LangResearchLab_NC at FinCausal 2020, Task 1: A Knowledge Induced Neural Net for Causality Detection

Raksha Agarwal
Indian Institute of Technology Delhi
Raksha.agarwal@maths.iitd.ac.in

Ishaan Verma
Manipal University Jaipur
Ishaanverma97@gmail.com

Niladri Chatterjee
Indian Institute of Technology Delhi
niladri@maths.iitd.ac.in

Abstract

Identifying causal relationships in a text is essential for achieving comprehensive natural language understanding. The present work proposes a combination of features derived from pre-trained BERT with linguistic features for training a supervised classifier for the task of Causality Detection. The Linguistic features help to inject knowledge about the semantic and syntactic structure of the input sentences. Experiments on the FinCausal Shared Task1 datasets indicate that the combination of Linguistic features with BERT improves overall performance for causality detection. The proposed system achieves a weighted average F1 score of 0.952 on the post-evaluation dataset.

1 Introduction

The understanding of cause-effect relation is an important NLP task because it appeals to human perception, reasoning, and decision-making. It has vast applications in the field of Information Extraction (Chan et al., 2002), Question Answering (Girju, 2003), and Event Prediction (Radinsky et al., 2012), among others. However, modeling causality relations between events is a non-trivial task because it requires a deeper analysis of the discourse and sometimes external knowledge to forge the relationship between separate events and entities.

In the present work, the focus is on detection of causal relationships in a given text, which is modeled as a binary classification task. Sometimes the presence of causal connectives, such as causes*, because of, leads to, after, due to* indicates causality. However, there may be cases when the causal relation is more implicit making the task of causality detection more challenging.

A causal relationship in a sentence involves the presence of a *cause* and an *effect*, where the cause triggers the effect. In other words, two events X and Y are considered to be causally related if the occurrence of X is triggered by the occurrence of Y, or vice versa. For illustration, consider the following: (1) *Fluctuations in exchange rates added to the risk factors*.
(2) *The company withdrew from bidding.*

It can be observed that the occurrence of (1) resulted in the occurrence of (2). Although there is no explicit marker, the association between the *risk* and *bidding* helps to forge a causal relationship.

Past studies revealed that two major approaches for causality detection involve the use of handcrafted features (Riaz and Girju, 2014) or deep neural networks (Liang et al., 2019). The proposed work integrates syntactic and semantic features of the input text with pre-trained embedding vectors to train a supervised neural network for the classification of causal relations.

The rest of the paper is organized as follows. The proposed system is described in Section 2, Section 3 contains implementation details, and experimental results are presented in Section 4.

2 Model Architecture

The proposed model aims to supplement pre-trained BERT (Devlin et al., 2019) embeddings with linguistic features in order to induce knowledge about the syntactic and semantic peculiarities of the

Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation, pages 33–39
Barcelona, Spain (Online), December 12, 2020.

input in the model. The model architecture is presented in Figure 1. The Linguistic feature vector and the BERT embeddings are concatenated together to generate a linguistically informed representation of the input text. The enhanced representations are processed using two identical layers of fully connected feed-forward network before applying a softmax classifier. The details of linguistic features are presented in Sec 2.1, and BERT features is described in Sec 2.2.

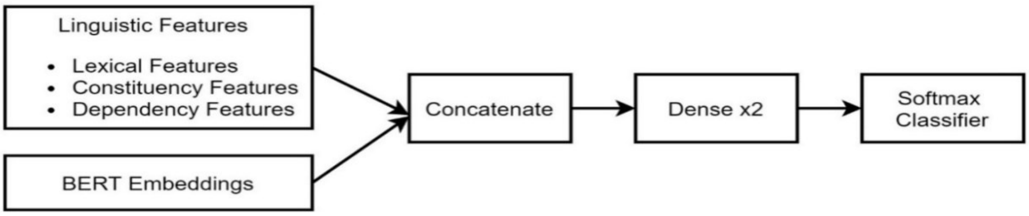


Figure 1: Proposed Model Architecture

2.1 Linguistic features

A combination of lexical and syntactical features is employed for the task of causality detection. These features enable us to encode the semantic information of the input tokens and the structural information of the input sentences.

Lexical Features

The lexical features proposed by Pitler et al. (2009) proved to be effective in sense prediction of implicit discourse relations (Prasad et al., 2008) between pair of input sentences. In the present work, a subset of lexical features is adapted for the task of detecting causality in an input text.

- **Polarity Tags:** The sentiment of each word in the input text is assigned according to the Multi-perspective Question Answering Opinion (MPQA) corpus (Wilson et al., 2005). The number of positive, negative, and neutral words in the input text were considered as features.
- **Inquirer Tags:** The General Inquirer lexicon (Stone et al., 1966) is used to assign semantic categories to the verbs present in the input text. The association between the different verb categories acts as an indicator of causality.
- **Money/Percent/Num:** This feature is used to determine whether the input text contains numbers, monetary amounts, or percentages. These entities frequently occur in financial texts and are useful in determining causality. The count of each such occurrence in the input is considered as a feature.
- **Verbs:** Levin Verb Classes (Levin, 1993) are used to identify verbs that belong to the same verb class. The average verb phrase lengths in the input text are also considered as a feature.
- **Modality:** Pitler et al. (2009) demonstrated that the presence of modal words such as *can, should, may* most likely relate to a contingency or causal relation between sentences. Therefore, a feature indicating the presence of a modal word is used in the present application.
- **Connective:** Text containing connectives belonging to the contingency class such as *because, as a result, consequently* are more likely to indicate causal relationships. A list of connectives is extracted from the Penn Discourse Treebank (Prasad et al., 2008), and a feature indicating the presence of connective in the input text is created.

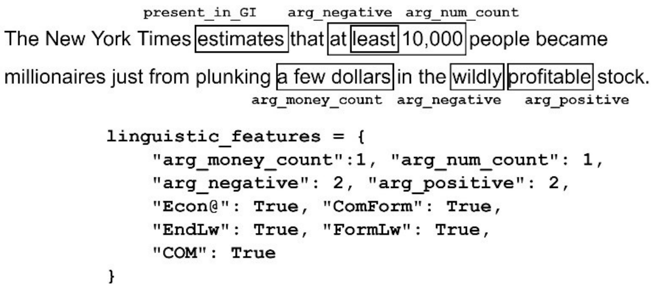
Syntactical Features

Lin et al. (2009) demonstrated the features extracted from syntactical trees of sentences helps in recognizing implicit discourse relations between pairs of sentences. Since the structure of a text can also indicate the presence of causal relations, two kinds of syntactical features are extracted from the input text.

- **Constituency Parse Features:** Production rules are extracted from the constituency tree of each sentence of the input text. A binary feature represents the presence of each production rule in a given input text.
- **Dependency Parse Features:** Dependency rules are extracted from the dependency parse trees of each sentence of the input text. For each word of the sentence, the dependency rule consists of the POS tag of the word along with a list of all dependency types from the dependents of the word. The presence of each dependency rule is indicated using a binary feature.

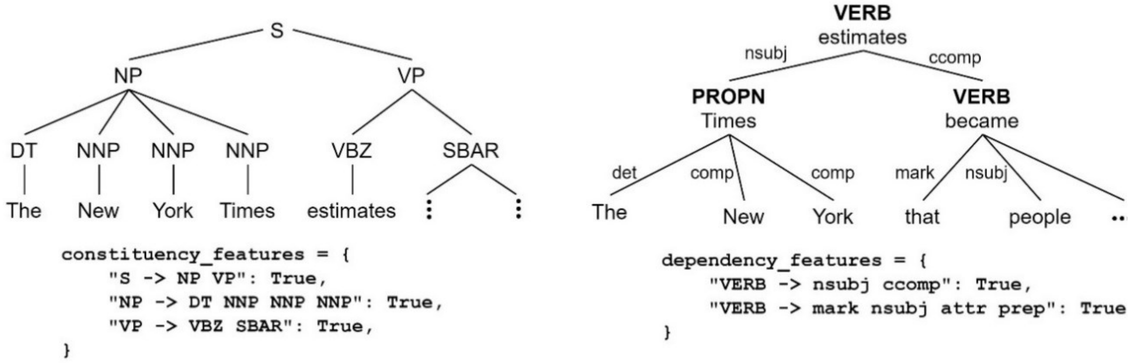
The rationale behind using both dependency and constituency features is that that precision and recall are increased when both parsing based features are used as the dependency trees encode additional information about the relationship between words of the sentences.

For illustration, consider the sentence *The New York Times estimates that at least 10,000 people became millionaires just from plunking a few dollars in the wildly profitable stock*[2]. A subset of linguistic features is described in Figure 2. Here, *Econ@, ComForm, EndLw, FormLw, COM* are the Inquirer tags for the word *estimate*. Figure 3 depicts a subset of constituency and dependency features extracted from the respective constituency and dependency parse trees.


present_in_GI
arg_negative
arg_num_count
The New York Times estimates that at least 10,000 people became
millionaires just from plunking a few dollars in the wildly profitable stock.
arg_money_count
arg_negative
arg_positive

```
linguistic_features = {
    "arg_money_count":1, "arg_num_count": 1,
    "arg_negative": 2, "arg_positive": 2,
    "Econ@": True, "ComForm": True,
    "EndLw": True, "FormLw": True,
    "COM": True
}
```

Figure 2: Linguistic Features



```
constituency_features = {
    "S -> NP VP": True,
    "NP -> DT NNP NNP NNP": True,
    "VP -> VBZ SBAR": True,
}
```

```
dependency_features = {
    "VERB -> nsubj ccomp": True,
    "VERB -> mark nsubj attr prep": True
}
```

Figure 3: Syntactic Features

It was observed that the linguistic features resulted in the creation of sparse feature space. Therefore, Singular Value Decomposition (SVD) is applied on three subsets of linguistic features, namely Lexical features, Constituency Parse Features, and Dependency Parse features. The Lexical features are reduced to a 100-dimensional space, and the Constituency and Dependency Parse features are reduced to 2000 dimensional space each, resulting in a 4100-dimensional feature vector.

2.2 BERT Embeddings

Bidirectional Transformers for Language Understanding (BERT) was introduced by Devlin et al. (2019), and the usage of BERT features has resulted in state-of-the-art performance for various downstream NLP tasks such as Question Answering, Textual Entailment and Paraphrase detection. In the present work, input embeddings are extracted from the pre-trained BERT-base-uncased[3] model. The output from the last layer corresponding to the [CLS] token is considered as the input text embedding.

3 Implementation Details

Pre-processing: In the pre-processing step, SpaCy[4] library is used to perform Tokenization, Lemmatization, Sentence Segmentation, Part-of-Speech (POS) tagging and Dependency Parsing. SpaCy's Named Entity Recognizer is employed to identify entities belonging to Cardinals, Monetary amounts, and Percentages. Constituency parsing is derived using the benepar_en2 model (Kitaev and Klein, 2018).

[2] present at Index 0026.00057 in the validation data

[3] https://storage.googleapis.com/bert_models/2018_10_18/uncased_L-12_H-768_A-12.zip

[4] https://github.com/explosion/spaCy

Dataset: The FinCausal Shared Task (Mariko et al., 2020) provides three datasets for Task1 viz., the *Practice-Task1*, the *Trial-Task1* and the blind dataset *Evaluation-Task1*. The aforementioned datasets are used for training, validation and testing, respectively. Oversampling of positive samples was performed during the training process to balance the dataset.

The proposed model is implemented on Python using keras[5] framework. ReLU activation is applied on the intermediate Dense layers along with dropout regularization. The proposed model is trained to minimize the cross entropy loss using the Adam optimizer (Kingma and Lei Ba, 2015). Hyperparameter optimization is performed using keras-tuner[6]. The encoding dimensions of the Dense layers are tuned on the set {256, 512, 1024, 2048, 4096} and dropout ratio is tuned between {0.1, 0.2, …, 0.9}. The optimal encoding dimension and dropout ratio was found to be 2048 and 0.1, respectively. All the experiments were conducted on Google Colab[7] using the Intel Xeon CPU @ 2.3GHz, the Nvidia Tesla P100 GPU and 25GB available RAM

4 Results and Analysis

In this section we present the results of our experiments on the blind Evaluation-Task1 dataset. The evaluation metrics are Precision, Recall and Weighted F1 score. Weighted F1 score is calculated by multiplying the class wise F1-scores with the class *support*, i.e. the number of examples in that class. The results corresponding to different subsets of the feature space is given in Table 1. It can be observed that linguistically enhanced input representations improve the ability of the supervised model to detect causal relationships in a given text.

Features	**Precision**	**Recall**	**Weighted F1 score**
Only Lexical Features	0.943	0.946	0.936
Only Constituency Features	0.934	0.940	0.936
Only Dependency Features	0.934	0.941	0.930
Only Syntactical Features	0.942	0.946	0.935
Only BERT Features	0.943	0.935	0.938
Only Linguistic Features	0.947	0.950	0.942
Linguistic + BERT Features(Evaluation)	0.950	0.949	0.951
Linguistic + BERT Features (Post Evaluation)	**0.951**	**0.954**	**0.952**

Table 1: Test Results on the blind dataset

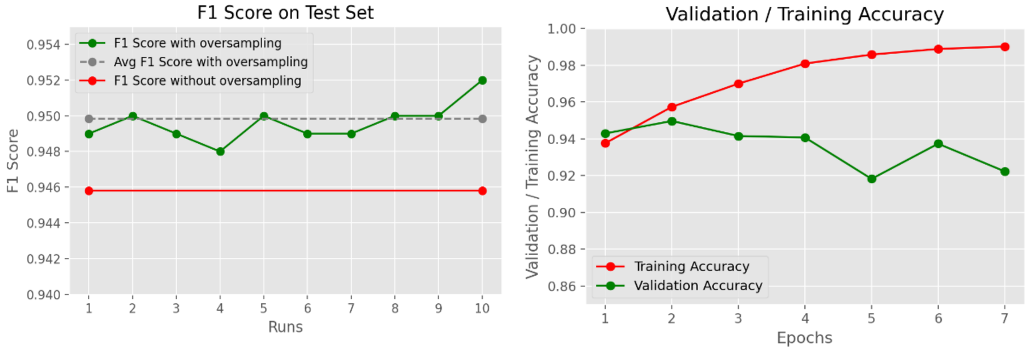


Figure 4:(a) F1 Score of the proposed model on Test Set with and without oversampling with different sampling seeds (b) Validation and Training accuracy for the proposed model (Early Stopping returns the weights of second epoch)

[5] https://keras.io/

[6] https://keras-team.github.io/keras-tuner/

[7] https://colab.research.google.com/

The effect of oversampling in the proposed model is demonstrated in Figure 4(a) by using different sampling seeds. The F1 score on the test set of the model trained without oversampling is lower than the average F1 score of the model trained with oversampling[8]. Training and Validation accuracies are shown in Figure 4(b). Early Stopping[9] call back is used to get the weights of the second epoch to avoid overfitting because validation accuracy stops increasing after this epoch.

Predictions of the proposed model for samples taken from the Validation set are described in Table 2. Example 1 and 2 are correctly classified while Examples 3-8 are incorrectly classified. The causality in Example 3 is between *capital out* (withdrawal of capital) and *refinancing*. However, the lexical and syntactical structures of the input sentence conceal the underlying causal relationship. In Examples 4 and 5 the difference in numeric quantities have an underlying causal effect. This indicates that knowledge with respect to variation of numeric quantities may help in improving performance of the model. The reason for unemployment in Example 6 is excluded from the input text and thus, the gold label is not causal. However, due to the presence of the connective phrase *as result* the proposed model assigns causality. The proposed model picks up on the complementary relationship between verbs *earn* and *pay* to predict causality in Example 7. The gold label for Example 8 indicates the absence of causality which is incorrect because the presence of the connective *as* strongly suggest the presence of a causal relation indicating that since 67 persons sold their share it dived. Thus, the proposed system is able to pick up on linguistic clues for meaningful predictions.

Input Text		**Index**	**Gold**	**Predicted**
1	Choice Hotels International has a consensus target price of $85.12, suggesting a potential downside of 8.49%.	0194.00006	Causal	Causal
2	Around the world fiduciaries are struggling with the challenging investment outlook.	0016.00011	Not Causal	Not Causal
3	I refinanced my apartment and took almost 30,000 euros of capital out of my home.	0311.00009	Causal	Not Causal
4	The S&P 500 returned 4.3%, after a 13.6% gain in the March quarter.	0088.00025	Causal	Not Causal
5	Keep in mind that an 8% annual return is really only a 5% annual return after 3% inflation.	0126.00018	Causal	Not Causal
6	It said 1,300 jobs would be lost as result, with a further 3,400 in the supply chain put at risk.	0366.00003	Not Causal	Causal
7	Anyone earning below $2 million a year will not pay a dime.	0102.00016	Not Causal	Causal
8	It dived, as 67 investors sold RTN shares while 352 reduced holdings.	0288.00031	Not Causal	Causal

Table 2: Predictions of the Proposed Model on the Validation Set (*Trial-Task1*)

5 Conclusion

Causality detection in a text is a challenging task due to the semantic peculiarities of the English language and also because it requires a deeper domain understanding. In the present work, semantic and structural knowledge of the input text is induced on the top of input embeddings to generate enhanced representation. The results indicate that the enhanced representations improve the performance across all the evaluation metrics. In future work we would like to experiment with more complex architectures such as LSTMs and Transformers. Additionally, we would also like to experiment with CNN and max-pooling based dimensional reduction for treatment of the sparse linguistic feature space.

[8] Best results are obtained for sampling seed 5050

[9] https://www.tensorflow.org/api_docs/python/tf/keras/callbacks/EarlyStopping

Acknowledgements

Raksha Agarwal acknowledges Council of Scientific and Industrial Research (CSIR), India for supporting the research under Grant no: SPM-06/086(0267)/2018-EMR-I

References

Ki Chan, Boon Toh Low, Wai Lam, and Kai Pui Lam. 2002. Extracting causation knowledge from natural language texts. *International Journal of Intelligent Systems*, 20:327–358.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Roxana Girju. 2003. Automatic Detection of Causal Relations for Question Answering. In *Proceedings of ACL 2003 Workshop on Multilingual Summarization and Question Answering*, pages 76–83.

Diederik P Kingma and Jimmy Lei Ba. 2015. Adam: A Method For Stochastic Optimization. In *Proceedings of the 3rd international conference for learning representations (ICLR'15)*, San Diego, California.

Nikita Kitaev and Dan Klein. 2018. Constituency Parsing with a Self-Attentive Encoder. In *Proceedings of the 56th Annual Meeting ofthe Association for Computational Linguistics*, pages 2676–2686.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.

Shining Liang, Wanli Zuo, Zhenkun Shi, and Sen Wang. 2019. A Multi-level Neural network for Implicit Causality Detection in Web Texts. arXiv: 1908.07822 v2 *http://arxiv.org/abs/1908.07822*

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing Implicit Discourse Relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 343–351, Singapore.

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.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 683–691, Singapore.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning to Predict from Textual Data. *Journal of Artificial Intelligence Research*, 45:641–684.

Mehwish Riaz and Roxana Girju. 2014. In-depth Exploitation of Noun and Verb Semantics to Identify Causation in Verb-Noun Pairs. In *Proceedings of the SIGDIAL 2014 Conference, The 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 161–170.

Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvia. 1966. The General Inquirer: A Computer Approach to Content Analysis. *MIT Press*.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354.