What Causes Causality? Explaining a Deep Causal Text Detection Model

No Author Given

No Institute Given

Abstract. Causal relation extraction is a challenging task with implications in text mining and information retrieval. Recent deep learning approaches have achieved performance improvements over traditional methods, but can be complex, require large training data, and lack interpretability. In this work, we conduct an explainability analysis on a deep causal text detection model trained on a large corpus of sentences from the web. By repeated forward propagation of noised word embeddings through our trained model, we constructed saliency maps that represent word importance in a sentence. Qualitative review of these saliency maps reveal trends in model behavior namely that effect phrases, connectives, and qualifying terms are all important in our causal detection model. These results can be used to inform of future causal relation extraction work.

Keywords: Causal Relations · NLP · Explainable AI

1 Introduction

Causal relationship extraction is an ongoing challenge in the field of Natural Language Processsing (NLP) due to its importance in downstream information retrieval (IR) tasks such as question answering [7]. Expressing causality through natural language can take many different forms. Causality can be stated explicitly (e.g. "mosquito bites cause malaria") where the cause and effect is explicitly stated [1, 12]. Moreover, the sense of causality can be *marked* (e.g. "Increased thirst is a symptom of diabetes") or *unmarked* (e.g. "Last week temperature rose significantly, there were several cases of heat stroke reported"). In the second example, although it is apparent that the rising temperature is the *cause* of the heat stroke cases, no causal marker is present. Past works on causality extraction from text has mostly focused on explicit and marked causality [8, 5, 3, 12]. Early approaches to address this problem relied on linguistic patterns and rule-based methods, subsequently, machine learning methods were proposed that involved manual feature engineering. Most of these methods are not scalable and may not work accurately for cases not seen in the training data [14].

Much like in other areas of NLP, more recent efforts in causal relation extraction have utilized larger datasets and increasingly complex model architectures. These

approaches generally provide marginal performance boosts but need substantially larger training data [15]. Model interpretability is also desirable because it improves trust, increases informativeness, and enables transferability, among other reasons [16]. The now dominant deep learning methods for causal relation extraction sacrifice the interpertability once afforded by rule-based approaches. This work addresses the complexity and interpretability issues simultaneously by investigating how a deep causal detection model is making successful decisions. We apply a method from explainable AI, saliency maps [23], to produce model explanations that we then qualitatively analyze to find patterns in model behaviour. Thus, we are able to give the reader a better idea of what "causes" causality in a deep learning model, and therefore inform future causal relation extraction model development. Additionally, the explainability method we use can be applied to any language model that uses word vector encoding, so future efforts can apply our method in the model development process.

We utilized large, web-based text data sources to train our model. Using the model we created saliency maps to show how individual words in sentences influenced the models prediction. We then categorized maps produced from the test data into one of four categories depending on the proportion of important words in each sentence. Within these categories, some words had intuitively high importance, but unexpectedly important words are also revealed. Namely, casual phrases, effect phrases, connectives, and qualifiers were important words across most sentence categories. Additionally, we observed that if a sentence contained off-label cause and effect phrases, these were found to be important as well. These qualitative results highlight the elements of causal sentences that are generally important. Applying this method to other causal text models could inform development via informed improvements to their architecture, transfer of potentially missing knowledge, or supplementation with rule-based methods.

2 Related Work

Early methods on causal relation extraction were rule-based [8, 2, 13]. One method used lexico-syntactic patterns and semantic constraints on causal terms to achieve an average of 65.6% accuracy relative to human reviewers [9]. This approach, and other early rule-based approaches, were limited by the syntactic templates implemented by the authors. This work was followed by approaches that combined rule based feature extraction with machine learning techniques [4]. For example, an approach using Naïve Bayes and linear SVMs applied to document-term matrices was able to classify causal relations in marine accident reports with an average F-measure of 0.65. CausalTriad [24] is a minimally supervised approach, based on focused distributional similarity methods and discourse connectives, for identifying causality. Other approaches focused on extracting causal relationships from text by exploiting linguistic structures. Roemmele and Gordon [21], on the other hand, addresses the task of predicting causally related events in stories according to a standard evaluation framework, the Choice of Plausible Alternatives (COPA). Paul et al. [20] used causal inference to find

causal relationships between word features and document labels for better feature engineering. Past works have also utilized several linguistic features to extract causality from text, such as multi-word expressions [22], N-grams, topics and sentiments [13], and lexical patterns [2, 8].

Recent causal relation extraction has moved towards deep learning approaches that outperform older methods substantially. One approach for causal classification used LSTMs fed by Glove word embeddings [17]. CNN based approaches have been applied successfully to causal relation detection as well [6]. A recent deep-learning based causal extraction approach used BiLSTM-CRF with transferred Flair embeddings to achieve a state-of-the-art F1 of 0.8455 on the SemEval 2010 task 8 dataset [15]. Dasgupta et al. [5] used a bi-directional LSTM model to extract causal relationships detection, where they used lexical feature-based k-means clustering to cluster Cause-Effect events. More recently, language models have been proposed that are able to detect causal association among events expressed in natural language text [14]. Additionally, some hybrid methods that rely on a combination of deep learning and rule-based methods have seen a degree of success [19].

3 Methods

3.1 Data Sources

CauseNet [10] is a collection of causal sentences mined from ClueWeb12 and Wikipedia. CauseNet offers a high precision (0.96) subset of nearly 200,000 causal relations supported by around 950,000 unique evidence sentences. We randomly selected 500,000 evidence sentences from the precision subset to use as example causal sentences in our dataset. To produce a set of non-causal sentences we used an export of English Wikipedia from Wikimedia. We filtered to exported pages in the "Main/Article" namespace and removed redirect pages. The remaining pages were parsed and sentence tokenized and then randomly sampled down to 500,000 sentences. The combined labeled dataset contained 1 million sentences consisting of 27,304,629 tokens. Since it was possible that our Wikipedia sampler could have selected causal sentences at random, we checked if any of the 500,000 non-causal sentences were in the CauseNet precision subset. Of the 500,000 non-causal sentences, there were 4 sentences that matched to CauseNet, so we excluded these from the explainability analysis.

3.2 Model Creation

Sentences in the final dataset were tokenized at the word level with spaCy's "en_core_web_sm" model [11]. Punctuation tokens were removed. To encode the lemmatized sentences as input to our model, we used the Gensim implementation of word2vec [18]. We used a vector size of 512, minimum word count of 1, window length of 8, and 20 training epochs. Each token in the lemmatized sentence data

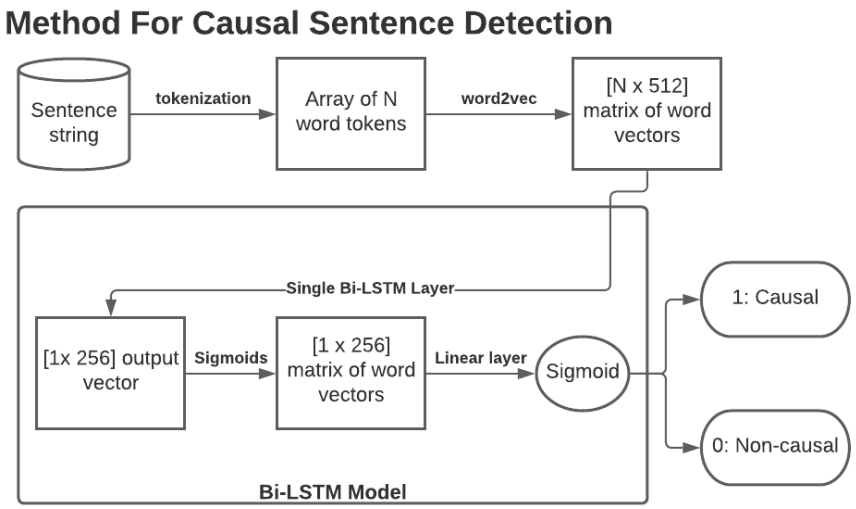

Fig. 1. Overall method for classifying a sentence as causal

was encoded using the trained word2vec model. Our final dataset consisted of sequences of word vectors labaled as either causal (label = 1) or noncausal (label = 0). The data was split into 60% training, 20% validation, and 20% test sets.

Past deep causal models have relied on various architectures like LSTMs [5] and CNNs [6], but recent work has indicated that Bi-LSTM models may better capture word context across at entire sentence, leading to improved performance [15]. Incorporating attention mechanisms has also improved the performance of causal relation extraction models. However, to more directly attribute the results of our explainability analysis to individual model components, we opted to use the simplest high performing model. Therefore, our classification model consisted of a Bi-LSTM layer followed by dense linear layer with a binary cross entropy loss function. Hyperparameters were chosen using iteration on the validation set, ultimately leading to a single layer Bi-LSTM with 3 training epochs. The trained model was evaluated on the test subset and achieved good performance at predicting causal sentences (F1 = 0.9859457, Sensitivity = 0.9999485, Specificity = 0.9723748, AU-ROC = 0.9865215). The overall model architecture is detailed in Figure 1.

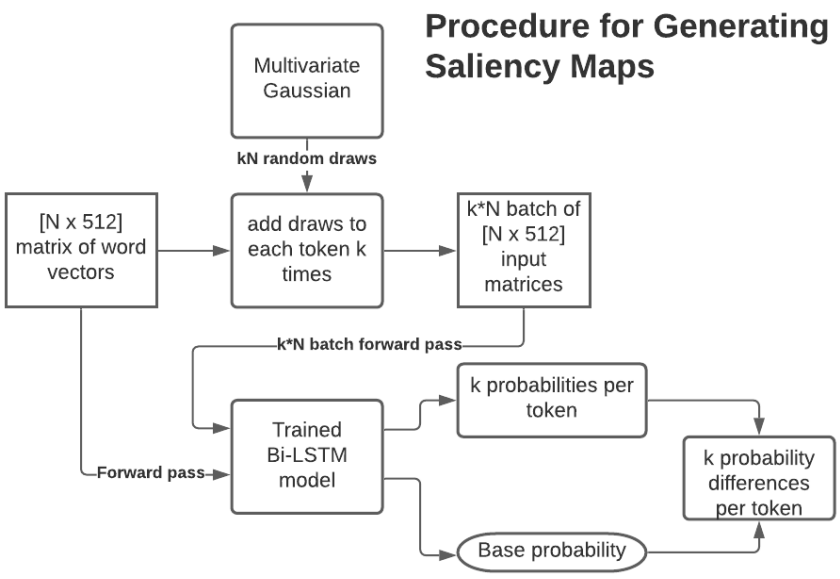

Fig. 2. Method for generating saliency maps

4 Experiments

4.1 Explainability Analysis

With an accurate model trained on a large dataset, we now have the goal of explaining how the model is making decisions. To conduct our explainability analysis, we adopted the idea of saliency maps [23]. The original saliency maps were developed for CNNs, and produced an importance value for each input pixel, which was then mapped back to an image with the same dimensions of the original. This process creates visual representations of relative pixel importance that show important regions in the original image. Rather than using a sole backpropagation pass to create a single valued saliency map, as described in [23], we developed a method to create saliency maps that, for each word, produce a distribution of output values via repeated forward propagation on noised inputs. From this distribution, we can see changes in probability relative to changes in a token, and thus gain some understanding of that token's importance in a sentence. By looking at importance values in tokens across a sentences, we can attempt to explain why the model may be classifying that sentence as causal.

For each token in each sentence, we took k noise samples from a multivariate normal distribution with mean $\mu = 0$ and covariance matrix $\Sigma = Is_d$ where I is the 512×512 identity matrix and s_d is a scaling parameter. Each noise

sample was added to the token's vector encoding, and a forward evaluation of the Bi-LSTM model was used to recalculate the classification probability. These probabilities were subtracted from the baseline classification probability of the un-noised sentence, resulting in k probability differences for each token. Finally, probability differences that were less than a fixed threshold th were set to 0. For computational expediency, all forward passes for a given sentence were batch calculated. This process is outlined in Figure 2.

The parameters s_d, k, and th, were determined through iteration. Since the noise is added to a word vector, we can think of s_d as controlling the magnitude of word vector "differences", so s_d should somewhat depend on properties of the word vector space. We took the element-wise standard deviations of every word vector in our vocabulary and observed that these were centered around 0.36 (min = 0.326, median = 0.362, mean = 0.364, max = 0.449, N = 512). We set s_d to 0.36 so that the variance of the noise distribution matched the variance of the word vector distribution, thus the addition of noise somewhat approximated choosing a random meaning from the vector space. Setting k to 10 seemed to produce stable results without imposing an unnecessarily high computational load. The purpose of using the threshold th is to ignore probability differences caused by computational noise. After some trial and error, this was set to 1e-6.

4.2 Visualization Results

We conducted the explainability analysis on a random subset of 4,880 causal sentences from the test data. For each sentence, the distribution of probability differences was plotted against each token in the sentence. Among these visualizations, were 100 randomly selected for qualitative analysis. Select plots are shown in Figure 3.

The visualizations should be read by comparing the distribution of probability differences against 0. A distribution with a near-0 mean and small standard deviation indicates very little probability difference when the corresponding token was noised, thus it was not important to the prediction. On the other hand, relatively high means with respect to zero are indicative of an average drop in probability when the token was noised, indicating that it was somewhat important to classification.

4.3 Sentence Categorizations

The plots were analyzed qualitatively to determine categories of behavior. A common feature of almost every plot is the disproportionately high importance of the last token. This is most likely due to the Bi-LSTM architecture used by the model. Therefore, we ignore the last token of each sentence for the remainder of the analysis. After manually reviewing the 100 plots, we found 4 general categories. Category counts are shown in Table 1. Category descriptions with examples are as follows

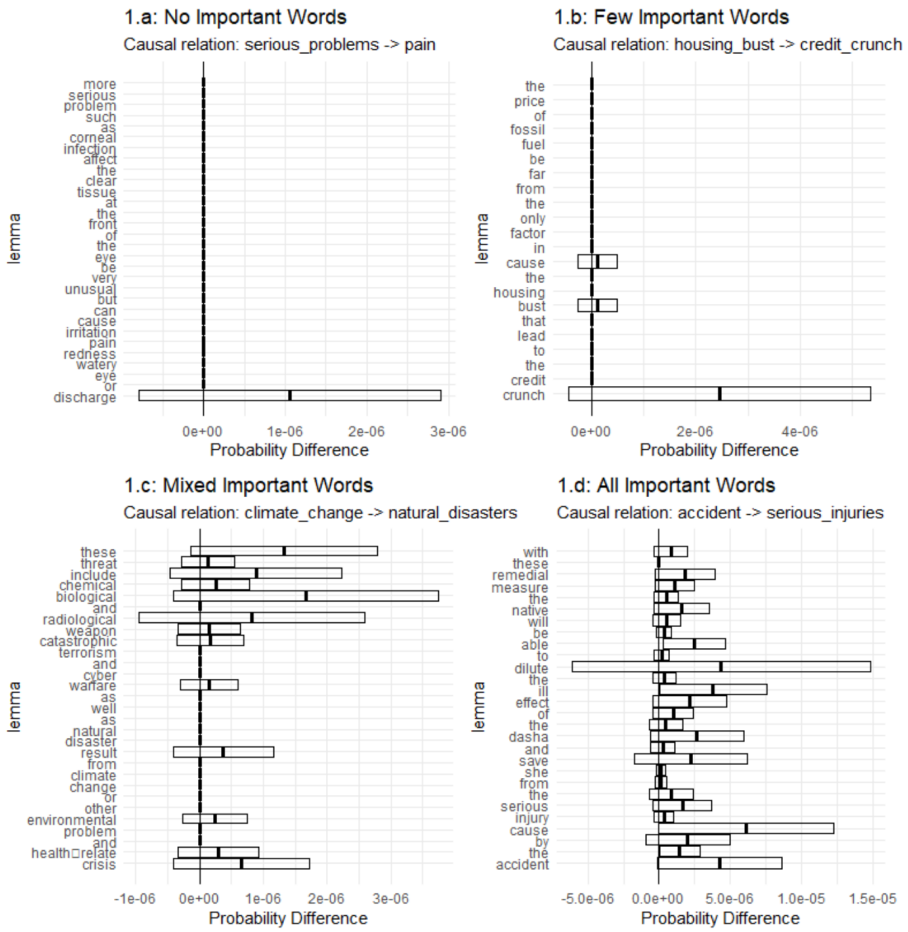

Fig. 3. Selected saliency maps. The horizontal axis corresponds to probability difference, while the vertical axis corresponds to the sequence of lemmatized tokens in the analyzed sentences. The center marks on the boxes correspond to the mean of probability differences, while the error bars indicate a standard deviation in each direction. A vertical line was drawn at a probability difference of zero.

Table 1. Distribution of sentence categorizations among the 100 reviewed sentences.

Categorization	**Count**
No Important Words	28
Few Important Words	27
Mixed Important Words	16
All Important Words	29
Total	100

- **No Important Words**: Noising of tokens caused no probability difference above the threshold th, other than for the final token. These cases are essentially uninformative.
  - "more serious problems (such as corneal infection, affecting the clear tissue at the front of the eye) are very unusual but can cause irritation, pain, redness, watery eyes or discharge."
- **Few Important Words**: Only a few noised token contributed to significant probability differences. The example in Figure 3b shows the words "cause" and "bust" to be important while all other words are not.
  - "the price of fossil fuels was far from the only factor in causing the housing bust that led to the credit crunch."
- **Mixed Important Words**: A roughly even combination of important and unimportant words.
  - "these threats include chemical, biological and radiological weapons, catastrophic terrorism and cyber warfare, as well as natural disasters resulting from climate change or other environmental problems, and health-related crises."
- **All Important Words**: All tokens contributed to probability differences when noised. Usually, there are a few tokens in these sentences that stand out as the most important.
  - "with these remedial measures the native will be able to dilute the ill effects of the dasha and save her from the serious injuries caused by the accident."

4.4 Discussion

Categories other than "No Important Words" were informative. Important words in the "Few Important Words" category were typically part of cause or phrases, or connectives like "cause" and "result", as seen in plot Figure 3b. Qualifying terms such as "believe", "important", "[in] turn", "general" were also important in many of these sequences. In several sentences, causal phrases unrelated to the labelled causal phrase were important. For example, in the following sentence

> "the article cites the following statistic: in scotland last year tobacco caused 13,000 deaths; alcohol, 2,052; illegal drugs 356; yet both alcohol and tobacco are legal."

The first occurrence of the word "alcohol" and the final word were the only important tokens, despite the fact that the labelled cause and effect were "tobacco" and "deaths" respectively.

Sentences in the "Mixed Important Words" category highlighted similar tokens to the "Few Important Words" category, but additionally highlighted regions of interest in a sentence, as seen in Figure 3c. Sentences in the "All Important Words" category were best read by looking at the the most important tokens, which were usually those with standard deviation bars that did not include 0. In the example sentence in Figure 3d, such tokens are "able", "ill", "cause" and "the". The token "ill" is interesting because it qualifies the effect phrase. In another example sentence "The slightest herbicide drift will cause damage", the words "slightest" and "cause" were the most important tokens in this sense. Again, the most important words were part of a cause phrase and it's qualifier. Multiple causal nouns phrases in a sentence, even those not in the CauseNet derived label, were often deemed important. For example, in the following sentence, the tokens "loud", "aging", "toxins", and "hearing" were all important:

> "loud noise, aging, drugs, and other toxins can all harm these cells and cause hearing loss."

In general, cause and effect phrases seem to be important, which should be expected. In all categories except "All Important Words", the last token in a cause or effect phrase is usually the most important, due to the Bi-LSTM model structure. Connectives such as "cause" and "result" are important as well. Qualifiers, appearing anywhere in the sentence, are also frequently important. An interesting, and perhaps unexpected feature, of this causal detection model is that multiple causal phrases in a sentence may be simultaneously detected as important. In some sense, the model is greedy. Although only one causal noun phrase is needed to qualify a sentence as causal, the model is somehow using additional noun phrases as well.

5 Conclusion

In this paper, we trained a highly accurate causal sentence detection model on a dataset of 1 million causal and non-causal sentences collected from the web. We then demonstrated how saliency maps could be used to analyze which word tokens were most important for causal sentence detection. By examining patterns of token importance across several sentences in the test data, we qualitatively determined some broad categories of model behaviors. Within most of these categories, many important tokens are as expected such as those in cause and effect phrases, connectives, and qualifiers. Additionally, the model has a greedy property. Our results provide insight into how this particular model is considering causality, and how our method can be applied in general to other models to aide in their development.

References

1. Blanco, E., Castell, N., Moldovan, D.I.: Causal relation extraction. In: Lrec (2008)

2. Bollegala, D., Maskell, S., Sloane, R., Hajne, J., Pirmohamed, M.: Causality patterns for detecting adverse drug reactions from social media: Text mining approach. JMIR public health and surveillance **4**(2) (2018)
3. Bui, Q.C., Nualláin, B.Ó., Boucher, C.A., Sloot, P.M.: Extracting causal relations on hiv drug resistance from literature. BMC bioinformatics **11**(1), 101 (2010)
4. Chang, D.S., Choi, K.S.: Causal relation extraction using cue phrase and lexical pair probabilities. In: International Conference on Natural Language Processing. pp. 61–70. Springer (2004)
5. Dasgupta, T., Saha, R., Dey, L., Naskar, A.: Automatic extraction of causal relations from text using linguistically informed deep neural networks. In: Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue. pp. 306–316 (2018)
6. De Silva, T.N., Zhibo, X., Rui, Z., Kezhi, M.: Causal relation identification using convolutional neural networks and knowledge based features. International Journal of Computer and Systems Engineering **11**(6), 696–701 (2017)
7. Girju, R.: Automatic detection of causal relations for question answering. In: Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12. pp. 76–83. Association for Computational Linguistics (2003)
8. Girju, R., Moldovan, D.I.: Text mining for causal relations. In: Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference. pp. 360–364. AAAI Press (2002), http://dl.acm.org/citation.cfm?id=646815.708596
9. Girju, R., Moldovan, D.I., others: Text mining for causal relations. In: FLAIRS conference. pp. 360–364 (2002)
10. Heindorf, S., Scholten, Y., Wachsmuth, H., Ngonga Ngomo, A.C., Potthast, M.: Causenet: Towards a causality graph extracted from the web. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. pp. 3023–3030 (2020)
11. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: Industrial-strength Natural Language Processing in Python (2020). https://doi.org/10.5281/zenodo.1212303, https://doi.org/10.5281/zenodo.1212303
12. Ittoo, A., Bouma, G.: Extracting explicit and implicit causal relations from sparse, domain-specific texts. In: International Conference on Application of Natural Language to Information Systems. pp. 52–63. Springer (2011)
13. Kang, D., Gangal, V., Lu, A., Chen, Z., Hovy, E.: Detecting and explaining causes from text for a time series event. arXiv preprint arXiv:1707.08852 (2017)
14. Khetan, V., Ramnani, R., Anand, M., Sengupta, S., Fano, A.E.: Causal bert: Language models for causality detection between events expressed in text. In: Intelligent Computing, pp. 965–980. Springer (2022)
15. Li, Z., Li, Q., Zou, X., Ren, J.: Causality extraction based on self-attentive bilstm-crf with transferred embeddings. Neurocomputing **423**, 207–219 (2021), publisher: Elsevier
16. Lipton, Z.C.: The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue **16**(3), 31–57 (2018)
17. Martínez-Cámara, E., Shwartz, V., Gurevych, I., Dagan, I.: Neural disambiguation of causal lexical markers based on context. In: IWCS 2017—12th International Conference on Computational Semantics—Short papers (2017)
18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)

19. Mueller, R.M., Huettemann, S.: Extracting causal claims from information systems papers with natural language processing for theory ontology learning. In: Proceedings of the 51st Hawaii international conference on system sciences (2018)
20. Paul, M.J.: Feature selection as causal inference: Experiments with text classification. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). pp. 163–172 (2017)
21. Roemmele, M., Gordon, A.: An encoder-decoder approach to predicting causal relations in stories. In: Proceedings of the First Workshop on Storytelling. pp. 50–59 (2018)
22. Sasaki, S., Takase, S., Inoue, N., Okazaki, N., Inui, K.: Handling multiword expressions in causality estimation. In: IWCS 2017—12th International Conference on Computational Semantics—Short papers (2017)
23. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
24. Zhao, S., Jiang, M., Liu, M., Qin, B., Liu, T.: Causaltriad: Toward pseudo causal relation discovery and hypotheses generation from medical text data (2018)