*Neural*NERE: Neural Named Entity Relationship Extraction for End-to-End Climate Change Knowledge Graph Construction

Prakamya Mishra [1] **Rohan Mittal** [1]

Abstract

This paper proposes an end-to-end Neural Named Entity Relationship Extraction model (called *Neural*NERE) for climate change knowledge graph (KG) construction, directly from the raw text of relevant news articles. The proposed model will not only remove the need for any kind of human supervision for building knowledge bases for climate change KG construction (used in the case of supervised or dictionary-based KG construction methods), but will also prove to be highly valuable for analyzing climate change by summarising relationships between different factors responsible for climate change, extracting useful insights & reasoning on pivotal events, and helping industry leaders in making more informed future decisions. Additionally, we also introduce the Science Daily Climate Change dataset (called *Sci*DCC) that contains over $11k$ climate change news articles scraped from the Science Daily website, which could be used for extracting prior knowledge for constructing climate change KGs.

1. Introduction

News outlets have played a crucial role in increasing awareness about climate change through their news articles, because of which more and more people have started to understand the consequences of climate change. The volume of news articles published regarding climate change has been increasing rapidly with the growth in news coverage. This has made it challenging to extract valuable information regarding climate change from these news articles. Algorithms that can extract and organize climate change information by condensing the relevant knowledge directly from a large collection of noisy and redundant news articles could prove to be highly valuable in analyzing relationships between different factors responsible for climate change. This would help in generating useful insight and reasoning about the pivotal events which in turn will help industry leaders in making more informed policies relating to climate change in future. Such knowledge can be distilled from the structured representation of knowledge graphs (KGs) generated from these news articles.

There has been a growing interest in generating high-quality KGs for information extraction from raw text (Yu et al., 2020; Saxena et al., 2020; Wang et al., 2019; Nathani et al., 2019; Lin et al., 2019; Wities et al., 2017). Previously, KG construction approaches were either supervised (Bollacker et al., 2008) or semi-supervised (Carlson et al., 2010; Dong et al., 2014), both of which were expensive and time-consuming due to the involvement of human supervision. Recently, neural models have enabled KG construction without any human involvement (Bosselut et al., 2019; Balazevic et al., 2019; Xiong et al., 2018; Trivedi et al., 2017; García-Durán et al., 2018). The problem with the existing approaches is that they all use some prior knowledge in the form of Knowledge Bases (KB) to learn to predict relationships between the subject-object entity phrases for constructing KGs. This is the major reason why not much work has been done for constructing climate change KGs since there doesn't exist any well-established KBs for climate change.

To solve this problem, we propose *Neural*NERE, an end-to-end Neural Named Entity Relationship Extraction model for constructing climate change knowledge graphs directly from the raw text of relevant news articles. Additionally, we introduce a new climate change news dataset (called *Sci*DCC dataset[1]) containing over $11k$ news articles scraped from the Science Daily website, which can be used for extracting prior knowledge for constructing climate change knowledge graphs using *Neural*NERE.

2. *Sci*DCC Dataset

The Science Daily Climate Change dataset called *Sci*DCC was created by web scraping news articles from the "Earth & Climate" and "Plant & Animals" topics in the environmental science section of the Science Daily (SD) website. The SD

[*]Equal contribution [1]Independent Researcher, India. Correspondence to: Prakamya Mishra <pkms.research@gmail.com>.

Tackling Climate Change with Machine Learning Workshop at ICML 2021.

[1]https://sites.google.com/view/scidccdataset

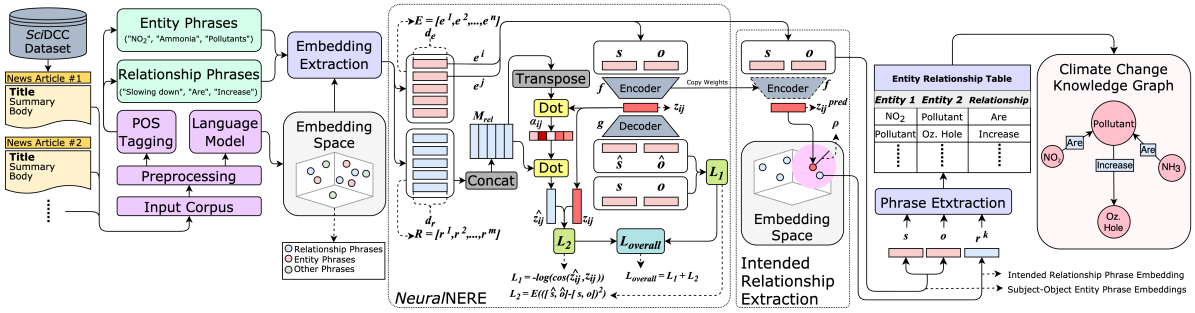

*Figure 1. Neural*NERE Model Architecture

Table 1. Key statistics of the *Sci*DCC dataset.

# NEWS ARTICLES	11,539	# NEWS CATEGORY	20
AVG. TITLE LEN.	9.32	MAX. TITLE LEN.	65
AVG. SUMMARY LEN.	47.28	MAX. SUMMARY LEN.	488
AVG. BODY LEN.	523.18	MAX. BODY LEN.	1968

Table 2. News Category Statistics

NO.	CATEGORY	# NEWS ARTICLES
1	EARTHQUAKES	986
2	POLLUTION	945
3	GENETICALLY MODIFIED	914
4	HURRICANES CYCLONES	844
5	AGRICULTURE & FOOD	844
6	ANIMALS	758
7	WEATHER	719
8	ENDANGERED ANIMALS	701
9	CLIMATE	700
10	OZONE HOLES	623
11	BIOLOGY	620
12	NEW SPECIES	527
13	ENVIRONMENT	478
14	BIOTECHNOLOGY	460
15	GEOGRAPHY	407
16	MICROBES	398
17	EXTINCTION	356
18	ZOOLOGY	210
19	GEOLOGY	28
20	GLOBAL WARMING	21

news articles are relatively more scientific when compared to other news outlets, which makes SD perfect for extracting scientific climate change news. In total, we extracted over 11k news articles from 20 categories relevant to climate change, where each article consists of a title, summary, and a body. For each category, we were able to extract a maximum of 1k news articles. The key statistics of the *Sci*DCC dataset are summarized in Table 1 and more detailed statistics can be found below. Here, we provide more detailed statistics about the *Sci*DCC dataset. Table 2 summarises the no. of news articles extracted per category. All the histogram plots (right) in Fig. 2, shows the length distribution of title, summary, and body of news articles, whereas all the density plots (left) in figure shows the cumulative distribution of these lengths over the years.

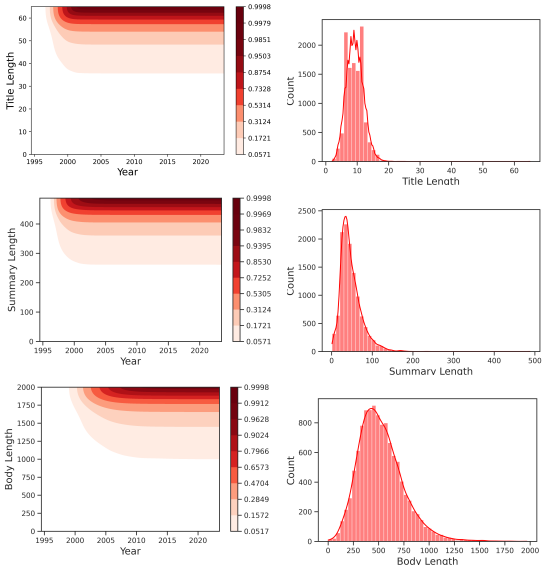

Figure 2. Length Distribution statistics of the *Sci*DND dataset.

3. *Neural*NERE

In the this section, we describe each component of the proposed *Neural*NERE model (see Fig. 1). The primary objective of the *Neural*NERE model is to learn the embedding representation of the intended relationship phrase that describes the relationship between any two named entities present in the previously introduced *Sci*DCC dataset. These learned embedding representations of the intended relationship phrase for every named entity pairs *[Entity 1 (subject), Entity 2 (Object)]* present in the *Sci*DCC dataset will later be used to generate a climate change knowledge graph using the *[Entity 1 (subject), Relationship, Entity 2 (Object)]* triplets.

We first create an input corpus by extracting the raw text from the summary part and body part of all the articles present in the *Sci*DCC dataset. This input corpus is first preprocessed (tokenization, lower-casing, stemming, lemmatization) and then used for: (1) fine-tuning a language model

(a character-based language like FastText[2] or GloVe[3]) for learning the word embedding representations corresponding to every word present in the corpus; (2) for extracting all the named entity phrases as well as all the possible relationship phrases. The named entity phrases and the possible relationship phrases are extracted by using Part-of-Speech Tagging (POS Tagging). All the word tokens are marked with their corresponding POS tags which are utilized to create (1) an entity phrase list by extracting all the named entity phrases using noun-phrase chunking, and (2) a relationship phrase list by extracting all the possible relationship phrases using verb-phrase chunking. The fine-tuned language model is then used to convert the extracted named entity phrases from the entity phrase list, and relationship phrases from the relationship phrase list into their corresponding embedding representations. We propose to use a character-based language model to avoid problems while generating embedding representations for multi-word entity/relationship phrases. All the extracted named entity phrases are represented by E, and all the extracted possible relationship phrases are represented by R.

$$E = [e^1, ..., e^n]; R = [r^1, ..., r^m] \quad d_e, d_r \in \mathbb{N} \tag{1}$$

Here in Equation (1), $e^i \in \mathbb{R}^{d_e}$ represents the d_e-dimensional embedding representation of the i^{th} named entity phrase in E; $r^i \in \mathbb{R}^{d_r}$ represents the d_r-dimensional embedding representation of i^{th} relationship phrase in R; n & m are the number of phrases in E and R respectively. Next *Neural*NERE uses E & R as input, and tries to learn the intended relationship representations between all the possible entity pairs that can be generated from the entity phrase list. These learned intended relationship phrase representation will later be used to construct the climate change KG. For training, *Neural*NERE uses (1) A pair of entity phrase representations represented by (s, o), where the $s \in \mathbb{R}^{d_e}$ is the entity phrase representation of the i^{th} entity phrase (in E) which acts as the subject in the subject-object relationship, and $o \in \mathbb{R}^{d_e}$ is the entity phrase representation of the j^{th} entity phrase (in E) which acts as the object in the subject-object relationship; (2) A relationship phrase matrix $M_{rel} \in \mathbb{R}^{d_r \times m}$, which is basically a matrix constructed by concatenating ($\otimes$) all the m relationship phrase representations together from the relationship phrase list, as shown in Equation (2).

$$M_{rel} = r^1 \otimes r^2 \otimes ... \otimes r^m; M_{rel} \in \mathbb{R}^{d_r \times m}, r^i \in \mathbb{R}^{d_r} \tag{2}$$

Next, *Neural*NERE uses an encoder-decoder network for encoding the relationship between the subject entity phrase represented by s and object entity phrase represented by o into an encoded representation $z_{ij} \in \mathbb{R}^{d_r}$, having the same embedding representation size as that of the relationship phrases. The encoder network f, first encodes the input subject-object entity phrase pairs (s, o) into an encoded vector z_{ij}, and the decoder network g then decodes the encoded vector z_{ij} into reconstructions represented by ($\hat{s}$, $\hat{o}$), as shown in Equation (3).

$$\hat{s}, \hat{o} = g(z_{ij}); \quad z_{ij} = f(s, o) \quad \hat{s}, \hat{o} \in \mathbb{R}^{d_e} \tag{3}$$

This encoded vector z_{ij} represents the embedding representation of the intended relationship phrase between subject-object entity phrase pairs that *Neural*NERE is trying to learn. Although we want the encoded vector z_{ij} to capture the relationship between phrases represented by s and o, but in reality, we don't really know much about the nature of information being captured in the encoded vector. In order to force z_{ij} to capture such relationship-based information, *Neural*NERE uses the relationship phrase matrix M_{rel} which contains embedding representations of all the existing relationship phrase. To do this *Neural*NERE first generates attention scores (α_{ij}) by taking a matrix-vector product ($\bullet$) between the transpose of normalized M_{rel} matrix and z_{ij}, as shown in Equation (4). The normalization of M_{rel} matrix is done by pre-multiplying it with a diagonal matrix D_{rel}=$diag(\frac{1}{|r^1|}, ..., \frac{1}{|r^m|})$, where the values at the diagonal are the reciprocal of the absolute values of r^ith column in M_{rel} matrix. These attention scores are then used for taking an attention-based weighted sum of all the relationship phrase embedding representations for generating a new encoded representation represented by $\hat{z_{ij}}$ that also captures the relationship-based information, as shown in Equation (5). These attention scores enables *Neural*NERE to give more attention to r^i's that better represents z_{ij}.

$$\alpha_{ij} = [\alpha^1, ..., \alpha^m]^T = D_{rel} \bullet M_{rel}^T \bullet z_{ij} \tag{4}$$

$$\hat{z_{ij}} = M_{rel} \bullet \alpha_{ij} \tag{5}$$

Now we will use the above generated encoded vector $\hat{z_{ij}}$ to enforce the encoded vector z_{ij} to capture relationship-based information. We will do so by modifying the overall loss function. The loss function of the proposed model will consist of two terms, (1) The first term will be the reconstruction loss represented by L_1 in Equation (6), which will ensure the reconstruction of input in the encoder-decoder network; (2) Second term will be the cosine similarity loss ($-\log cos()$) between the two encoded vectors $\hat{z_{ij}}$ & z_{ij} represented by L_2 in Equation (6), which will ensure the learned encoded representation to capture the relationship-based information from the existing relationship phrase representations. The overall loss function ($L_{overall}$) of the *Neural*NERE model will be the addition of the above mentioned individual losses, as shown in Equation (7).

$$L_1 = E((\hat{[s, o]} - [s, o])^2); L_2 = -\log cos(\hat{z_{ij}}, z_{ij}) \tag{6}$$

[2]https://fasttext.cc/

[3]https://nlp.stanford.edu/projects/glove/

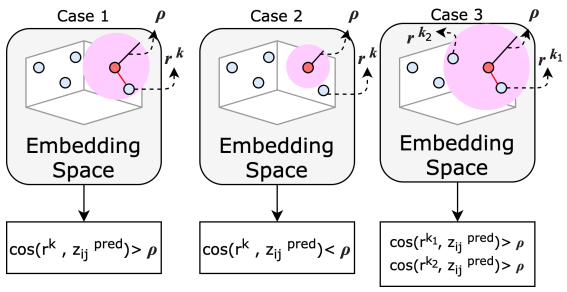

Figure 3. Effect of ρ on Knowledge Graph Construction

$$L_{overall} = L_1 + L_2 \tag{7}$$

Next after training the *Neural*NERE model with all the possible combinations of subject-object entity phrase pairs that can be constructed from the entity phrase list using our custom loss function $L_{overall}$, the encoder network f learns to generate the embedding representation of the intended representation phrase between the subject-object entity phrases. Now we use the trained encoder network f to predict the embedding representation z_{ij}^{pre} of the intended relationship phrase for all the subject-object entity phrase pairs. To extract the actual phrase corresponding to z_{ij}^{pre}, we first compute the cosine similarities between z_{ij}^{pre} and all $r^i \in R$. Then the relationship phrase corresponding to the r^k, which has the highest cosine similarities with z_{ij}^{pre} is chosen as the intended relationship phrase. We don't choose any if all the computed cosine similarity values fall below a threshold ρ. Such a threshold keeps the model in check and prohibits the generation of useless relationships (as shown in Fig. 3). Finally, these triplets comprising of a subject entity phrase, an object entity phrase, and the predicted relationship phrase (from *Neural*NERE) are used to construct the climate change knowledge graph.

In the proposed model, we introduce a threshold ρ to keep the model in check during the relationship generation phase. For intuition, ρ limits the proximity of search for the intended relationship phrase from the predicted representation of the intended relationship phrase. Decreasing the value of threshold parameter ρ enables the exploration of relationship phrases that are relatively distant from the predicted representation of the intended relationship phrase, whereas increasing the value of threshold parameter ρ prohibits the exploration relationship phrases that are relatively distant. This is illustrated in the Figure above. As shown in case 1, for some value of ρ if there only exist a single $r^k \in R$ such that the cosine similarity of r^k and z_{ij}^{pred} is more than the threshold value ρ, then the relationship phrase corresponding to r^k will be chosen as the intended relationship phrase. Whereas as shown in case 2, for some value of ρ if there does not exist any $r^k \in R$ such that the cosine similarity of r^k and z_{ij}^{pred} is more than the threshold value ρ, then no relationship phrase will be extracted between the subject and object entities. Case 2 demonstrates the example in which the proposed model prohibits the generation of useless relationships. Finally as shown in case 3, for some value of ρ if there exist $r^{k_1}, r^{k_2} \in R$ (in other words more than one relationship phrases) such that the cosine similarities of both r^{k_1} & r^{k_2} with z_{ij}^{pred} are more than the threshold value ρ, then the relationship phrase corresponding to r^{k_i} with the highest cosine similarity with z_{ij}^{pred} will be chosen as the intended relationship phrase. Case 3 demonstrates the example wherein the proposed model chooses the intended relationship phrases whose representation is in the closest proximity to the predicted representation of the intended relationship phrase.

4. Projected Results

Using the proposed *Sci*DCC dataset and *Neural*NERE model we aim to give industry leaders, analyst, and policymakers a tool for:

- Extracting and organizing climate change information from a large collection of news articles.
- Analyzing relationships between different factors responsible for climate change.
- Gathering insight/reasoning about the pivotal events for more informed climate change policy making.

In conclusion, we proposed *Neural*NERE, an end-to-end Neural Named Entity Relationship Extraction model for constructing climate change knowledge graphs directly from the raw text of relevant news articles. We also introduced a new climate change news dataset (called *Sci*DCC dataset) for extracting prior knowledge for constructing climate change knowledge graphs using *Neural*NERE.

References

Balazevic, I., Allen, C., and Hospedales, T. TuckER: Tensor factorization for knowledge graph completion. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5185–5194, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1522. URL https://www.aclweb.org/anthology/D19-1522.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pp.

1247–1250, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581026. doi: 10.1145/1376616.1376746. URL https://doi.org/10.1145/1376616.1376746.

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., and Choi, Y. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4762–4779, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1470. URL https://www.aclweb.org/anthology/P19-1470.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E., and Mitchell, T. Toward an architecture for never-ending language learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 24(1), Jul. 2010. URL https://ojs.aaai.org/index.php/AAAI/article/view/7519.

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., and Zhang, W. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pp. 601–610, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450329569. doi: 10.1145/2623330.2623623. URL https://doi.org/10.1145/2623330.2623623.

García-Durán, A., Dumančić, S., and Niepert, M. Learning sequence encoders for temporal knowledge graph completion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4816–4821, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1516. URL https://www.aclweb.org/anthology/D18-1516.

Lin, B. Y., Chen, X., Chen, J., and Ren, X. KagNet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2829–2839, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1282. URL https://www.aclweb.org/anthology/D19-1282.

Nathani, D., Chauhan, J., Sharma, C., and Kaul, M. Learning attention-based embeddings for relation prediction in knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4710–4723, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1466. URL https://www.aclweb.org/anthology/P19-1466.

Saxena, A., Tripathi, A., and Talukdar, P. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4498–4507, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.412. URL https://www.aclweb.org/anthology/2020.acl-main.412.

Trivedi, R., Dai, H., Wang, Y., and Song, L. Know-evolve: Deep temporal reasoning for dynamic knowledge graphs. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 3462–3471. JMLR.org, 2017.

Wang, Q., Huang, L., Jiang, Z., Knight, K., Ji, H., Bansal, M., and Luan, Y. PaperRobot: Incremental draft generation of scientific ideas. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1980–1991, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1191. URL https://www.aclweb.org/anthology/P19-1191.

Wities, R., Shwartz, V., Stanovsky, G., Adler, M., Shapira, O., Upadhyay, S., Roth, D., Martinez Camara, E., Gurevych, I., and Dagan, I. A consolidated open knowledge representation for multiple texts. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pp. 12–24, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-0902. URL https://www.aclweb.org/anthology/W17-0902.

Xiong, W., Yu, M., Chang, S., Guo, X., and Wang, W. Y. One-shot relational learning for knowledge graphs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1980–1990, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1223. URL https://www.aclweb.org/anthology/D18-1223.

Yu, H., Li, H., Mao, D., and Cai, Q. A relationship extraction method for domain knowledge graph construction. *World Wide Web*, 23(2):735–753, 2020. doi: 10.1007/s11280-019-00765-y. URL https://doi.org/10.1007/s11280-019-00765-y.