ISSN 2007-9737

On Causality Problem in Natural Language Processing Field

Altynay Yerkhassym[1], Alexandr A. Pak[1,2], Iskander Akhmetov[1,2], Amir Yelenov[1,2], Alexander Gelbukh[3]

[1] Institute of Information and Computational Technologies, Kazakhstan

[2] Kazakh-British Technical University, Kazakhstan

[3] Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico

{yerkhassym.altynay, aa.pak83, iskander.akhmetov, greamdesu, gelbukh}@gmail.com

Abstract. Natural language processing (NLP) field has been developing rapidly recently. This article consists mainly of literature review of the basic understanding and solving the causality problem in natural language processing field. Existing models may benefit from the concept of causality because conventional language models are brittle and spurious [10]. Incorporating the principle of causality could assist in resolving this issue. Since this issue affects seriously on the accuracy value of NLP methods and algorithms, it is worth paying attention to. Content of the article includes the authors who have been covered this topic and have made researches respecting mentioned problem, the results that have been achieved, the methods and approached that have been used and the data that was used in researches.

Keywords. Natural language processing, neural network, causality.

1 Introduction

Natural language processing is a subfield of Artificial intelligence branch focused on allowing computers to perceive human language. NLP-based systems are primarily designed to comprehend and interact with human voice and text. Companies and organizations throughout the world are increasingly utilizing NLP-enabled solutions to obtain client information and enhance the automation of regular procedures.

These tools do numerous tasks, including translation, keyword extraction, subject classification, etc. To automate these procedures and provide precise results, however, machine learning is required. Machine learning is the application of algorithms that train machines to automatically learn from experience and improve without being explicitly programmed.

AI-powered chatbots, for instance, employ natural language processing to read what users say and what they mean to do, and machine learning to automatically provide more correct responses by learning from previous interactions.

However, this accuracy is never 100 percent, as determining causation remains a challenging task for machine learning algorithms and, consequently, natural language processing. More examples are used to train natural language processing models in an effort to tackle these issues.

As the environment becomes increasingly complicated, however, it becomes impossible to cover the full distribution by adding more training instances. Due to a lack of comprehension of cause-and-effect interactions, it is extremely challenging to generate accurate predictions and successfully adapt to novel situations.

The machine can forecast the outcome of every action, but this does not mean its predictions are always accurate.

Since there is always a possibility that certain events do not fit particular patterns. In this instance, the outcome is erroneous. In contrast to humans, who can construct a causal logic and forecast a more accurate output based on collected data, machines are incapable of doing so.

2 Authors who Have Addressed this Topic in their Works and Articles

In contrast to numerous challenges in natural language processing, the causal relationship has not been thoroughly examined. Recently released studies and works on this topic provide additional evidence. To comprehend how to implement this term into the work of algorithms and models, it is necessary to comprehend this element itself.

Determining the causal relationship in natural language, including analysis and psycholinguistics, is therefore the initial step in the investigation of this subject. Torgrim Solstad and Oliver Bott had made some researches on this topic and had written the article named Causality and causal reasoning in natural language [15].

This article offers a synopsis of theoretical and psycholinguistic approaches to causation in language. The primary phenomenological focus of the paper is on causal relations as articulated intra-clausally by verbs (such as break and open) and inter-clausally by discourse markers (e.g. because, therefore).

Special consideration is given to Implicit Causality verbs that elicit explanation expectations in the succeeding conversation. The article also analyzes linguistic terms, such as counterfactual conditionals, that do not convey causation as such but appear to require a causal model for their proper interpretation.

The study of the phenomena is supplemented with a summary of key characteristics of their cognitive processing as revealed by psycholinguistic research. Due to the strong relationship between machine learning and natural language processing, the problem of causality in machine learning is reflected in the NLP discipline.

Bernhard Scholkopf wrote an article [13] that can serve as an introduction to some relevant concepts of graphical or structural causal models for a machine learning. Algorithms and methods of Artificial Intelligence cannot reason and make decisions like humans or animals. They neglect numerous variables that can influence pattern formation and depend solely on generalized models based on uniformly distributed data. In addition, these models are poor in imagining and navigating imagined spaces.

The author thinks that causality, with its emphasis on modeling and reasoning about treatments, can make a significant contribution to understanding and resolving these challenges, thereby advancing the science.

Further continuation of the previous article can be found in the work of Bernhard Scholkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner Anirudh Goyal, Yoshua Bengio named Towards Causal Representation Learning [14].

In this article, the authors describe different levels of causal and statistical modeling, investigate the Independent Causal Mechanisms (ICM) principle as a key component that enables the estimation of causal relations in artificial intelligence agents, and examine existing methods for learning causal relations.

Primarily, authors present examples of causality and machine learning in scientific applications and hypothesize on the benefits of merging the skills of both domains to create a more adaptable AI. The extraction of causal patterns from natural language texts and using it in methods of natural language processing systems are described in the article written by P. Maslov [11].

This research proposes a technique for extracting and characterizing causal facts from Russian business prose documents. In addition, the implementation of the derived cause-and-effect relationships within the algorithm for anticipating severe scenarios is offered.

The majority of modern techniques are either lexico-semantic pattern matching or feature-driven

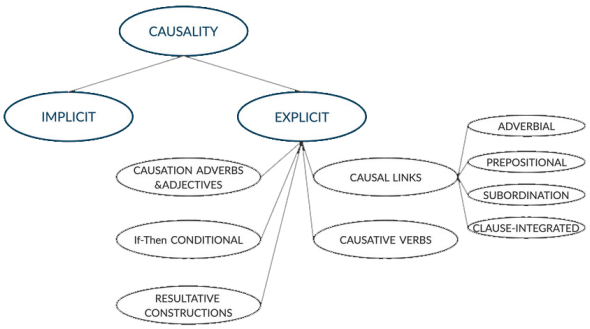

Fig. 1. Causality in natural language

supervised techniques. Consequently, as anticipated, these methods are better suited for managing explicit causal links, with limited coverage for implicit relationships, and are difficult to generalize.

In the paper written by Vivek Khetan, Roshni Ramnani, Mayuresh Anand, Subhashis Sengupta, Andrew E. Fano [9] they investigate the language models capabilities for causal association among events expressed in natural language text using sentence context combined with event information, and by leveraging masked event context with in-domain and out-of-domain data distribution.

A more specific application of causal relations is given in the work written by Son Doan, Elly W.Yang, Sameer Tilak and Manabu Torri [3]. Using natural language processing techniques, the authors assessed a method for extracting health-related causal linkages from Twitter conversations.

The analysis of health-related tweets would assist us comprehend the health conditions and worries we face on a regular basis, especially in the present day.

3 Methods and Approaches

Understanding causal relationships between commonplace events is crucial for common sense language comprehension. The majority of existing causality comprehension techniques rely on language pattern-matching rules or feature engineering to train supervised machine learning algorithms.

The types and structure of causality are illustrated in Figure 1. This section focuses on the methods and approaches utilized in works that directly address the topic in the field of natural language processing.

The paper [9] focuses on understanding the causality between events expressed in natural language text. The intent is simply to identify possible causal relationships between marked events implied by a given sequence of text.

Authors causality understanding approach can be simplified as a binary classification of “Cause-Effect” / “Other” relationship between events expressed in natural language text. The methodology in this work involves:

— Fine-tuning BERT based feed forward network for Cause Effect/Other relationship label between events expressed in natural language text. In this network architecture, authors feed the input sentence as a sequence of tokens to the BERT model and take the overall sentence context vector from the BERT models [2] output, feed it to a non-linear activation layer followed by two fully connected layers. Mathematical formulation for C-BERT model is:

$$H_0' = W_0(\tanh(H_0)) + b_0, \quad (1)$$

$$h'' = W_1(H_0) + b_1, \quad (2)$$

$$p = \mathrm{softmax}(h''), \quad (3)$$

where W_0, W_1, H_0 is the output token of bi-directional context (i.e. [CLS]) of BERT, and L = 2 (Cause-Effect, Other).

— Combining both the events context and BERTs sentence context to predict “Cause-Effect” / “Other” relationship label between events. This methodology works on the intuition that the interaction between two events is result of the information in the sentence as well as in the events.

They can be more than a single token, resulting in many vectors when the input sentence is fed into a pre-trained BERT model. Authors averaged them to get the final context of each event expression and passed the

sentence context as well as both the events context to a non-linear activation layer followed by a fully connected layer. The sentence context is concatenated with both the events averaged context and is feed to another fully connected layer followed by a softmax layer.

The model is trained using backpropagation with Adam-optimizer on a binary loss function to predict the "Cause-Effect" / "Other" relationship between events.

— Combining both the events masked context with BERTs sentence context to predict "Cause-Effect" / "Other" relationship label between events. This network architecture is very similar to the event aware C-BERT network architecture, where the whole span of event text is replaced with a "BLANK" token.

As each event is just a single blank token, unlike Event aware C-BERT we dont need to take an average to get the final context of any event. Each model trained by this approach is then fine tuned using actual event information using the Event Aware C-BERT model described above.

P. P. Markov in his article [11] facts describing cause-and-effect patterns are understood as text objects $s_i \in S$ (the set of vertexes of noun groups, predicates, and definitions that are syntactically consistent with subjects), semantically related by relationships $R_C \subseteq S \times S$, $R_A \subseteq S \times S \times S$ and by relationship groups $RE \subseteq S \times S$.

More detailed description of the connections between objects is also provided. The relationship properties are specified by means of attributes $A = a\{r,\ v\} \in R \times V$, where V is the set of valid attribute values.

Attributes are divided into $A_A \in A$ to describe the properties of symmetry, transitivity, reflexivity, etc. and $AV \in A$ to indicate the values of standard types, for example, to indicate the probabilistic characteristics of cause-and-effect relationships.

The result is formed by searching for all possible substitutions in the arguments of cause-and-effect relationships R_C , taking into account the ordering of objects. In this case, first of all, the facts whose arguments have the maximum weight are output, then, respectively, by reducing the weight. Authors in [3] use two methods for extraction the health related causality from tweets. They are:

— Natural Language Processing (NLP) pipeline. The NLP pipeline for extracting causal relation is summarized as follows: First, the corpus is filtered using the target keywords. Next, a series of basic NLP components are applied: sentence splitter, Part-of-Speech (POS) tagger, and dependency parser. Finally, causal relations are identified based on syntactic relations generated by the dependency parser.

— Cause-Effect Relation Extraction. Authors created a set of six general rules to identify cause-effect relationship from verb and noun phrase. Those rules are based on syntactic relations derived from a dependency graph generated by a dependency parser. For example, a Semgrex [16] pattern =subj < subj (word: /cause/=target > dobj =cause) finds a match in a sentence Stress caused my insomnia, where Stress is matched with the pattern =subj and insomnia is matched with the pattern =cause. Using Semgrex, we extracted the triple <cause, relation, effect> from tweets, where effect is one of the three health-related topics of our focus: insomnia, stress and headache.

The final step is to extract causality from extracted cause effect relations. To do so, we extracted the triple <cause, relation, effect>, where effect is one of the three health-related topics of our focus: insomnia, stress and headache.

4 Data Used in Researches

Authors in work [9] use three different datasets to train and evaluate the models described in previous section. Semeval 2007 [4] and Semeval 2010 [7] is curated using pattern-based web search while ADE is curated from a biomedical text as in [5].

Table 1. Statistics for curated datasets

		Train Dataset		
Dataset	**Max Sentence Length**	**Total**	**Cause-Effect**	**Other**
Semeval2010	(85, 60)	8000	1003	6997
Semeval2007	(82, 62)	980	80	900
ADE	(135, 93)	8947	5379	3568
		Train Dataset		
Dataset	**Max Sentence Length**	**Total**	**Cause-Effect**	**Other**
Semeval2010	(85, 60)	2717	134	2389
Semeval2007	(82, 62)	549	46	503
ADE	(135, 93)	2276	1341	935

Table 2. Comparison of F1 score of models

	Semeval2007	**Semeval2010**	**ADE**
C-BERT	93,78	97,68	97,10
Event Aware C-BERT	94,94	98,35	97,85
Masked Event C-BERT + Event Aware C-BERT	95,31	97,85	97,85

1. SemEval 2007 is an evalution task designed to provide a framework for comparing different approaches to classifying semantic relations between nominals in a sentence. For this work, authors use part of the SemEval 2007 dataset with the Cause-Effect relationship.

   For a given sentence, if the interaction between marked events is causal, they label it as “Cause-Effect” else the sentence is labeled as “Other”.

2. Similar to the above dataset, authors use SemEval 2010 dataset with causal interaction between events labeled as “Cause-Effect”, and all the other types of interactions between events in rest of the sentences are labeled as “Other”.

3. ADE dataset [6] is a collection of biomedical text annotated with drugs and their adverse effects.

The first corpus of this dataset has drugs as well as effects annotated. In the second corpus, where drugs are not causing any side-effect, the drug and its effect name are not manually annotated.

Authors curated a list of unique drugs and affect names using the first corpus data and use this set to annotate the drugs and effect names in the second corpus.

While they take sentences with two or more drugs/effect mention in them; for simplicity, we do not replicate the sentence in our final corpus.

To evaluate the precision of causal relation extraction, authors compared system outputs with human annotations. Table 4 provides us with the comparison results. P. Maslov [11] does not allocate a particular dataset. His work is designed using texts of the Russian business prose genre.

Business prose is defined by its strict means of expression, unambiguity of the transmitted information, economy of language means, clarity of the function of each communication, and other advantageous characteristics.

This genre provides information on objects (events, phenomena, people, etc.) that can be represented by an abbreviation of facts provided directly in the examined text. 24 million tweets were employed in the job of identifying health-related causality from Twitter messages [3].

This information was collected over a four-month period from four cities (New York, Los Angeles, San Francisco, and San Diego) (Sep 30, 2013 and Feb 10, 2014). The Twitter Streaming API was utilized to retrieve 1% of all tweets from these cities during the specified time frame. Three terms were chosen as the intended “effects”: stress, sleeplessness, and headache.

5 Achievements in this Field

To be more accurate about the outcomes of applying causality phenomena in natural language processing, It was more suitable to present the results of [11, 9, 3] as the methodologies and datasets have already been described.

The authors of [9] constructed three distinct BERT-based network architectures on each of the datasets to evaluate the language model’s ability to understand the “cause - effect” relationship between events.

Table 2 compares the performance of our models developed utilizing three distinct network architectures and trained on in/out of domain

Table 3. F1 score after pre-training on masked event C-BERT model (dataset 1) and fine-tuning on event aware C-BERT (dataset 2)

Dataset1 \ Dataset2	Semeval2007	Semeval2010	ADE
Semeval2007	95,31	98,42	97,27
Semeval2010	97,14	98,38	97,47
ADE	96,42	98,49	97,85

Table 4. Precision of extracted causal relations when comparing to human annotators

	Strict evaluation	Relax evaluation
Insomnia	73,81%	88,10%
Stress	82,65%	96,04%
Headache	56,10%	85,37%
Micro-average	74,59%	92,27%

data distribution to previously reported F1 performance metrics.

Table 3 shows the result of another set of experiments where authors examine the performance of the models [8] when pretraining and fine-tuning are conducted using in-domain data distribution rather than when pretraining is performed using out-of-domain data distribution.

They pre-trained three models for each of the target data distributions using the other out of domain data distribution. In general, pre-training on a dataset distinct from the target data distribution resulted in either comparable or enhanced performance.

According to [11] the presented method is at the stage of practical implementation and is made in the form of a system of logical inference of cause-and-effect patterns. The weights of objects and attributes are also partially taken into account.

To receive the results the authors in [3] observed that the number of tweets containing specific health-related cause-effect relationships is small in comparison to the overall corpus. The number of sentences matched by the rules is 501 from 29705 tweets for stress (1.6 %), 72/3827 (1.8 %) for insomnia, and 94/11252 (0.8 %) for headache.

The final causality extracted from the matched sentences are 41, 98 and 42 for insomnia, stress and headache, respectively. To evaluate the precision of causal relation extraction, authors compared system outputs with human annotations. Table 4 shows the comparison results.

6 Conclusion

In conclusion, various theories and methods of causal relationships have already been created. However, we currently face the challenge of incorporating these methods and approaches into natural language processing algorithms.

This paper attempted to highlight the most pertinent and focused papers on causality in natural language processing. As shown by the outcomes of various ways, exploiting causality links can improve the accuracy of algorithms' work, although it remains challenging and problematic to manage this duty entirely.

[9] shows that the network architectures built on top of the contextualized language model can learn causal relations in the text using sentence context, event information, and masked event context. For a comprehensive causal comprehension of events stated in natural language text, we must be able to recognize sentences containing causal events, identify those events and their causal linkages, and comprehend the impacts between those events.

Consequently, there is a target to test the other the other language models as XLNet [18], GPT-2 [12], ELECTRA [1], MT5 [17] and to try the other different approaches to fully implement the causality in NLP methods and work on improving the accuracy of algorithms.

Acknowledgments

This research is funded by the Aerospace Committee of the Ministry of Digital Development, Innovations and Aerospace Industry of the Republic of Kazakhstan (BR11265420).

References

1. **Clark, K., Luong, M. T., Le, Q. V., Manning, C. D. (2020).** Electra: Pre-training text encoders as discriminators rather than generators. DOI: 10.48550/ARXIV.2003.10555.
2. **Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2019).** BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.
3. **Doan, S., Yang, E. W., Tilak, S. S., Li, P. W., Zisook, D. S., Torii, M. (2019).** Extracting health-related causality from twitter messages using natural language processing. BMC Medical Informatics and Decision Making, Vol. 19. DOI: 10.1186/s12911-019-0785-0.
4. **Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P., Yuret, D. (2007).** SemEval-2007 task 04: Classification of semantic relations between nominals. Proceedings of the Fourth International Workshop on Semantic Evaluations, Association for Computational Linguistics, pp. 13–18.
5. **Gopalan, S., Lalithadevi, S. (2018).** Cause and effect extraction from biomedical corpus. Computación y Sistemas, Vol. 21. DOI: 10.13053/cys-21-4-2854.
6. **Gurulingappa, H., Rajput, A. M., Roberts, A., Fluck, J., Hofmann-Apitius, M., Toldo, L. (2012).** Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. Journal of Biomedical Informatics, Vol. 45, No. 5, pp. 885–892. DOI: 10.1016/j.jbi.2012.04.008.
7. **Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., Pennacchiotti, M., Romano, L., Szpakowicz, S. (2010).** SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. Proceedings of the 5th International Workshop on Semantic Evaluation, Association for Computational Linguistics, pp. 33–38.
8. **Khetan, V., Ramnani, R., Anand, M., Sengupta, S., Fano, A. (2020).** Causal-BERT: Language models for causality detection between events expressed in text.
9. **Khetan, V., Ramnani, R., Anand, M., Sengupta, S., Fano, A. E. (2021).** Causal BERT: Language models for causality detection between events expressed in text. Lecture Notes in Networks and Systems, Springer International Publishing, pp. 965–980. DOI: 10.1007/978-3-030-80119-9$_64$.
10. **Marasovi, A. (2018).** NLPs generalization problem, and how researchers are tackling it. The Gradient.
11. **Maslov, P. (2008).** Extracting causal patterns from natural language texts. Tavrichesky Bulletin of Informatics and Mathematics, Vol. 13, No. 2.
12. **Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. (2019).** Language models are unsupervised multitask learners.
13. **Schölkopf, B. (2019).** Causality for machine learning. DOI: 10.48550/ARXIV.1911.10500.
14. **Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., Bengio, Y. (2021).** Toward causal representation learning. Proceedings of the IEEE, Vol. 109, No. 5, pp. 612–634. DOI: 10.1109/JPROC.2021.3058954.
15. **Solstad, T., Bott, O. (2017).** Causality and causal reasoning in natural language. pp. 619–644.
16. **Tamburini, F. (2017).** Semgrex-plus: a tool for automatic dependency-graph rewriting. Proceedings of the Fourth International Conference on Dependency Linguistics, Linköping University Electronic Press, pp. 248–254.
17. **Xue, L., Constant, N., Roberts, A., Kale, M., Al Rfou, R., Siddhant, A.,**

Barua, A., Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, pp. 483–498. DOI: 10.18653/v1/2021.naacl-main.41.

18. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. Advances in Neural Information Processing Systems, Curran Associates, Inc., Vol. 32.

Article received on 20/05/2022; accepted on 14/10/2022.
Corresponding author is Alexander Gelbukh.