ICML

International Conference
On Machine Learning

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

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

[1]Independent Researcher

Outline

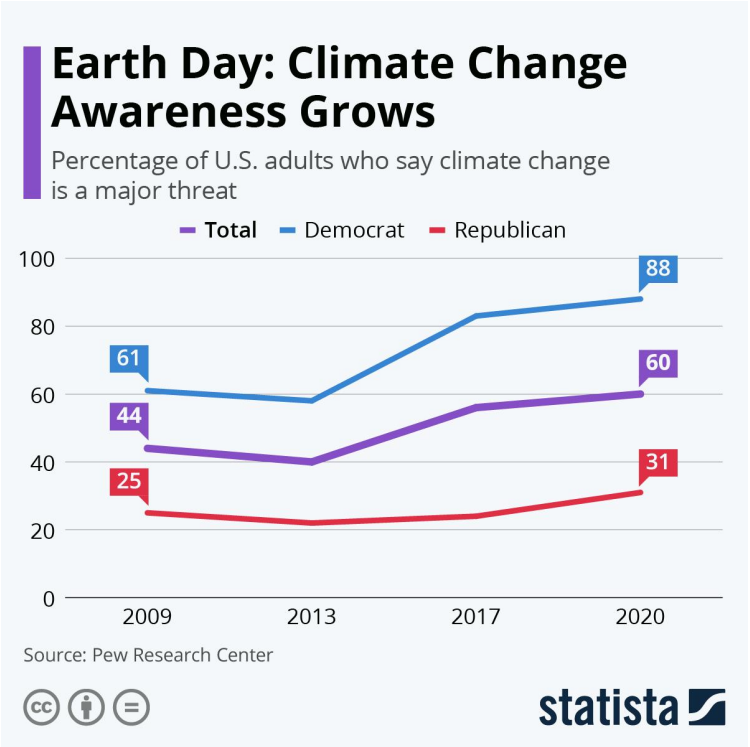
1. Motivation
2. *Sci*DCC Dataset
3. *Neural*NERE
4. Conclusion & Future Work

Motivation

Why?

- News outlets have increased awareness about climate change through their news articles.
- Volume of news articles have been increasing rapidly.
- Difficult to extract useful information.

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.



Source: https://www.statista.com/chart/21415/climate-change-awareness-earth-day/

Motivation

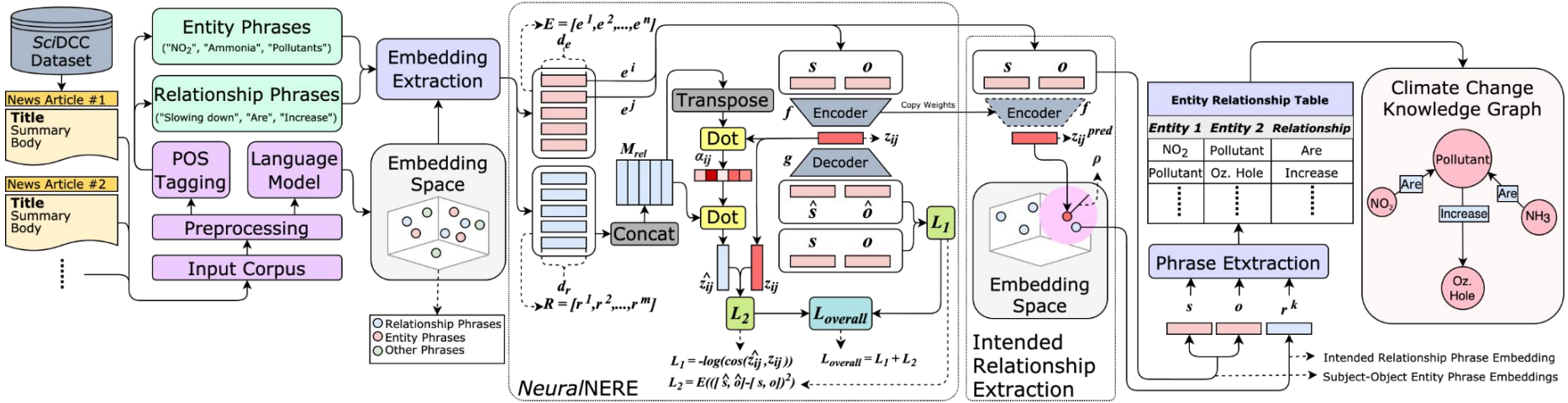
Knowledge Graph Construction

- Supervised or Semi-Supervised [Bollacker et al., 2008; Carlson et al., 2010; Dong et al., 2014]
- Neural models [Bosselut et al., 2019; Balazevic et al., 2019; Xiong et al., 2018; Trivedi et al., 2017; Garc´ıa-Duran 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.

Our Solution

- **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 SciDCC dataset) containing over 11k news articles scraped from the Science Daily website.**



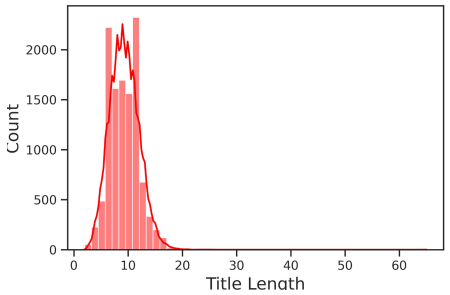
*Sci*DCC Dataset

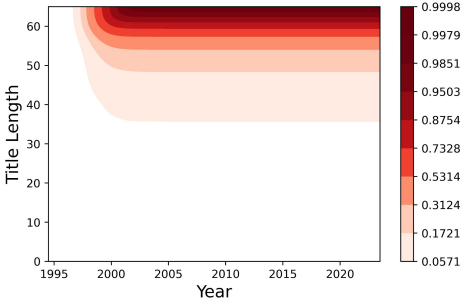
Table 1. Key statistics of the *Sci*DCC dataset.

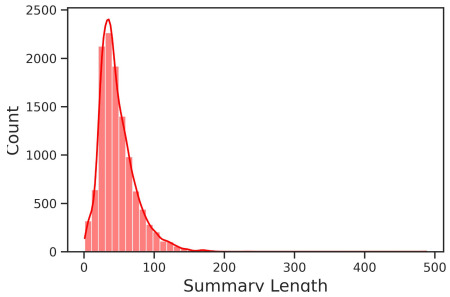
# News articles	11,539	# News category	20
Avg. title len.	9.32	Max. title len.	65
Avg. summary len.	47.28	Max. summary len.	488
Avg. body len.	523.18	Max. body len.	1968

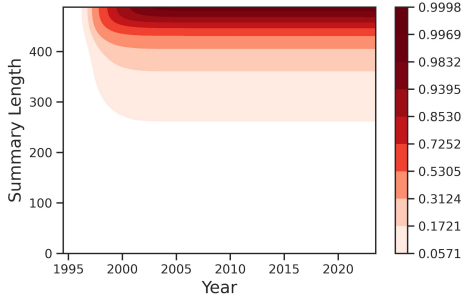
Table 2. News Category Statistics

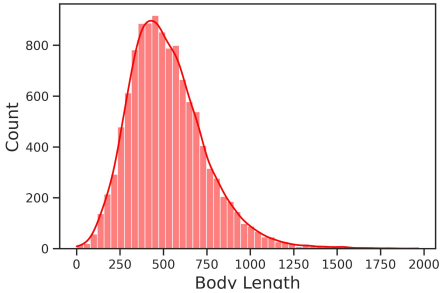
No.	Category	# News Articles
1	Earthquakes	986
2	Pollution	945
3	Genetically Modified	914
4	Hurricanes Cyclones	844
5	Agriculture & Food	844
6	Animals	758
7	Weather	719
8	Endangered Animals	701
9	Climate	700
10	Ozone Holes	623
11	Biology	620
12	New Species	527
13	Environment	478
14	Biotechnology	460
15	Geography	407
16	Microbes	398
17	Extinction	356
18	Zoology	210
19	Geology	28
20	Global Warming	21

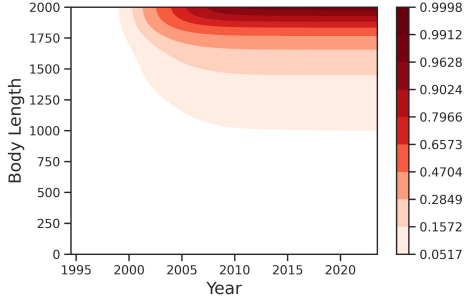




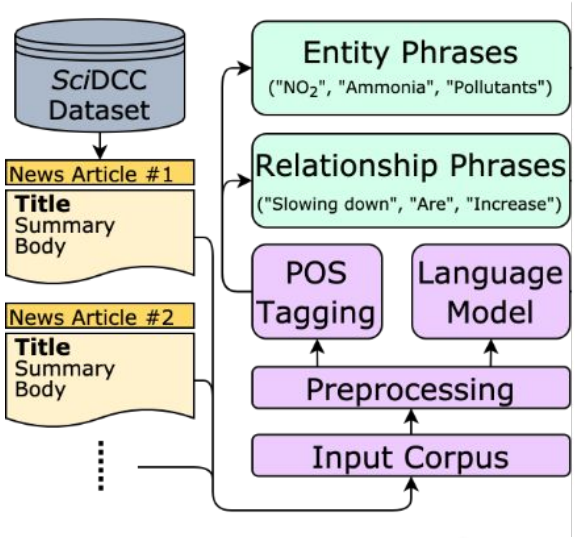






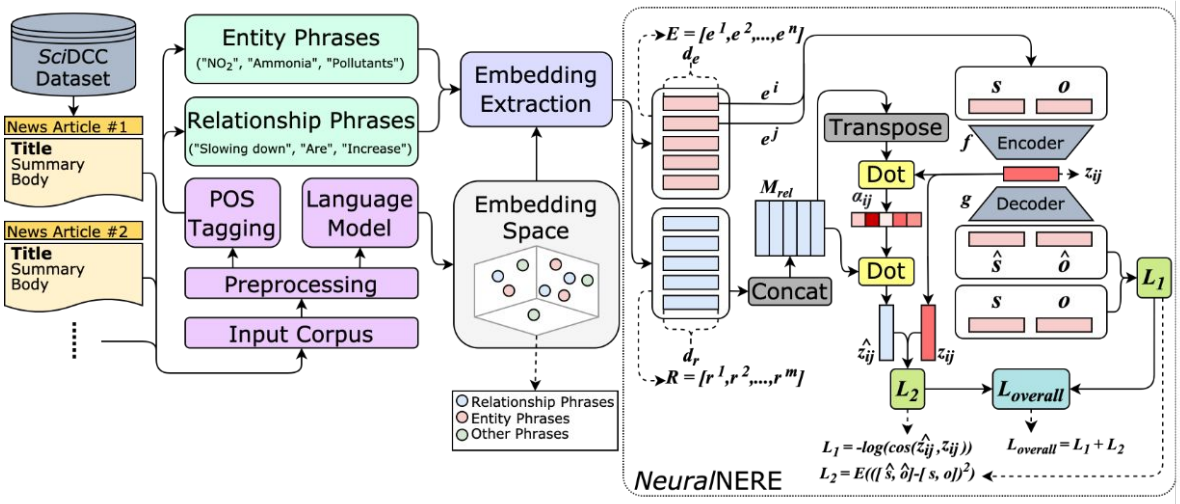


*Neural*NERE



- 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 SciDCC dataset.
- This input corpus is first preprocessed (tokenization, lower-casing, stemming, lemmatization) and then used for: (1) Fine-tuning a language model for learning the word embedding representations corresponding to every word present in the corpus; (2) Extracting all the named entity phrases as well as all the possible relationship phrases using noun & verb phrase chunking.

*Neural*NERE: Inputs



- The fine-tuned language model is 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. [Extracted named entity phrases are represented by E, and all the extracted possible relationship phrases are represented by R]
- NeuralNERE 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.

*Neural*NERE: Intended Relationship Representation

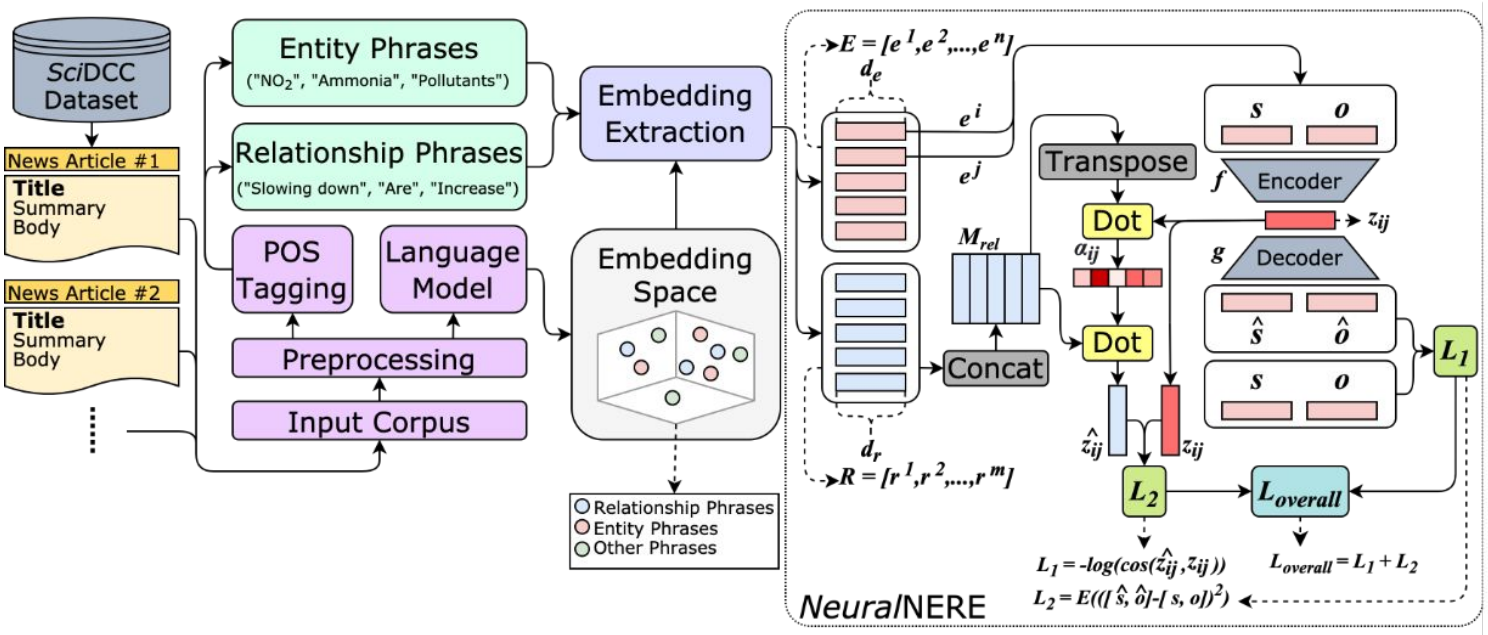
For training, NeuralNERE uses:

- A pair of entity phrase representations represented by (s, o), where the s $\in$ R de 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 $\in$ R de is the entity phrase representation of the j^{th} entity phrase (in E) which acts as the object in the subject-object relationship.
- A relationship phrase matrix $M_{rel} \in R^{dr \times m,}$ which is basically a matrix constructed by concatenating ($\otimes$) all the m relationship phrase representations together from the relationship phrase list.

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

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

*Neural*NERE: Intended Relationship Representation



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

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

*Neural*NERE: Intended Relationship Representation

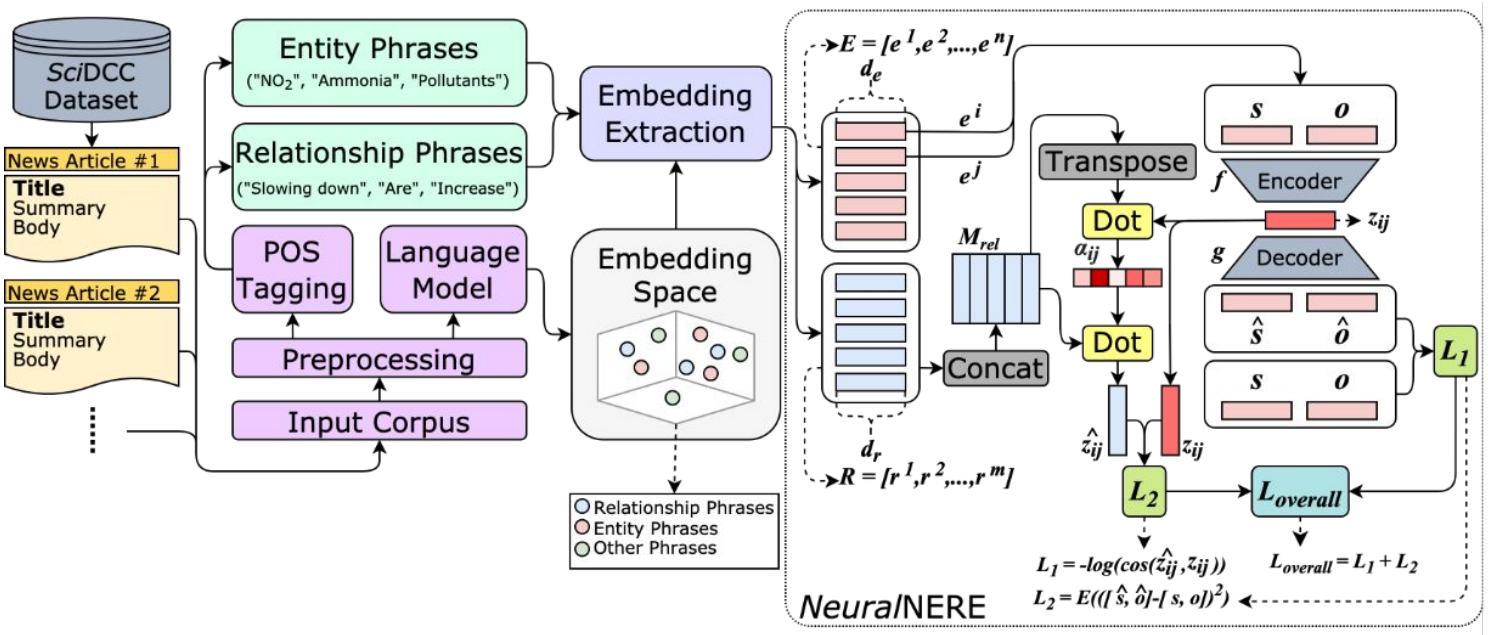
- Next, NeuralNERE uses an encoder-decoder network for encoding the relationship between the subject entity phrase and object entity phrase into an encoded representation $z_{ij} \in R^{dr}$, having the same embedding representation size as that of the relationship phrases.
- This encoded vector z_{ij} represents the embedding representation of the intended relationship phrase between subject-object entity phrase pairs that NeuralNERE 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, NeuralNERE uses the relationship phrase matrix Mrel which contains embedding representations of all the existing relationship phrase.

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

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

$$\hat{z_{ij}} = M_{rel} \bullet \alpha_{ij} \qquad (5)$$

*Neural*NERE: Intended Relationship Representation



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

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

$$\hat{z_{ij}} = M_{rel} \bullet \alpha_{ij} \qquad (5)$$

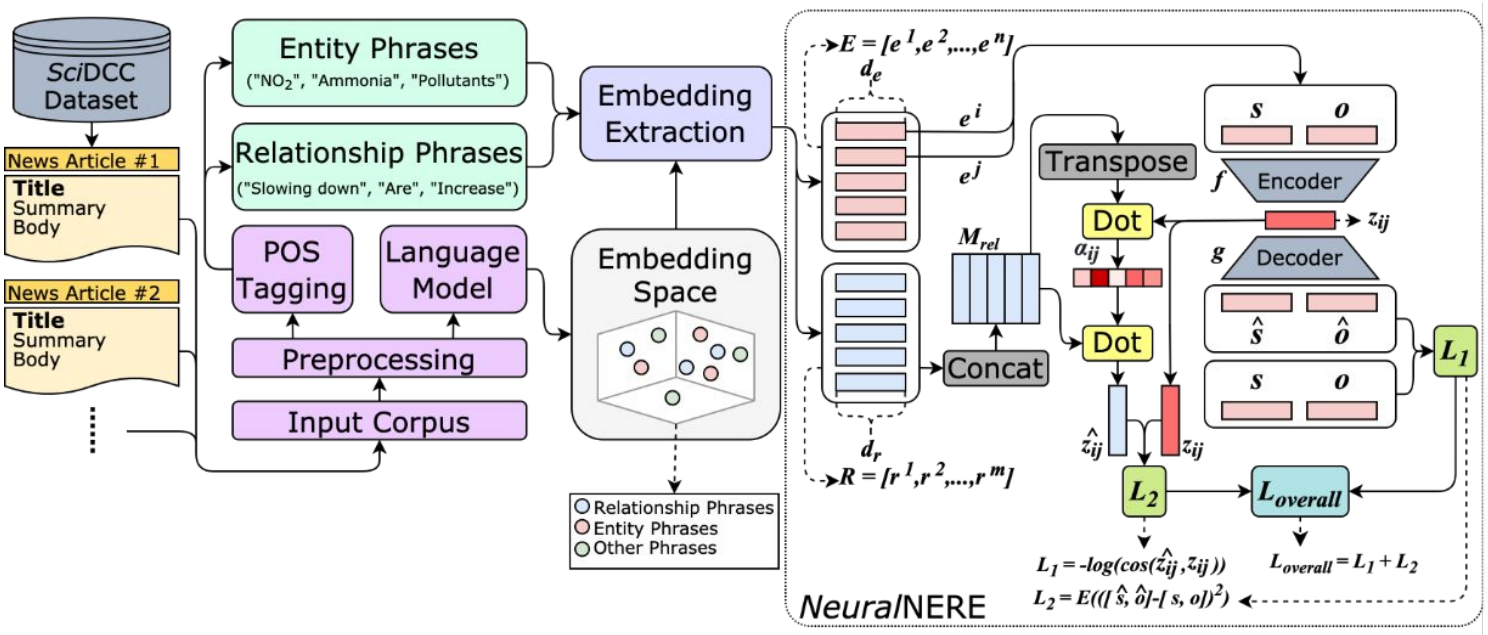
*Neural*NERE: Loss Function

- 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:
  - 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.
  - 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 (Loverall) of the NeuralNERE model will be the addition of the above mentioned individual losses, as shown in Equation (7).

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

$$L_{overall} = L_1 + L_2 \quad (7)$$

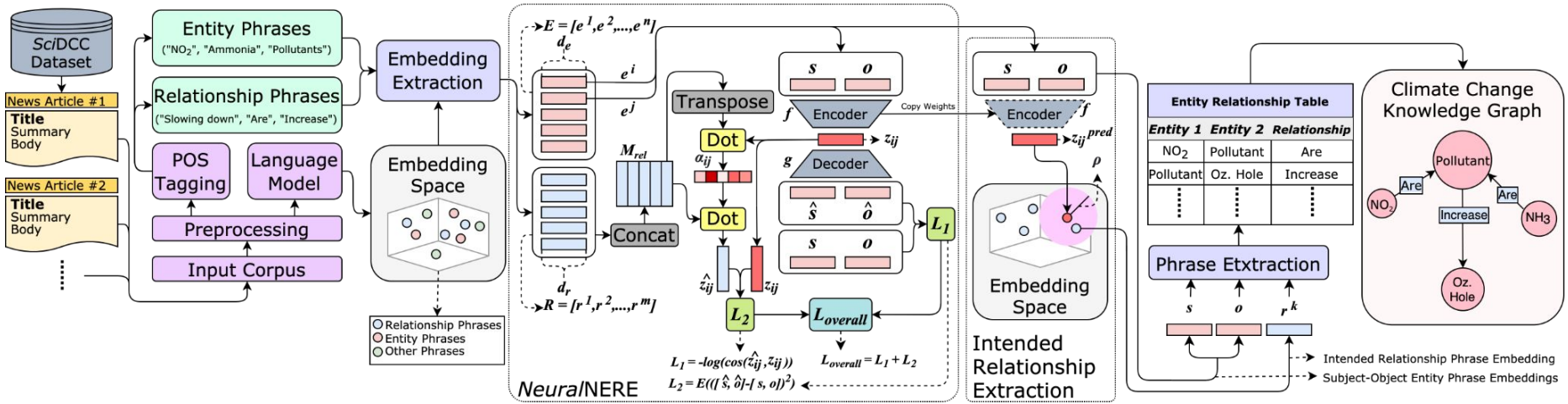
*Neural*NERE: Loss Function



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

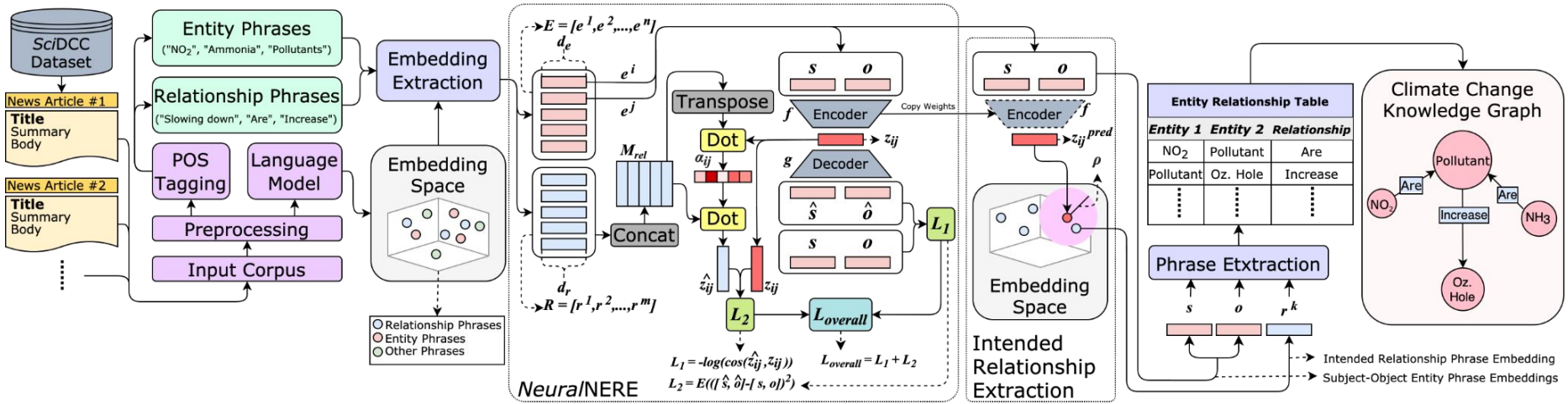
$$L_{overall} = L_1 + L_2 \quad (7)$$

*Neural*NERE



- Now we use the trained encoder network f to predict the embedding representation z pre ij of the intended relationship phrase for all the subject-object entity phrase pairs.
- Then the relationship phrase corresponding to the r^k , which has the highest cosine similarities with z^{pre}_{ij} 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

*Neural*NERE



- Finally, these triplets comprising of a subject entity phrase, an object entity phrase, and the predicted relationship phrase (from NeuralNERE) are used to construct the climate change knowledge graph.

Conclusion & Future Work

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

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

Thank You