data set with 5-fold cross validation and the results on

the test data set are similar.

6 Conclusion

In this study, we have presented two systems that are implemented in the scope of the FinCausal 2020 Shared Task. The aim of the first system is the binary classification of the sentences in a financial text as carrying causality relation or not-carrying causality relation, whereas the goal of the second system is the extraction of the cause-effect relations in the financial text. In conclusion, compared to the baseline, for the Task-1, there is a small difference (1%) between the results of the proposed approach based on linear support vector machines which requires less time compared to the BERT-based baseline. For the Task-2, on the other hand, better results are obtained compared to the baseline.

Acknowledgements

We would like to thank the FinCausal shared task organizers for organizing the shared task and for their help with the questions.

References

Nabiha Asghar. 2016. Automatic extraction of causal relations from natural language texts: a comprehensive survey. *arXiv preprint arXiv:1605.07895*.

Eduardo Blanco, Nuria Castell, and Dan I Moldovan. 2008. Causal relation extraction. In *Lrec*, volume 66, page 74.

Arantza Casillas, Koldo Gojenola, Alicia Pérez, and Maite Oronoz. 2016. Clinical text mining for efficient extraction of drug-allergy reactions. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 946–952. IEEE.

Christopher SG Khoo, Syin Chan, and Yun Niu. 2000. Extracting causal knowledge from a medical database using graphical patterns. In *Proceedings of the 38th annual meeting of the association for computational linguistics*, pages 336–343.

B Shravan Kumar and Vadlamani Ravi. 2016. A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, 114:128–147.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Edward Loper and Steven Bird. 2002. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.

Dominique Mariko, Hanna Abi Akl, Estelle Labidurie, Stephane Durfort, Hugues de Mazancourt, and Mahmoud El-Haj. 2020. The Financial Document Causality Detection Shared Task (FinCausal 2020). In *The 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation (FNP-FNS 2020*, Barcelona, Spain.

Andrew McCallum, Kamal Nigam, et al. 1998. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer.

Claudiu Mihăilă and Sophia Ananiadou. 2014. Semi-supervised learning of causal relations in biomedical scientific discourse. *Biomedical engineering online*, 13(2):1–24.

Claudiu Mihăilă, Tomoko Ohta, Sampo Pyysalo, and Sophia Ananiadou. 2013. Biocause: Annotating and analysing causality in the biomedical domain. *BMC bioinformatics*, 14(1):2.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.