Natural Language Engineering (2021), **27**, pp. 427–454
doi:10.1017/S1351324920000194

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Is your document novel? Let attention guide you. An attention-based model for document-level novelty detection

Tirthankar Ghosal[1,*], Vignesh Edithal[1], Asif Ekbal[1], Pushpak Bhattacharyya[1], Srinivasa Satya Sameer Kumar Chivukula[2] and George Tsatsaronis[2]

[1]Indian Institute of Technology Patna, Bihta, Bihar, Patna 801103, India and [2]Elsevier, Amsterdam, Netherlands
*Corresponding author. E-mail: tirthankar.slg@gmail.com

(Received 16 November 2018; revised 10 November 2019; accepted 16 January 2020; first published online 24 April 2020)

Abstract

Detecting, whether a document contains sufficient new information to be deemed as *novel*, is of immense significance in this age of data duplication. Existing techniques for document-level novelty detection mostly perform at the lexical level and are unable to address the semantic-level redundancy. These techniques usually rely on handcrafted features extracted from the documents in a rule-based or traditional feature-based machine learning setup. Here, we present an effective approach based on neural attention mechanism to detect document-level novelty without any manual feature engineering. We contend that the simple alignment of texts between the source and target document(s) could identify the state of *novelty* of a target document. Our deep neural architecture elicits inference knowledge from a large-scale natural language inference dataset, which proves crucial to the novelty detection task. Our approach is effective and outperforms the standard baselines and recent work on document-level novelty detection by a margin of ∼3% in terms of accuracy.

Keywords: Document-Level Novelty Detection; Decomposable Attention; Natural Language Inference; Document Classification

1. Introduction

Novelty detection implies finding elements that have not appeared before or are new/original with respect to relevant prior references. Document-level novelty detection implies characterizing a document as novel or non-novel based on the amount of new information contained in the document. Research in this field could be attributed mostly to signal processing domain in detecting novel patterns from time series data (Dasgupta and Forrest 1996), mammograms (Tarassenko *et al*. 1995), fault detection (King *et al*. 2002), radar target detection (Carpenter, Rubin, and Streilein 1997), handwritten digit recognition (Tax and Duin 1998), Internet and e-commerce (Manikopoulos and Papavassiliou 2002), statistical process control (Guh *et al*. 1999), and several others. However there had been attempts to pursue the problem in detecting new information from texts as well (Wayne 1997; Zhang, Callan, and Minka 2002; Soboroff and Harman 2003). Textual novelty detection has not been attempted rigorously except a few explorations such as the one introduced at TAC.[a] Most of the existing works on novelty detection in text till date have focused primarily at the sentence level, that is, to extract the sentences that carry new information with respect to a reference set. The current work aims to detect the novelty of an entire document,

[a]https://tac.nist.gov/2011/RTE/.

that is, to find the relevant documents that carry new information with respect to whatever the intended reader is already aware of. In our work, we view the problem of novelty detection as a binary classification problem with the judgment that whether an incoming document bears sufficiently new information to be labeled as novel with respect to a set of source documents. We view the source document set as the memory of the reader, which stores the known information about the topic/event. In all of our subsequent discussions, we would refer to source documents as the manifestation of information already known about a topic or the knowledge base of the evaluator. Novelty detection is a well-studied problem in information retrieval literature and has widespread natural language processing (NLP) applications such as: text summarization (Allan, Gupta, and Khandelwal 2001; Bysani 2010), event detection from news or tracking development of news items (Karkali *et al.* 2013), predicting impact of scholarly articles (Mishra and Torvik 2016), etc. However, we find that most of the investigations (Gamon 2006; Zhang and Tsai 2009; Lee 2015) and exercises/shared tasks (Soboroff 2004; Bentivogli *et al.* 2011) till date are directed toward sentence-level novelty mining. But considering the present context and exponential growth of redundant documents across the web, we deem document-level novelty detection as a very well-timed problem. Redundancy today is not just limited to the lexical surface form of text but takes into account the semantic and pragmatic aspects as well (e.g., paraphrasing, natural language inference (NLI)/entailment, etc.). Here in this work, we investigate a deep learning architecture to detect novelty of a target document with respect to a set of documents already seen by the system (which we refer to as the source document(s)).

Existing methods often employ representations from a deep neural network to address the semantic characteristics of texts. Detecting whether a document (*the target*) is novel or redundant is a complex problem and could not be approached in a straightforward manner (only from a single text). The target document should always be judged with respect to a set *source document(s)* or *previously known information*. Hence, deducing a joint vector representation of the source and the target document(s) which could effectively capture the semantic interactions necessary for deciding the novelty of the target document would be computationally very expensive, considering that the number of the source document(s) could exponentially rise.

Here, we arrive at a potentially feasible solution inspired by the phrase/sentence alignment problem in machine translation. We investigate which portion of texts in the source document(s) makes the target document appear *non-novel* to the reader and align the corresponding source–target text pairs to learn their interactions by a neural network. We hypothesize that for a *novel* document, there would be very little or no alignment with any portion of the source text. Eventually, we consider the sentence as the unit of information conveyance and look for their corresponding alignments in the respective texts (source and target). Let us consider the following ***example*** (Ghosal *et al.* 2018b):

- ***d1***: *Singapore is an island city-state with a population of around 5.61 millions. Singapore's territory consists of one main island along with 62 other islets.*
- ***d2***: *The Republic of Singapore is a sovereign country in Southeast Asia. The island city-state lies 137 kilometers north of the equator and has a dense population of approximately 5.6 million.*
- ***d3***: *Singapore is a global commerce, finance, and transport hub. Singapore has a tropical rainforest climate with no distinctive seasons, uniform temperature and pressure, high humidity, and abundant rainfall.*
- ***d4***: *Singapore, an island city-state off southern Malaysia, lies one degree north of the equator. As of June 2017, the island's population stood at 5.61 million.*

It is fairly easy to conclude that document *d4* follows from *d1* and *d2*, by simply aligning the two sentences in *d4* with the first sentence of *d1* and the second sentence of *d2*. However, considering only *d3* as the source, no such alignment is possible with *d4*. Thus, *d4* would be non-novel w.r.t *d1* and *d2* combined, but would appear novel if we consider *d3* as the only source. Hence, we could say that once the reader goes through *d1* and *d2*, they could infer the contents of *d4*. The inference is not exactly one-to-one as is the usual case in entailment literature. The example above describes a more likely multiple premise entailment (Lai, Bisk, and Hockenmaier 2017) scenario. Here, the premise of the first target sentence in *d4* is both the first sentence of *d1* and the second sentence of *d2*. However, if the reader goes through only *d3*, then document *d4* would seem to contain new information to them. Quite interestingly, here we could see a relation brewing between text alignment and sentence inference while judging the novelty of a piece of text. To effectively model this relation, we need efficient sentence representations. Hence, we train our sentence representations on a large-scale NLI dataset (the Stanford Natural Language Inference (SNLI) corpus) to capture the essence of the inference knowledge in our sentence embeddings. As discussed earlier, we manifest the alignment perspective via the neural attention mechanism as substantiated in our further discussions.

We set out to investigate this idea and see whether a neural network could learn text alignment and correctly identify a document as novel or redundant. We achieve this alignment via the attention mechanism (Bahdanau, Cho, and Bengio 2014), popular in deep neural networks. Here, novel document *d3* is relevant, yet diverse w.r.t others. Word/phrase-level alignment is handled with reasonable accuracy via attention in entailment literature (Parikh *et al.* 2016). We leverage their idea for sentence-level entailment to work in our case for ascertaining novelty of documents; a kind of transfer learning (textual entailment (TE) to novelty detection) approach to the problem concerned. One key challenge in this work is to generate a sentence representation that could effectively model this alignment perspective. We do so via the *inner-attention*-based sentence encoder (Liu *et al.* 2016) trained on the very large and semantically rich SNLI corpus (Bowman *et al.* 2015). Our proposed approach demonstrates significant performance improvement over the *state-of-the-art* systems and the reported baselines.

2. Related work

We trace the first significant concern on novelty detection back to the new event/first story detection task of the topic detection and tracking (TDT) campaigns (Wayne 1997). Techniques mostly involved grouping the news stories into clusters and then measuring the belongingness of an incoming story to any of the clusters based on some preset similarity threshold. If a story does not belong to any of the existing clusters, it is treated as the first story of a new event, and a new cluster is started. Vector space model, language model, lexical chain, etc. were used to represent each incoming news story/document. Some notable contributions from TDT are by Allan, Papka, and Lavrenko (1998), Yang *et al.* (2002), Stokes and Carthy (2001), Franz *et al.* (2001), Allan, Lavrenko, and Jin (2000), Yang, Pierce, and Carbonell (1998), Brants, Chen, and Farahat (2003). A close approximation of event-level document clustering via cross document event tracking can be found in Bagga and Baldwin (1999).

Research on sentence-level novelty detection gained prominence in the novelty tracks of Text Retrieval Conferences (TREC) from 2002 to 2004 (Harman 2002; Soboroff and Harman 2003; Soboroff 2004; Soboroff and Harman 2005). The goal of these tracks was to highlight relevant sentences that contain new information, given a topic and an ordered list of relevant documents. A significant amount of work came out on sentence-level novelty detection from Allan, Wade, and Bolivar (2003), Kwee, Tsai, and Tang (2009), Li and Croft (2005) based on TREC data. Language model measures, vector space models with cosine similarity, and word count measures were the dominant paradigms. Some other notable works on finding effective features to represent natural language sentences for novelty computation were based on the sets of terms (Zhang *et al.* 2003), term translations (Collins-Thompson *et al.* 2002), named entities or NE patterns (Gabrilovich, Dumais, and Horvitz 2004; Zhang and Tsai 2009), principal component analysis vectors (Ru *et al.* 2004), contexts (Schiffman and McKeown 2005), and graphs (Gamon 2006). Tsai, Tang, and Chan

https://doi.org/10.1017/S1351324920000194 Published online by Cambridge University Press

(2010), Tang, Tsai, and Chen (2010) presented an evaluation of metrics for sentence-level novelty mining.

Next came the novelty subtracks of Recognizing Textual Entailment-Text Analytics Conference (RTE-TAC) 6 and 7 (Bentivogli *et al.* 2011) where TE (Dagan *et al.* 2013) was viewed as one close neighbor to sentence-level novelty detection.

At the document level, pioneering work was conducted by Yang *et al.* (2002) via topical classification of online document streams and then detecting novelty of documents in each topic exploiting the NEs. Another work by Zhang *et al.* (2002) viewed novelty as an opposite characteristic to redundancy and proposed a set of five redundancy measures ranging from the set difference, geometric mean, distributional similarity to calculate the novelty of an incoming document with respect to a set of memorized documents. They also presented the first publicly available Associated Press-Wall Street Journal (APWSJ) news dataset for document-level novelty detection. Tsai and Zhang (2011) applied a document to sentence-level framework to calculate the novelty of each sentence of a document which aggregates to detect novelty of the entire document. Karkali *et al.* (2013) computed novelty score based on the inverse document frequency scoring function. Another work by Verheij *et al.* (2012) presents a comparison study of different novelty detection methods evaluated on news articles where language model-based methods perform better than the cosine similarity-based ones. More recently, Dasgupta and Dey (2016) conducted experiments with information entropy measure to calculate the innovativeness of a document. Zhao and Lee (2016) proposed an intriguing idea of assessing the novelty appetite of a user based on a curiosity distribution function derived from curiosity arousal theory and Wundt curve in psychology research.

Novelty detection is also studied in information retrieval literature for diversity detection. The idea was to retrieve relevant, yet diverse documents in response to a user query. The work on Maximal Marginal Relevance by Carbonell and Goldstein (1998) was the first to explore diversity and relevance for novelty. Some other notable works along this line are Chandar and Carterette (2013), Clarke *et al.* (2011, 2008).

Our proposed work significantly differs from the existing literature in terms of the methodology adopted and the problem addressed. The major contributions of the current work are

1. Proposing an efficient deep learning architecture for document-level novelty detection, outperforming existing methods on two benchmark datasets;
2. Presenting an effective architecture based on an interpretable intuition: *alignment of sentences via attention to detect novelty of a document with inference knowledge gained from a large-scale NLI corpus.*

We discuss our approach in the subsequent section.

3. Proposed method

As discussed earlier, we intend to make use of attention mechanism to identify the potential contributing sections in the source document(s) that makes a target document appear *non-novel* to the reader. We draw inspiration from the work of Parikh *et al.* (2016) on decomposable attention for NLI. Using *Attend, Compare,* and *Aggregate* steps in the model, they successfully show the importance of attention to identify the contributing sections in a sentence for inference decisions. NLI is one such task that closely resembles the notion of *non-novelty*. The relation of novelty detection with entailment is extensively studied in the RTE-TAC novelty subtasks and associated literature (Bentivogli *et al.* 2011). This is also why we are motivated to base our model on the knowledge learned from an entailment dataset. Hence, we proceed to transfer inference knowledge from an external NLI dataset to our text representations, train our proposed model on the novelty detection datasets, and thereby investigate the performance. We hypothesize that for a piece of *novel* text, there should be no contributing section (i.e., text having similar/overlapping information

https://doi.org/10.1017/S1351324920000194 Published online by Cambridge University Press

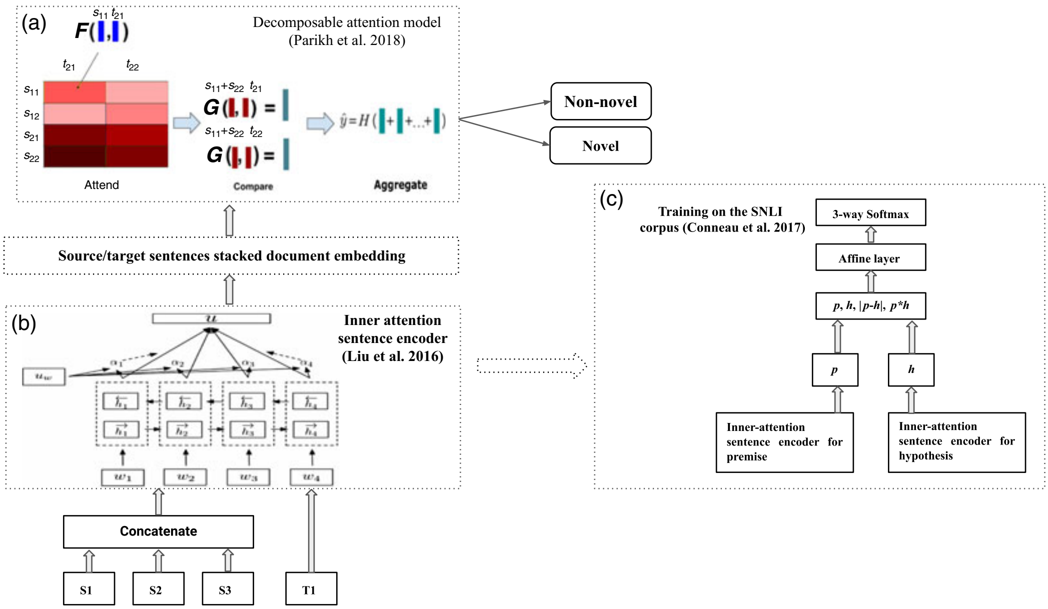

Figure 1. Overall architecture for document-level novelty detection. Component (b) is the inner-attention sentence encoder. Component (c) shows how the inner-attention sentence encoder is trained on the SNLI corpus. Component (a) is the sentence-level decomposable attention model we use in our work (Section 3.2) for document-level novelty detection. s_{11}, s_{12} represent the two sentences in source document d_1 in the example introduced in Section 1. s_{21}, s_{22} are the two sentences in source document d_2. d_1 and d_2 are concatenated to form a single source document. t_{21}, t_{22} are the two sentences in target document d_4. Simply reading the example, we can conclude that t_{21} and t_{22} directly follow from s_{11} and s_{22}. d_4 is redundant if we consider d_1 and d_2 as the source documents.

content: either lexically or semantically) in the source texts. We leverage the word-level attention model to sentence-level attention to effectively model document-level novelty detection. Instead of generating a complex joint representation of the source and target texts with dominant deep neural paradigms, the decomposable attention model (Parikh *et al.* 2016) relies on simple *alignment* of target text with the source text. The decomposable model is also found to be efficient in terms of the number of parameters as compared to the other models for modeling entailment pairs.

Initially, we encode our sentences using intrasentence attention trained on the SNLI corpus to capture the rich semantic perspectives involved in sentence-level inference decisions (Section 3.1). Then, we make use of the decomposable attention model to learn the notion of document-level novelty and redundancy from the aggregated representation (Section 3.2). We depict the overall architecture in Figure 1.

Here, T_1 is the *target* document whose state of *novelty* is to be determined against the source document(s) S_1, S_2, S_3. Although we investigate novelty at the document-level, we rely on sentence-level interactions among the source and target texts. Hence, we split the source and target texts into component sentences and generate the sentence encodings. Please note that our architecture is not end-to-end trained. In the first phase, we train the sentence encodings on the semantically rich SNLI dataset. In the second phase, we train the decomposable attention model on the novelty detection datasets (APWSJ, TAP-DLND 1.0). The sentences in the novelty datasets are vectorized with the sentence encodings from SNLI training.

3.1 Sentence encoding

The task of *novelty detection* requires high-level understanding and reasoning about semantic relationships within texts. TE or NLI is one such task which exhibits such complex semantic

https://doi.org/10.1017/S1351324920000194 Published online by Cambridge University Press

interactions. Hence, we train our sentence encodings on the very large (about 570k text pairs) and semantically rich SNLI dataset (Bowman *et al.* 2015). We use open-source GloVe vectors trained on Common Crawl 840B with 300 dimensions as fixed word embeddings. There are several other state-of-the-art sentence embedding techniques like the Universal Sentence Encoder (Cer *et al.* 2018), self-attentive sentence encoder (Lin *et al.* 2017), unsupervised sentence embedding using weighted average of word vectors (Arora, Liang, and Ma 2016), etc., but we use the sentence encoder by Conneau *et al.* (2017) in order to have the NLI semantic interactions in our sentence embeddings.

3.1.1 SNLI dataset

The SNLI corpus (Bowman *et al.* 2015) is a collection of 570k human-written English sentence pairs manually labeled for balanced classification with the labels entailment, contradiction, and neutral, supporting the task of NLI, also known as RTE. As discussed earlier, we seek to transfer the inference knowledge of SNLI to our sentence embeddings.

3.1.2 Training the sentence encoder on external linguistic resource: SNLI

We already explained the reason why we chose SNLI to generate our sentence embeddings in preceding sections. Conneau *et al.* (2017) demonstrated that sentence encoder trained on NLI corpus could learn sentence representations that capture universally useful features. We follow the idea of a siamese neural network (Figure 1(c)). It denotes that two identical sentence encoders share the same set of weights during training, and the two sentence representations (premise p and the hypothesis h) are then combined to generate a "relative vector" for classification. A typical architecture of this kind uses a shared sentence encoder that outputs a representation for the premise p and the hypothesis h. Three matching methods (Mou *et al.* 2015) are applied to extract the relations between p and h: (i) concatenation of the two representations (p, h); (ii) element-wise product $p * h$; and (iii) absolute element-wise difference $|p - h|$. The resulting vector, which captures information from both premise and the hypothesis, is fed into a three-class classifier (neutral, entailment, contradiction: class labels in SNLI) consisting of multiple fully connected layers culminating in a softmax layer.

3.1.3 Bi-LSTM + inner-attention sentence encoder

We use an inner-attention sentence encoder (Liu *et al.* 2016) to generate a representation u of an input sentence. This encoder employs attention mechanism on the representation produced in previous hidden state of a Bi-Directional Long Short Term Memory (Bi-LSTM) to attend important words in the sentence itself. This is inspired from the *inner-attention* idea from Liu *et al.* (2016) where the author says that readers usually form a rough intuition about which part of the sentence is more important, usually from their past experiences. For a sequence of T words $\{w_t\}_{t-1,...,T}$, a Bi-LSTM computes a set of T vectors $\{h_t\}_t$. For $t \in [1, ..., T]$, h_t is the concatenation of a forward LSTM and a backward LSTM that read the sentences in two opposite directions:

$$\overrightarrow{h_t} = \overrightarrow{LSTM}_t(w_1, ..., w_T)$$

$$\overleftarrow{h_t} = \overleftarrow{LSTM}_t(w_1, ..., w_T)$$

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$$

The attention mechanism is formalized as follows:

$$\bar{h_i} = tanh(Wh_i + b_w)$$

$$\alpha_i = \frac{e^{\bar{h_i}^T u_w}}{\sum_i e^{\bar{h_i}^T u_w}} \qquad u = \sum_i \alpha_i h_i$$

https://doi.org/10.1017/S1351324920000194 Published online by Cambridge University Press

where $(h_1, ..., h_T)$ are the output hidden vectors of a Bi-LSTM. These are fed to an affine transformation (W, b_w) which outputs a set of keys $(\bar{h_1}, ..., \bar{h_T})$. The α_i represents the score of similarity between the keys and a learned context query vector u_w. These weights are used to produce the final representation u, which is a weighted linear combination of the hidden vectors.

3.2 The decomposable attention model

The decomposable attention model was first proposed by Parikh *et al.* (2016) for word-level alignment to model sentence-level inference. We take inspiration from their work and make use of decomposable attention for sentence alignment to model identification of document-level novelty. The objective of the decomposable model is to decompose the task into the subproblems that are solved separately. In contrast to recent dominant complex approaches for NLI, this approach only relies on the alignment of local text substructure to aggregate to a prediction.

Let $x = [x_1, x_2, ...x_m]$ denote the set of all sentences concatenated from the source documents and $y = [y_1, y_2, ...y_n]$ denote the set of sentences in the target document. The corresponding sentence vector representation (c.f. Section 3.1.2) of these sentences is denoted by $\bar{x} = [\bar{x_1}, \bar{x_2}, ...\bar{x_m}]$ and $\bar{y} = [\bar{y_1}, \bar{y_2}, ...\bar{y_n}]$. The core decomposable model (Figure 1(c)) consists of the following three components, which are trained jointly:

1. **Attend:** The attention layer uses a variant of neural attention proposed in Bahdanau *et al.* (2014). We implement it using a feed-forward neural network F that is applied to both sentences separately. We soft align the elements of $\bar{x}$ and $\bar{y}$ via attention mechanism and decompose the problem into the comparison of aligned sentences. We obtain the unnormalized attention weights e_{ij}, computed by a function F', which decomposes as follows:

$$e_{ij} := F'(\bar{x}_i, \bar{y}_j) := F(\bar{x}_i)^T F(\bar{y}_j)$$

F is a feed-forward neural network with Rectified Linear Unit (ReLU) activations. These attention weights are normalized as follows:

$$\beta_i = \sum_{j=1}^{n} \frac{exp(e_{ij})}{\sum_{k=1}^{n} exp(e_{ik})} \bar{y}_j$$

$$\alpha_j = \sum_{i=1}^{m} \frac{exp(e_{ij})}{\sum_{k=1}^{m} exp(e_{kj})} \bar{x}_i$$

Here, β_i is the text segment in $\bar{y}$ that is (softly) aligned to $\bar{x}_i$ and the vice versa for α_j.

2. **Compare:** Next step is to compare the soft-aligned document matrices. Similarly to the previous step, we use a feed-forward neural network G to compare the aligned text segments $\{(\bar{x}_i, \beta_i)\}_{i=1}^{m}$ and $\{(\bar{y}_j, \alpha_j)\}_{j=1}^{n}$

$$\mathbf{v}_{1,i} = G([\bar{x}_i, \beta_i]) \qquad \mathbf{v}_{2,j} = G([\bar{y}_j, \alpha_j])$$

where the brackets [., .] denote concatenation.

3. **Aggregate:** The last part of the core model architecture is the aggregation layer. All this layer does is a column-wise sum over the output of the comparison network:

$$\mathbf{v}_1 = \sum_{i=1}^{m} \mathbf{v}_{1,i} \qquad \mathbf{v}_2 = \sum_{j=1}^{n} \mathbf{v}_{2,j}$$

and feed the output through a final classifier H, which is again a feed forward network followed by a linear layer:

https://doi.org/10.1017/S1351324920000194 Published online by Cambridge University Press

$$\hat{y} = H([\mathbf{v}_1, \mathbf{v}_2])$$

where $\hat{y}$ represents the predicted scores for each class (novel or non-novel). Output layer of the classifier is softmax normalized so that we obtain a probability distribution over target classes and consequently the predicted class is given by $\hat{y} = argmax_i \hat{y}_i$

3.2.1 Hyperparameter settings and parameter tuning

The training of the sentence encoder on SNLI is carried out with 20 epochs, batch size = 64 with Stochastic Gradient Descent optimizer and learning rate set to 0.001. The output sentence vector dimension *SENT_DIM* is 2048. For the decomposable model, we use epochs = 30, batch size = 25, Adam optimizer with learning rate set to 0.001 and the loss function as categorical cross entropy. The comparison layer compares each target sentence with its corresponding alignment, finally outputs a vector of dimension 2**SENT_DIM*. Next, the entailment layer is a fully connected layer with two hidden layers and *SENT_DIM* neurons with ReLU activation. Finally, softmax layer is used for the classification.

The Bi-LSTM inner-attention sentence encoder has two affine layers without bias. The first and second layer in the SNLI training has 512 neurons each and the classification layer has three neurons (entailed, contradiction, neutral). The attention layer in the decomposable model is actually a Bi-Directional Recurrent Neural Network (Bi-RNN) over the sentence vector (2048 dimension) and then dot product. The compare step is a Bi-RNN over concatenation of original and the aligned sentence. Finally, in the aggregate layer, we have an affine transformation to 2048 neurons and finally a prediction layer of two neurons.

4. Dataset description

The most popular datasets for novelty detection are the ones released in TREC 2002–2004 (Harman 2002; Soboroff and Harman 2003) and RTE-TAC 2010–2011 (Bentivogli *et al.* 2011). Although these datasets are for sentence-level novelty mining, and hence do not cater to our document-level investigation needs. However, we experiment with two document-level novelty detection datasets: the APWSJ (Zhang *et al.* 2002) and on our recently released TAP-DLND 1.0 (Ghosal *et al.* 2018b) with which we perform our experiments. Both these datasets are in the newswire domain.

The SNLI dataset is the basis of our knowledge transfer from the inference task to the novelty detection problem. It is not our testbed for the problem in hand, rather a linguistic resource that encodes rich semantic knowledge for NLI or TE, which is closely related to textual novelty detection (Bentivogli *et al.* 2011). As discussed earlier, we use SNLI for the creation of efficient sentence embeddings (eventually document representations) that inherently manifest the semantic interactions needed in an inference decision. This inference knowledge helps us to better discover traces of *novel* and *non-novel* patterns in a target document given that our neural architecture has already seen some relevant documents.

4.1 APWSJ corpus

The APWSJ data consist of news articles from the AP and WSJ corpus covering the same period (1988–1990) and many on the same topics, guaranteeing some redundancy in the document stream. There are 11,896 documents on 50 topics (Q101–Q150 TREC topics). After sentence segmentation, these documents have 319,616 sentences in all. The APWSJ data contain a total of 10,839 (91.1%) novel documents and 1057 (8.9%) non-novel documents. However, similar to Zhang *et al.* (2002), we use the documents within the designated 33 topics[b] with redundancy judgments by the assessors. The dataset was meant to filter superfluous documents in a retrieval

[b] http://www.cs.cmu.edu/~yiz/research/NoveltyData/.

https://doi.org/10.1017/S1351324920000194 Published online by Cambridge University Press

scenario to deliver only the documents having *redundancy score* below a calculated threshold. Documents for each topic were delivered chronologically, and the assessors provided two degrees of judgments on the non-novel documents: *absolute redundant* or *somewhat redundant* based on the preceding documents. The unmarked documents are treated as *novel*. However, since there is a huge class imbalance, we follow Zhang *et al*. (2002) and include the *somewhat redundant* documents also as *redundant* and finally arrive to ∼37% *non-novel* documents in APWSJ. Finally, we had the total number of instances: 5789, total number of new instances: 3656, total number of non-novel instances: 2133. The percentage of novel instances for the actual experiments now stands at 63.15%.

4.2 TAP-DLND 1.0 corpus

We experiment with our new benchmark resource for document-level novelty detection (Ghosal *et al*. 2018b). The dataset[c] is balanced and consists of 2736 novel documents and 2704 non-novel documents. There are several categories of events and we track the development of a news item across time. For a particular event, we select a set of documents as the source and the rest as target. We asked the annotators to judge the information content in the target (for novelty) against the source documents only. For each novel/non-novel document, there are three source documents against which the target documents are annotated. The state of *novelty* for each target document is measured against those source documents only, that is, once the system has already seen the designated source documents for a particular event, it is to judge whether an incoming on-topic document is novel or not. We follow the following annotation principles:

(a) To annotate a document as non-novel whose semantic content significantly overlaps with the source document(s) (maximum redundant information).
(b) To annotate a document as novel if its semantic content as well as intent (direction of reporting) significantly differs from the source document(s)(minimum or no information overlap). It could be an update on the same event or describing a post-event situation.
(c) We left out the ambiguous cases (for which the human annotators were not sure about the label).

That the dataset manifests semantic-level redundancy, goes beyond lexical similarity, makes it an ideal candidate for our experiments. The inter-rater agreement is 0.82 in terms of Fleiss Kappa (Fleiss 1971), and the average length of documents is 15 sentences/353 words.

5. Evaluation

We evaluate our proposed approach on the two benchmark datasets discussed above and present the results in the following sections. As discussed before, we have two segments in our proposed model:

1. The inner-attention sentence encoder trained on the very large scale SNLI corpus.
2. The decomposable attention model trained on the document-level novelty detection datasets separately with the proposed method.

Thus, our model is not end-to-end trained. We leverage the knowledge of NLI (first segment) to classify documents as novel or non-novel (core task; second segment).

5.1 Baselines and comparing systems

We adopt the following baselines to investigate the strength of various factors in our model as well as help in ablation studies. We also compare with the available up-to-date results on APWSJ and TAP-DLND 1.0 datasets to better scrutinize our performance.

[c] http://www.iitp.ac.in/~ai-nlp-ml/resources.html.

https://doi.org/10.1017/S1351324920000194 Published online by Cambridge University Press

- **Baseline 1 (without decomposable attention model):** We choose this baseline to see the effect of withholding the second segment (i.e., the decomposable attention module) of our model. Here, we train the sentence encoder (based on inner-attention) on SNLI. We generate the document matrix for the source[d] and target documents by stacking the corresponding sentence encodings. A Bi-LSTM layer encodes each document matrix and produces a fixed-sized vector of dimension *SENT_DIM* for the source and target document(s) separately. The two resultant vectors are then concatenated and passed to a fully connected entailment layer with two hidden layers having *SENT_DIM* neurons (basically a Multi-Layered Perceptron) and *ReLU* activation followed by classification via softmax.
- **Baseline 2 (without inner-attention sentence encoder):** Instead of using attention mechanism over the hidden states of the Bi-LSTM to generate a representation u of an input sentence, we select the maximum value over each dimension of the hidden units (max pooling) (Conneau *et al.* 2017).
- **Baseline 3 (without pre-training on SNLI):** To see how much pre-training of the sentence encoder with the large-scale SNLI affects our system performance, we ablate the training of the sentence encoder on SNLI. We use the pre-trained paragraph vector[e] (*doc2vec*; Le and Mikolov 2014) to generate the sentence encodings. Here, we want to see how much of the contribution of the pre-training step comes from training on SNLI as opposed to simply using pre-trained embeddings (doc2vec) on a news dataset (Associated Press News DBOW). We then train the decomposable model (Section 3.2) and finally classify the target document.
- **Comparing systems (previously published results):** We compare the performance of our proposed system against the novelty measures proposed by Zhang *et al.* (2002) for APWSJ. The three novelty detection measures from Zhang *et al.* (2002) are set difference (*Novelty Measure 1*), geometric distance (*Novelty Measure 2*), and language model (*Novelty Measure 3*). For TAP-DLND 1.0, we re-implement these measures along with another one (*Novelty Measure 4*) which is based on inverse document frequency (Karkali *et al.* 2013). For comparison, we also consider another approach based on calculating the relative entropy of a document (Dasgupta and Dey 2016). Instead of setting a fixed threshold[f] as in these works, we train a logistic regression (LR) classifier based on those measures to automatically determine the decision boundary.

  We also compare our approach with the current benchmark on TAP-DLND 1.0 by Ghosal *et al.* (2018b) and the RDV-CNN architecture (Ghosal *et al.* 2018a) for document-level novelty detection. In our feature-based solution (Ghosal *et al.* 2018b) to the problem, we make use of several features like document similarity, divergence, NEs, keywords, etc. In our subsequent work (Ghosal *et al.* 2018a), we employ a CNN-based deep model that learns the notion of novelty and redundancy from a derived vector representation from source and target documents which we term as the relative document vector (RDV). We trained the sentence embeddings on the SNLI dataset using a Bi-LSTM with max pooling (Conneau *et al.* 2017) technique. We then pulled the nearest source sentence to a given target sentence and concatenated them using the representation of Mou *et al.* (2015). Finally, we stacked the source-encapsulated target sentence representations to form the RDV. We used a CNN to extract features from the RDV to classify a given target documents as novel or non-novel.

  It is to be noted here that except Ghosal *et al.* (2018a,b), Dasgupta and Dey (2016), and our baselines, all the other comparing systems were developed from an information retrieval perspective. Whereas in our earlier attempt (Ghosal *et al.* 2018a), we use a CNN-based model

[d] Concatenated for multiple source documents.
[e] https://github.com/jhlau/doc2vec.
[f] The weak thresholding algorithm reported in these works yield poor results.

Table 1. Document-level Novelty Classification Results on TAP-DLND 1.0. We achieve better performance than the earlier reported works (indicated by figures in bold faces)

System	P(N)	R(N)	P(NN)	R(NN)	F_1	A(%)
Baseline 1†	0.74	0.85	0.88	0.78	0.81	81.4
Baseline 2†	0.74	0.46	0.69	0.88	0.69	70.4
Baseline 3†	0.74	0.43	0.68	0.89	0.68	69.5
Novelty Measure 1*	0.74	0.71	0.72	0.74	0.72	73.2
Novelty Measure 2*	0.65	0.84	0.84	0.55	0.69	69.8
Novelty Measure 3*	0.73	0.74	0.74	0.72	0.73	73.6
Novelty Measure 4†	0.52	0.92	0.66	0.16	0.45	54.2
Dasgupta and Dey (2016)	0.63	0.72	0.77	0.66	0.69	68.2
Ghosal *et al*. (2018b)†	0.77	0.82	0.80	0.76	0.78	79.3
Ghosal *et al*. (2018a)†	0.86	0.87	0.84	0.83	0.85	84.5
Proposed approach†	0.85	0.85	**0.89**	**0.89**	**0.87**	**87.4**

P, Precision; R, Recall; F1, Average F-Score; N, Novel; NN, Non-Novel; LR, Logistic Regression; IDF, Inverse Document Frequency.
†10-*fold* cross-validation output.
*Measures from Zhang *et al*. (2002) with LR.
†Measure from Karkali *et al*. (2013) with LR.

Table 2. Document-level Redundancy Classification (Non-Novel) Results on APWSJ. We achieve better performance than the earlier reported works (indicated by figures in bold faces)

Measure	R(NN)	P(NN)	Mistake (%)
Set distance*	0.52	0.44	43.5
Cosine distance*	0 62	0 63	28 1
LM: Shrinkage*	0.80	0.45	44.3
LM: Dirichlet prior*	0.76	0.47	42.4
LM: Mixture model*	0.56	0.67	27.4
Ghosal *et al*. (2018a,b)†	0.58	0.76	22.9
Proposed approach†	**0.86**	**0.92**	**7.8**

LM, Language Model; Mistake, 100-Accuracy.
*Results taken from Zhang *et al*. (2002).
10-*fold* cross-validation output.

on a derived source–target representation, here in our current work, we use an attention-based model to aggregate aligned information from source and target documents to identify novelty and redundancy.

5.2 Results and discussions

We show the comparatively better performance of our approach in the evaluation figures of Table 1 (TAP-DLND 1.0) and Table 2 (APWSJ). The current approach outperforms our previous two *state-of-the-art* approaches by a margin of 8.1% and 2.9% in terms of accuracy. Our

https://doi.org/10.1017/S1351324920000194 Published online by Cambridge University Press

initial system is feature-based (Ghosal *et al.* 2018b). Although our second system is a deep neural architecture (Ghosal *et al.* 2018a), the current approach can surpass its performance with lesser complexity in terms of the number of parameters used. Our baselines also serve as a good means of ablation study.

- We observe that the novelty measures (Zhang *et al.* 2002; Karkali *et al.* 2013) do not perform well in identifying novel or non-novel documents. This is because these measures do not consider the semantics involved in recognizing novelty. These lexical measures also do not consider context information which is of utmost importance to operating at the discourse level.
- Without the inner-attention sentence encoder (Baseline 2), we see that the system does not perform good to classify the target documents. This resonates the observation made by Liu *et al.* (2016) that it is necessary to emphasize certain sections of texts (achieved via inner-attention) based on the already seen data or neighborhood data (i.e., the context).
- **Impact of inference knowledge from SNLI:** When we remove the SNLI pre-training of the sentence encoder (Baseline 3), the performance of the proposed system drops by a considerable amount (∼18% in terms of accuracy). This behavior supports our intuition that upon training on the large-scale SNLI corpus, the sentence embeddings can capture the semantics involved in understanding NLI. As discussed in the preceding sections, NLI is one such phenomenon that closely simulates our reckoning of non-novelty at the semantic level. Thus, we could say that the inference knowledge transfer from SNLI highly correlates with the text alignment perspective for novelty detection.
- Baseline 1 shows the effect of not having the decomposable attention module in our system. Although we have the inner-attention sentence vectors pre-trained on SNLI, it is evident that it costs us F_1 by a margin of ∼6 points. This is because, in this model, the resultant vectors obtained from the source and target document(s) do not have the information: which source text should a target sentence be attending to determine the novelty of the target sentence. We deem that this step is essential to adjudge whether a document is novel/non-novel w.r.t. a set of given documents (or known information). Eventually, the novelty of constituent sentences would lead us to the judgment of novelty of the given document. Particularly, we see that the recall for non-novel (NN) documents is low which points that the pairing of target sentence with its corresponding source sentence (which we achieve via attention) is important to create the resultant vector before feature extraction and classification via a neural network.
- Our current approach supersedes our previous hand-crafted feature-based system on TAP-DLND 1.0 (Ghosal *et al.* 2018b) by a margin of ∼8% in terms of accuracy. This is particularly encouraging in the sense that our proposed model can capture the textual characteristics required to understand document novelty from the data itself. Also, our current proposed approach shows improvement over our recently proposed RDV-CNN (Ghosal *et al.* 2018a) architecture for document-level novelty detection (∼3% in terms of accuracy). The current attention-based architecture is also simple as compared to RDV-CNN in terms of implementation and in the order of parameters used.
- Table 2 is a strong testimony that the proposed approach is effective to identify the non-novel documents. The alignment of the target sentence to the corresponding source sentence(s) and subsequent feature discovery via neural layers proves to be appropriate for recognizing redundant documents in APWSJ. It is particularly encouraging to see that even though there are a lesser number of non-novel documents in APWSJ (which is more likely a practical scenario), attention mechanism enables to identify the sentences which actually contribute to making a target sentence *non-novel*.

https://doi.org/10.1017/S1351324920000194 Published online by Cambridge University Press

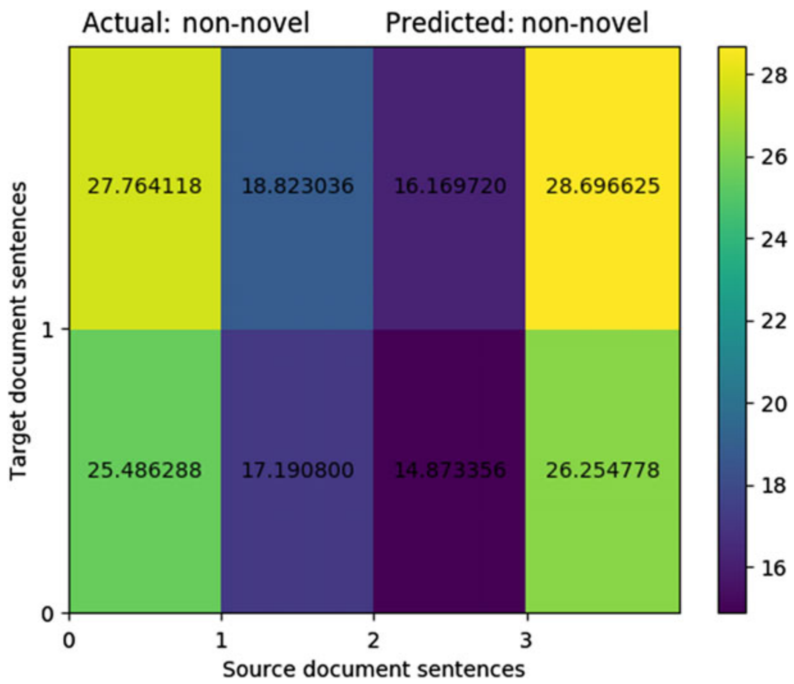

Figure 2. Attention matrix visualization via heat map for the example in Section 1. *d1* and *d2* are concatenated to form the source document. *d4* is the target document.

- The results strongly corroborate our intuition that textual alignment of sentences (between the source and target documents) could lead to a better understanding of document novelty. The alignment should not necessarily be on a one-to-one sentence basis. Rather a target sentence may have multiple information sources (sentences). The multiple premise cases are well handled with the distributed attention weights over multiple sentences, as discussed in the Attend step in Section 3.2 of the decomposable attention model. But for efficient attending of source document sentences (identifying the premise) by a target document sentence (the hypothesis), we need sentence embeddings that could manifest the inference perspective. We achieve that via our inner-attention sentence encoder trained on the large-scale SNLI corpus. Thus, a relation between textual inference and alignment is established, which proves crucial to novelty detection.

The accuracy of the proposed model is statistically significant over the baselines (two-tailed t-test, $p < 0.05$).

5.3 Analysis

The strength of our approach stems from the very objective we start to investigate: *alignment* via *attention* with the value addition of *inference*. Considering the example in Section 1, we employ our approach and plot the attention weights obtained from the *Attend* step (Section 3.2) in Figure 2. From the example, we consider *d4* as the target document against the source documents *d1* and *d2*. We concatenate *d1* and *d2* to form one single source having four sentences. From the attention weights, we can clearly see that the two target sentences (in *d4*) are highly attending the first and fourth source sentences (the corresponding attention weights are high), signifying a multiple premise scenario. However, the second and third source sentence having diverse information finds less alignment or attention of the target sentences. Our method correctly predicts the class label, as well. Thus, if there is less attention (less attention weights for most sentence pairs between

https://doi.org/10.1017/S1351324920000194 Published online by Cambridge University Press

source and target), which indicates fewer appropriately aligned sentences, which in turn signifies that the majority of the sentences in the target are distant from each sentence in the source, we arrive to the conclusion that the tendency of the target document is towards novelty. The target document sentences are not highly attending or aligned to any source document sentence or the target sentences found no premise in the source document(s); the target document supposedly has new information. The corresponding heat map would look like Figure 3(a) signifying that the document is *novel*. Figure 3(b) is the heatmap of a correctly predicted non-novel document from TAP-DLND 1.0 pitched against the corresponding source document(s).

5.4 Error analysis

We perform a thorough analysis of our predictions. Figure 3(c), (d), (e) shows some error instances. Majority of the errors committed by our system are due to the presence of:

1. Multiple complex premise sentences in the source documents for a target sentence. The attention weights are evenly distributed, sometimes making it hard to identify the exact premises.
2. Long compound sentences and conveying a greater amount of information, resulting in misalignment.
3. Higher number of NEs, which are often over-emphasized by the attention model. This often contributes to false negatives. In spite of conveying different information but due to the presence of same NEs, sentences are misaligned (given higher attention weights).
4. Dichotomy in annotation judgments in both the datasets. We manually went through the annotation judgments and in certain cases were not sure about the ground-truth. Subjectivity plays a crucial role here. The novelty appetite is not the same for all readers (Zhao and Lee 2016). The amount of new information that makes a document appear novel to one reader may not be the same to another reader.
5. We cannot always establish a simple mapping between sentences in the source and target documents. Sometimes target documents consist of background information (world knowledge; pragmatics) that has relevance with the source but are not explicitly mentioned.
6. Errors in sentence splitting (see in Appendix) and the difference in document lengths. Some source documents after concatenation are too long and manifest all the information in the target document. Hence, the target document should be non-novel. But, the information in the concatenated source document is distributed all across the source sentences. Thus, there is no emphasis on one or two source sentences. Hence finding a suitable source text segment to get aligned with a target sentence is difficult. Therefore, although the target is non-novel, but the model predicted it as a novel.

We pulled out some examples from TAP-DLND 1.0 and demonstrate the error categories (See heatmaps in Figure 3(c), (d), (e)). Kindly consult the Appendix for full-text. The following instances are actually non-novel but predicted as novel.

- **Instance #45:** Observations (Figure 3(c))
  - — Target sentence #15 has many premises so its attention values are spread throughout the source documents decreasing its importance.
  - — Target sentence #11 conveys important information but its complex structure and long length made it difficult for the model to capture its attention values.
  - — Target sentence #6 has some NEs which unusually increased its attention values.
- **Instance #381:** Observations (Figure 3(d))
  - — Target sentence #2 conveys important information but due to its long length resulted in low attention values

https://doi.org/10.1017/S1351324920000194 Published online by Cambridge University Press

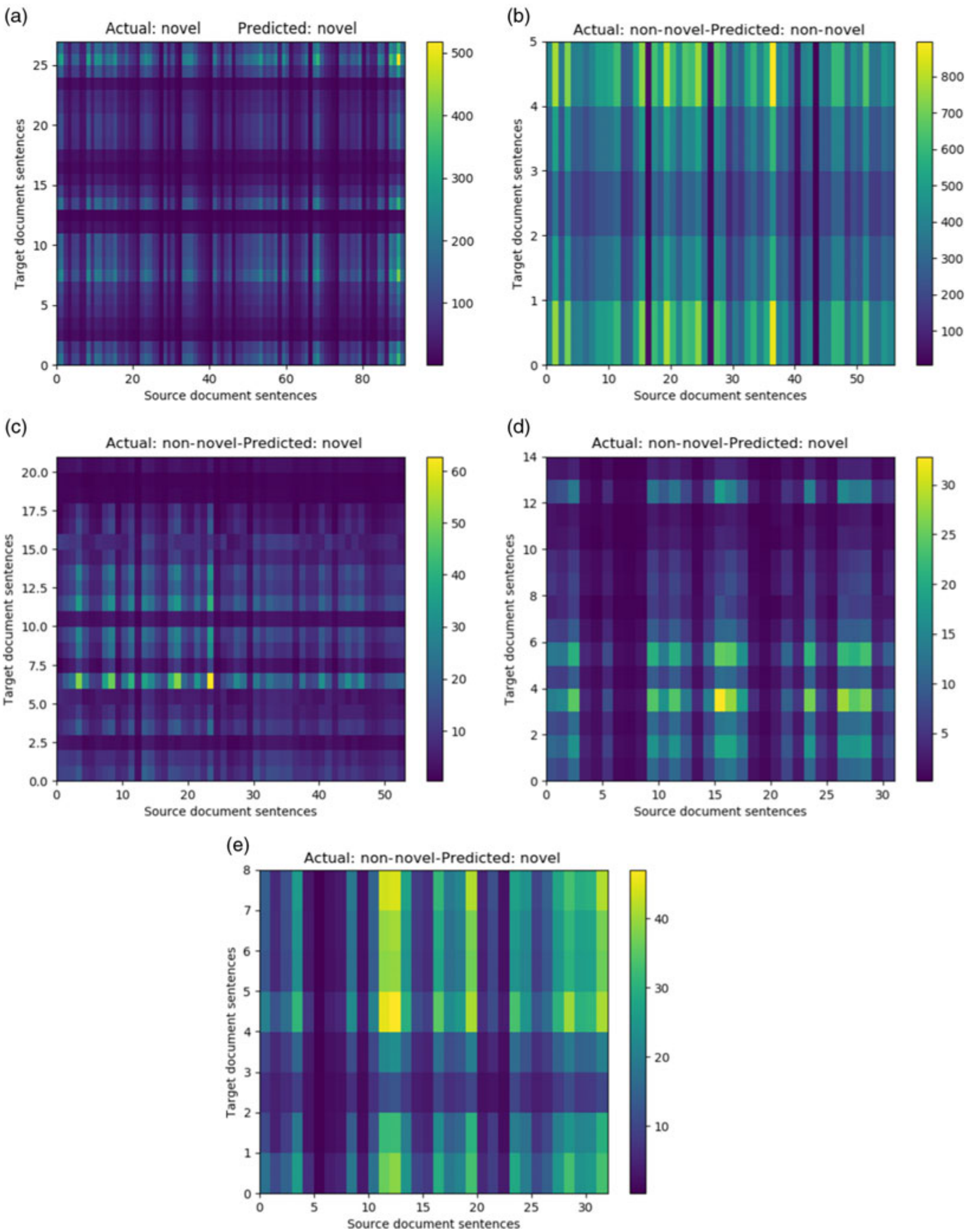

Figure 3. Attention matrix visualization via heat map for a correctly predicted (a) novel and (b) non-novel document from TAP-DLND 1.0. (a) Many dark patches signify that most of the target sentences are not highly attending to any source sentence. Hence, may contain sufficient new information. (b) Lesser dark patches indicate that the target sentences are highly aligned to the source sentences and may contain redundant information. Wrongly predicted instances → (c), (d), (e).

— Target sentence #5 is over emphasized due to high NEs

- **Instance #495:** Observations (Figure 3(e))

  — Target sentence #0 and sentence #2 are important, but due to their long length and complexity, the model could not capture there attention values appropriately.

We analyzed 250 false positive and false negative cases. Category 1 and Category 2 errors together contributed to nearly ∼71% of the misclassified instances.

6. Conclusion

In this paper, we present an efficient attention-based deep neural architecture on categorizing a document as novel or non-novel based on its information content. The foundation idea of our proposed model is *alignment via attention to detect novelty*, which is relatively less complex than the other existing deep neural architecture for document-level novelty detection. With our model, we are able to outperform the baselines and state-of-the-art on two datasets by a good margin. Results and analysis establish that by simply aligning target document sentences with source document sentences, it is possible to conclude upon the state of the novelty of the target document. We empirically show that incorporation of inference knowledge in our sentence representations from an external NLI corpus is essential to the performance of our system. Our proposed architecture and corresponding results also establish the relation of alignment and inference toward the novelty detection task. In the future, we would like to investigate better semantic representations of multiple premises (in source documents), tackle NEs and longer sentences before the attention step in the network for more accurate correspondence between source and target document information.

Acknowledgements. The first author, Tirthankar Ghosal, acknowledges Visvesvaraya PhD Scheme for Electronics and IT, an initiative of Ministry of Electronics and Information Technology (MeitY), Government of India for fellowship support. The third author, Asif Ekbal, acknowledges Young Faculty Research Fellowship (YFRF), supported by Visvesvaraya PhD scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia). The authors also thank Elsevier Center of Excellence for Natural Language Processing, Indian Institute of Technology Patna for adequate infrastructural support to carry out this research. Finally, the authors appreciate the anonymous reviewers for their critical evaluation of our work and suggestions to carry forward from here.

References

Allan J., Gupta R. and Khandelwal V. (2001). Topic models for summarizing novelty. In *ARDA Workshop on Language Modeling and Information Retrieval. Pittsburgh, Pennsylvania.*

Allan J., Lavrenko V. and Jin H. (2000). First story detection in TDT is hard. In *Proceedings of the Ninth International Conference on Information and Knowledge Management*. ACM, pp. 374–381.

Allan J., Papka R. and Lavrenko V. (1998). On-line new event detection and tracking. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, pp. 37–45.

Allan J., Wade C. and Bolivar A. (2003). Retrieval and novelty detection at the sentence level. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*. ACM, pp. 314–321.

Arora S., Liang Y. and Ma T. (2016). A simple but tough-to-beat baseline for sentence embeddings. In *5th International Conference on Learning Representations, 2017, Toulon, France, April 24–26, 2017*, Conference Track Proceedings.

Bagga A. and Baldwin B. (1999). Cross-document event coreference: Annotations, experiments, and observations. In *Proceedings of the Workshop on Coreference and its Applications*. Association for Computational Linguistics, pp. 1–8.

Bahdanau D., Cho K. and Bengio Y. (2014). Neural machine translation by jointly learning to align and translate. CoRR abs/1409.0473.

Bentivogli L., Clark P., Dagan I. and Giampiccolo D. (2011). The seventh pascal recognizing textual entailment challenge. In *TAC*.

Bowman S.R., Angeli G., Potts C. and Manning C.D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17–21,* 2015, pp. 632–642.

Brants T., Chen F. and Farahat A. (2003). A system for new event detection. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*. ACM, pp. 330–337.

Bysani P. (2010). Detecting novelty in the context of progressive summarization. In *Proceedings of the NAACL HLT 2010 Student Research Workshop*. Association for Computational Linguistics, pp. 13–18.

Carbonell J. and Goldstein J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 335–336.

Carpenter G.A., Rubin M.A. and Streilein W.W. (1997). Artmap-fd: Familiarity discrimination applied to radar target recognition. In *Proceedings of International Conference on Neural Networks (ICNN'97)*, vol. 3. IEEE, pp. 1459–1464.

Cer D., Yang Y., Kong S., Hua N., Limtiaco N., John R.S., Constant N., Guajardo-Cespedes M., Yuan S., Tar C., Strope B. and Kurzweil R. (2018). Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: System Demonstrations*, Brussels, Belgium, October 31–November 4, 2018, pp. 169–174.

Chandar P. and Carterette B. (2013). Preference based evaluation measures for novelty and diversity. In *SIGIR*.

Clarke C.L.A., Craswell N., Soboroff I. and Ashkan A. (2011). A comparative analysis of cascade measures for novelty and diversity. In *WSDM*.

Clarke C.L.A., Kolla M., Cormack G.V., Vechtomova O., Ashkan A., Büttcher S. and MacKinnon I. (2008). Novelty and diversity in information retrieval evaluation. In *SIGIR*.

Collins-Thompson K., Ogilvie P., Zhang Y. and Callan J. (2002). Information filtering, novelty detection, and named-page finding. In *TREC*.

Conneau A., Kiela D., Schwenk H., Barrault L. and Bordes A. (2017). Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017,* Copenhagen, Denmark, September 9–11, 2017, pp. 670–680.

Dagan I., Roth D., Sammons M. and Zanzotto F.M. (2013). *Recognizing Textual Entailment: Models and Applications*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, pp. 1–220.

Dasgupta D. and Forrest S. (1996). Novelty detection in time series data using ideas from immunology. In *Proceedings of the International Conference on Intelligent Systems*, pp. 82–87.

Dasgupta T. and Dey L. (2016). Automatic scoring for innovativeness of textual ideas. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.

Fleiss J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin* **76**(5), 378.

Franz M., Ittycheriah A., McCarley J.S. and Ward T. (2001). First story detection: Combining similarity and novelty based approaches. In *Topic Detection and Tracking Workshop Report*, pp. 193–206.

Gabrilovich E., Dumais S. and Horvitz E. (2004). Newsjunkie: Providing personalized newsfeeds via analysis of information novelty. In *Proceedings of the 13th International Conference on World Wide Web*. ACM, pp. 482–490.

Gamon M. (2006). Graph-based text representation for novelty detection. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*. Association for Computational Linguistics, pp. 17–24.

Ghosal T., Edithal V., Ekbal A., Bhattacharyya P., Tsatsaronis G. and Chivukula S.S.S.K. (2018a). Novelty goes deep. A deep neural solution to document level novelty detection. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20–26,* 2018, pp. 2802–2813.

Ghosal T., Salam A., Tiwary S., Ekbal A. and Bhattacharyya P. (2018b). TAP-DLND 1.0 : A corpus for document level novelty detection. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7–12, 2018*.

Guh R.-S., Zorriassatine F., Tannock J. and O'Brien C. (1999). On-line control chart pattern detection and discriminationa neural network approach. *Artificial Intelligence in Engineering* **13**(4), 413–425.

Harman D. (2002). Overview of the TREC 2002 novelty track. In *Proceedings of The Eleventh Text REtrieval Conference, TREC 2002, Gaithersburg, Maryland, USA, November 19–22, 2002*.

Karkali M., Rousseau F., Ntoulas A. and Vazirgiannis M. (2013). Efficient online novelty detection in news streams. In *WISE (1)*, pp. 57–71.

King S., King D., Astley K., Tarassenko L., Hayton P. and Utete S. (2002). The use of novelty detection techniques for monitoring high-integrity plant. In *Proceedings of the International Conference on Control Applications*, vol. 1. IEEE, pp. 221–226.

Kwee A.T., Tsai F.S. and Tang W. (2009). Sentence-level novelty detection in english and malay. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, pp. 40–51.

Lai A., Bisk Y. and Hockenmaier J. (2017). Natural language inference from multiple premises. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27–December 1, 2017 - Volume 1: Long Papers,* pp. 100–109.

Le Q. and Mikolov T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1188–1196.

https://doi.org/10.1017/S1351324920000194 Published online by Cambridge University Press

Lee S. (2015). Online sentence novelty scoring for topical document streams. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 567–572.

Li X. and Croft W.B. (2005). Novelty detection based on sentence level patterns. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*. ACM, pp. 744–751.

Lin Z., Feng M., Santos C.N.d., Yu M., Xiang B., Zhou B. and Bengio Y. (2017). A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130.

Liu Y., Sun C., Lin L. and Wang X. (2016). Learning natural language inference using bidirectional LSTM model and inner-attention. CoRR abs/1605.09090.

Manikopoulos C. and Papavassiliou S. (2002). Network intrusion and fault detection: A statistical anomaly approach. *IEEE Communications Magazine* **40**(10), 76–82.

Mishra S. and Torvik V.I. (2016). Quantifying conceptual novelty in the biomedical literature. *D-Lib Magazine* **22**(9/10).

Mou L., Men R., Li G., Xu Y., Zhang L., Yan R. and Jin Z. (2015). Recognizing entailment and contradiction by tree-based convolution. CoRR abs/1512.08422.

Parikh A.P., Täckström O., Das D. and Uszkoreit J. (2016). A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1–4,* 2016, pp. 2249–2255.

Ru L., Zhao L., Zhang M. and Ma S. (2004). Improved feature selection and redundance computing-thuir at trec 2004 novelty track. In *TREC*.

Schiffman B. and McKeown K.R. (2005). Context and learning in novelty detection. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 716–723.

Soboroff I. (2004). Overview of the TREC 2004 novelty track. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16–19, 2004.*

Soboroff I. and Harman D. (2003). Overview of the TREC 2003 novelty track. In *Proceedings of The Twelfth Text REtrieval Conference, TREC* 2003, *Gaithersburg, Maryland, USA, November 18–21, 2003,* pp. 38–53.

Soboroff I. and Harman D. (2005). Novelty detection: The trec experience. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 105–112.

Stokes N. and Carthy J. (2001). First story detection using a composite document representation. In *Proceedings of the First International Conference on Human Language Technology Research*. Association for Computational Linguistics, pp. 1–8.

Tang W., Tsai F.S. and Chen L. (2010). Blended metrics for novel sentence mining. *Expert Systems with Applications* **37**(7), 5172–5177.

Tarassenko L., Hayton P., Cerneaz N. and Brady M. (1995). Novelty detection for the identification of masses in mammograms. In *1995 Fourth International Conference on Artificial Neural Networks*, Cambridge, UK, pp. 442–447.

Tax D.M. and Duin R.P. (1998). Outlier detection using classifier instability. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*. Springer, pp. 593–601.

Tsai F.S., Tang W. and Chan K.L. (2010). Evaluation of novelty metrics for sentence-level novelty mining. *Information Sciences* **180**(12), 2359–2374.

Tsai F.S. and Zhang Y. (2011). D2s: Document-to-sentence framework for novelty detection. *Knowledge and Information Systems* **29**(2), 419–433.

Verheij A., Kleijn A., Frasincar F. and Hogenboom F. (2012). A comparison study for novelty control mechanisms applied to web news stories. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT),* vol. 1. IEEE, pp. 431–436.

Wayne C.L. (1997). Topic detection and tracking (tdt). In *Workshop Held at the University of Maryland on,* vol. 27. Citeseer, pp. 28.

Yang Y., Pierce T. and Carbonell J. (1998). A study of retrospective and on-line event detection. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 28–36.

Yang Y., Zhang J., Carbonell J. and Jin C. (2002). Topic-conditioned novelty detection. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 688–693. ACM.

Zhang M., Song R., Lin C., Ma S., Jiang Z., Jin Y., Liu Y., Zhao L. and Ma S. (2003). Expansion-based technologies in finding relevant and new information: Thu TREC 2002: Novelty track experiments. In *NIST SPECIAL PUBLICATION SP* (251), 586–590.

Zhang Y., Callan J. and Minka T. (2002). Novelty and redundancy detection in adaptive filtering. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 81–88.

Zhang Y. and Tsai F.S. (2009). Combining named entities and tags for novel sentence detection. In *Proceedings of the WSDM'09 Workshop on Exploiting Semantic Annotations in Information Retrieval*. ACM, pp. 30–34.

Zhao P. and Lee D.L. (2016). How much novelty is relevant? it depends on your curiosity. In *39th International ACM SIGIR Conference on Research and Development, Pisa, Italy,* pp. 100.

A. Appendix

This section consists of the full-text examples of error instances in Figure 3(c), (d), (e). Kindly refer to Section 5.4 in the paper for more details. These documents are from the TAP-DLND 1.0 dataset.

Instance #45

Concatenated Source

0: Raebareli (Uttar Pradesh) [India], November 1 (ANI): At least 10 people were killed and 70 others sustained burn injuries after an ash-pipe exploded due to pressure at National Thermal Power Corporation (NTPC) plant in Unchahar area of Uttar Pradesh's Raebareli district.

1: Speaking to ANI, Additional Director General (Law and Order) Anand Kumar said, "As of now ten deaths have been confirmed by the district administration, while about 70 people have sustained burn injuries."

2: Following the incident, Uttar Pradesh Chief Minister Yogi Adityanath has announced an ex-gratia of Rs 2 lakh for the kin of deceased, Rs 50,000 for critically injured and Rs 25,000 for injured.

3: A 32-member team of National Disaster Response Force (NDRF) has left for Unchahar in Raebareli.

4: The authorities of NTPC said, "Rescue operations are underway in close coordination with District Administration.

5: Injured people have been shifted to nearby district hospitals.

6: An unfortunate accident in the boiler of 500 MW under trial unit of NTPC in Unchahar occurred this afternoon."

7: "NTPC's senior management is rushing to the site to coordinate the efforts," NTPC Corporation Communication Department said.

8: Union Health Minister J P Nadda spoke to UP Health Minister and Union Health Secretary to extend all possible help.

9: The blast reportedly took place when a boiler tube exploded at the unit.

10: "Ash-pipe exploded due to pressure at NTPC plant in Rae Bareli," the District Magistrate informed.

11: NTPC operates a 1550 MW power plant of Uttar Pradesh, which is named after Feroze Gandhi, the husband of former Prime Minister Indira Gandhi.

12: (ANI).

13: Englishmate Rahul Gandhi demands judicial probe into NTPC blast, meets victims 29 people died and over 60 injured in the boiler blast at the state-run power giant NTPCs Unchahar plant india Updated: November 02, 2017 19:15 IST Kenneth John Hindustan Times, Rae Bareli

(Unchahar) Congress vice-president Rahul Gandhi arrives to meet the family members of the victims of Unchahar NTPC boiler blast in Raebareli on Thursday.

14: (PTI Photo) Congress vice president Rahul Gandhi on Thursday demanded a judicial probe into the boiler blast at the state-run power giant NTPCs Unchahar plant in which 29 people died and over 60 injured.

15: He made the demand after visiting the district hospital in Rae Bareli, where the injured were undergoing treatment.

16: The accident showed serious lapse in working of the unit and to know the reason behind it a judicial probe was needed, Gandhi said during a brief interaction with media at the hospital.

17: Congress vice-president rushed to Rae Bareli taking a break from his ongoing Navsarjan Yatra in poll-bound Gujarat.

18: Rae Bareli is the parliamentary constituency of his mother and Congress president Sonia Gandhi.

19: To ascertain the cause of the blast, the NTPC has initiated a probe amid allegations from labourers that they had warned about the possible disaster at the ill-fated unit-six as the temperature near the furnace was steadily rising.

20: Gandhi consoled family members of victims who lost their lives in the blast and enquired about the condition of those admitted in the hospital.

21: Later, the Amethi member of parliament also visited private hospitals, SIMHANS and Nirmal, and enquired about condition of blast victims admitted there.

22: He also visited the site of boiler blast on NTPC premises and enquired about the accident from officials.

23: Senior Congress leader Ghulam Nabi Azad and UP Congress chief Raj Babbar accompanied the Congress vice president.

24: The laxities in construction should be probed thoroughly as without a major lapse accident of such magnitude was not possible, said Babbar.

25: The 500 MW unit 6 of the power plant was commissioned in April but due to technical fault in the boiler it failed to produce power.

26: We demand a judicial probe into the tragedy.

27: Setting up of an enquiry committee by government is just eyewash, reiterated Azad.

28: During his visit to the blast site Gandhi came face to face with union power minister RK Singh, who also visited the plant and took stock of the situation.

29: Singh denied any human negligence led to the blast.

---

https://doi.org/10.1017/S1351324920000194 Published online by Cambridge University Press

30: Army pays tribute to 2 soldiers killed in Pulwama encounter November 03, 2017 16:55 IST.

31: NTPC shuts Unchahar plant unit after 26 die in blast; warnings ignored?

32: No human negligence behind boiler blast: Power Minister Agency Report — New Delhi/Rae Bareli — 2 November, 2017 — 11:30 PM Power producer NTPC has shut down a 500 MW unit at its Feroze Gandhi Unchahar Thermal Power Station in Rae Bareli in Uttar Pradesh following the accident on Wednesday that claimed the lives of 30 people.

33: Share this: Print The company in a regulatory filing on the BSE said: This is to inform that Unit 6 (500 MW) of Feroze Gandhi Unchahar Thermal Power Station, Rae Bareli, is under shut down after an accident in the evening of November 1, 2017.

34: The other five units of the station are operating normally. The death toll in the NTPC boiler blast here in Uttar Pradesh rose to 26 on Thursday, with more injured workers succumbing to their burn injuries, officials said.

35: The massive explosion took place in a 500 MW boiler unit in Unchahar town on the Lucknow-Allahabad highway.

36: Many were trapped when a fire erupted in the boiler and a huge ball of dust rose after the blast, making the rescue operations difficult.

37: On Thursday, contractual labourers at the plant raised slogans against the NTPC management.

38: They claimed they had warned about an impending disaster at unit six as the temperature near the furnace had been steadily rising.

39: The NTPC has launched a probe into the incident, which it said took place due to excess ash deposition in the furnace.

40: The state government has ordered a magisterial probe.

41: Union Power Minister R.K. Singh on Thursday denied claims by some political leaders and families of the deceased that human negligence was to blame for the boiler blast in NTPCs Unchahar unit here that left 30 dead and dozens seriously injured.

42: I have seen everything during my physical inspection of the accident scene and I can say that there is no human negligence in the unfortunate incident, Singh told reporters, after visiting the accident site along with state Power Minister Shrikant Sharma.

43: He also announced that the Central government has decided to give financial assistance of Rs 20 lakh to the families of the deceased and Rs 10 lakh each to the critically injured, while those who sustained minor injuries would get Rs 2 lakh each, the Union Minister announced.

44: This compensation would be in addition to the ex-gratia and financial assistance announced by Uttar Pradesh Chief Minister Yogi Adityanath.

45: The state-run National Thermal Power Corp (NTPC) has also announced a financial assistance of Rs 5 lakh each to the families of the dead.

---

https://doi.org/10.1017/S1351324920000194 Published online by Cambridge University Press

46: The Prime Ministers Relief Fund will also give Rs 2 lakh each to the next of kin of the deceased.

47: R.K. Singh also said that the priority of the government, both at the Centre and the state, was to save as many lives as possible, provide the best, prompt and adequate treatment.

48: Singh also said that NTPCs Unchahar unit was among the best in the country and that rumours that there was an extra load on it or that it was under pressure to increase production were unfounded and baseless.

49: How and why the accident happened would be conclusively found and detailed in the probe ordered by the Ministry which would be completed in 30 days, he added.

50: State Deputy Chief Minister Dinesh Sharma, who also visited Rae Bareli on Thursday, urged the opposition parties not to make political currency out of the tragic incident.

51: Both the state and Union governments are saddened by the tragedy and are doing all they can to bring relief to the affected, he said and added that Prime Minister Narendra Modi was personally very sad at the loss of lives in the accident.

52: (IANS) Share this:

Target Document

0: The death toll from a blast at a coal-fired power plant in northern India rose to at least 29 on Thursday, as authorities launched an investigation into the cause of one of the countrys worst industrial accidents in years.

1: More than 20 survivors were battling for their lives with severe burns following Wednesdays blast in a newly-operated unit at the 1550 MW plant run by state-owned NTPC, officials in Uttar Pradesh state said.

2: More than 80 others suffered injuries in the explosion.

3: Arvind Kumar, a principal secretary, said some of the severely injured had been taken to a hospital in the state capital Lucknow.

4: Blockages in the flue gas pipe in a unit led to the blast.

5: Hot flue gases and steam let out by the blast severely injured several workers

---

6: Sanjay Kumar Khatri, the top government official of Rae Bareli district where the plant is located, told Reuters on Thursday.

7: A magisterial inquiry has been initiated.

8: This two-member technical team will submit findings within seven days, Khatri said.

9: In a statement, the National Human Rights Commission said an investigation was needed to ascertain whether negligence or errors had caused the explosion, and asked the state government to submit a detailed report within six weeks.

---

https://doi.org/10.1017/S1351324920000194 Published online by Cambridge University Press

10: NTPC is the countrys top power producer and accidents have been rare at its facilities.

11: Senior state police official Anand Kumar said on Wednesday ash had piled up in the furnace beneath the boiler, which led to a build-up of pressure resulting in the explosion.

12: The power ministry and state government have both offered cash compensation to the families of the deceased and to the injured.

13: The plant in the town of Unchahar supplies electricity to nine states, NTPCs website showed.

14: The company said other facilities would make up for the shortfall and outages were unlikely.

15: The 500 MW unit had been operating since April and was shut down after the accident.

16: The other five units of the station are operating normally, NTPC said in a statement.

17: NTPC has initiated an inquiry into the incident.

18: We are not a company that will take any risk.

19: We have so many units that if power cannot be supplied by one, it can be given by the other.

20: It was a sudden accident, an NTPC official, who did not wish to be named, said.

Instance #381

Concatenated Source

0: As many as 42 clerics have issued a fatwa against reality singing star Nahid Afrin, who was the first runner-up of a musical reality TV show, asking her to stop performing in public.

1: According to reports, the fatwa has been issued against her keeping in view an upcoming event that was to be held on March 25.

2: Since the venue of the event was in the vicinity of a mosque and a graveyard, the clerics have announced to boycott her singing programmes and have asked her to stop performing in public.

3: The young singer was shocked when she first heard of the fatwa against her.

4: I was shocked and broken from inside at first.

5: But, many Muslim singers gave me inspiration to not quit music.

6: I will never do so, she said.

7: I think my music is Gods gift to me.

8: I believe it must be properly utilised; not doing so is ignoring God, she added.

9: Nahid, who made her Bollywood singing debut for actress Sonakshi Sinha in the 2016 movie Akira, first rose to stardom after a successful stint on a reality singing show in the year 2015.

10: Police are investigating whether the fatwa was a reaction to Nahid recently performing songs against terrorism Leaflets bearing the fatwa in Assamese and the names of the clerics were distributed across Hojai and Nagaon districts of Indian state of Assam.

11: Forty-six Muslim clerics in Assam have issued a fatwa against up-and-coming singer Nahid Afrin, who was the first runner-up in the 2015 season of a musical reality TV show, asking her to stop performing in public.

12: Police said they were investigating whether the fatwa was a reaction to Nahid recently performing songs against terrorism, including the Islamic State terror group.

13: “We are looking at this angle as well,” ADG (special branch) Pallab Bhattacharya said.

14: Leaflets bearing the fatwa in Assamese and the names of the clerics were distributed across Hojai and Nagaon districts in central Assam on Tuesday.

15: According to the fatwa, a March 25 programme at Udali Sonai Bibi College in Lanka, Assam, where Nahid, 16, is scheduled to perform is “against the Sharia”.

16: “If anti-Sharia acts like musical nights are held on grounds surrounded by masjids, idgahs, madrassas and graveyards, our future generations will attract the wrath of Allah,” it said.

17: The young singer, a Class X student who lives in Biswanath Chariali, broke down on hearing news of the fatwa.

18: “I am speechless.

19: I think my music is God’s gift to me.

20: I will never bow down to it (such warnings) and never leave singing,” she said.

21: Her mother added, “The organisers of the musical night told us that the programme on March 25 will not be cancelled.”

22: Police said Nahid and her family would be provided security cover.

23: Nahid, who made her Bollywood debut singing for Sonakshi Sinha in the 2016 film “Akira”, first rose to stardom after her successful innings on reality TV.

24: Her beautiful renditions of songs written and composed by the Vaishnavite saint Srimanta Sankardeva have made her especially popular in Assam.

25:.

26: 10,656 For those that are wondering why: According to reports, the fatwa has been issued against her keeping in view an upcoming event that was to be held on March 25.

27: Since the venue of the event was in the vicinity of a mosque and a graveyard, the clerics have announced to boycott her singing programmes and have asked her to stop performing in public.

---

https://doi.org/10.1017/S1351324920000194 Published online by Cambridge University Press

28: If anti-Sharia acts like musical nights are held on grounds surrounded by masjids, idgahs, madrassas and graveyards, our future generations will attract the wrath of Allah, the fatwa read.

29: The wrath of Allah to be with you.

30: Theres so many different worlds So many differents suns And we have just one world But we live in different ones

Target Document

0: Nahid Afrin, who was the first runner-up of a singing talent-hunt show Indian Idol Junior, have banned from doing stage shows.

1: As many as 46 Muslim clerics in Assam have issued a fatwa against the 16-year-old girl asking her to stop performing in public citing Sharia laws.

2: According to reports, the fatwa has been issued in view of an upcoming event to be held on March 25, the venue of which is in the vicinity of a mosque and a graveyard.

3: The clerics have announced to boycott her singing programmes and have asked her to stop performing in public.

4: Leaflets bearing the fatwa in Assamese and the names of the clerics were distributed across Hojai and Nagaon districts in central Assam on Tuesday.

5: "If anti-Sharia acts like musical nights are held on grounds surrounded by masjids, idgahs, madrassas and graveyards, our future generations will attract the wrath of Allah," it said, according to a report in The Times of India.

6: Police said they were investigating whether the fatwa was a reaction to Nahid recently performing songs against terrorism, including the Islamic State terror group.

7: "We are looking at this angle as well," the report quoted ADG (special branch) Pallab Bhattacharya as saying.

8: Nahid, who made her Bollywood playback singing debut with the song Rajj Rajj Ke from Sonakshi Sinha starrer Akira, said she is not afraid of the threat and will continue to pursue her dream of making it big in the music industry.

9: "I was shocked and broken from inside at first, but many Muslim singers gave me inspiration to not quit music, will never do so.

10: My singing is gift of God.

11: I believe it must be properly utilised, not doing so is ignoring God," Nahid said when asked about the fatwa issued against her.

12: Assam Chief Minister Sarbananda Sonowal has assured the young singer of proper security cover, following the fatwa and threat reports.

13: Other prominent persons of the state too have come out in support of Nahid.

https://doi.org/10.1017/S1351324920000194 Published online by Cambridge University Press

Instance #495

Concatenated Source

0: New Delhi, October 23 (IANS) Condemning the police firing on a tribal protest in Jharkhand in which a man was killed, the CPI-M on Sunday demanded the Centres intervention to protect tribals rights, including their right to protest.

1: The man was killed and three others were injured in Khunti district of the state on Saturday when the police opened fire on tribals protesting the Raghubar Das-led Bharatiya Janata Party governments decision to amend two state legislations on lands, which would have significant impact on the tribals.

2: According to reports, the police fired around a hundred rounds to disperse the gathering, resulting in the death of one tribal leader, Abraham Munda, and critical injuries to five others, the Communist Party of India-Marxist said in a statement here.

3: The BJP state government has during the last one month resorted to repeated police firing against popular protests.

4: There have also been two custodial deaths during this period, the party said in the statement.

5: Jharkhand is a state carved out in the name of protecting the interests of the tribals.

6: The BJP government on the contrary is continuously threatening the tribals and other weaker sections of Jharkhand to negate the very purpose upon which the state was formed, said the party.

7: The killing of tribals, according to established parliamentary practice, is not considered merely as a state law and order subject.

8: The central government must intervene to ensure that the rights of the tribals are protected, including their democratic right to protest, and no further encroachment on their rights to land and forest produce must be permitted, the party added.

9:.

10: Ranchi, October 23: The Catholic Bishops Conference of India (CBCI) today said it deplored the recent "unconstitutional" and "divisive" statements made by chief minister Raghubar Das where he controversially asked villagers in Dumka gram sabhas to protect their identity without falling prey to allurements, a barb apparently directed at missionaries.

11: The CBCI is the highest policy making body of the Roman Catholic Church in India.

12: "The Catholic Bishops Conference of India is saddened by the unconstitutional and divisive statements being made by the chief minister of Jharkhand Mr Raghubar Das against Christian tribals and Christian leaders. . .we strongly deplore and condemn the comments made by Mr Raghubar Das against Christians," a media communiqu signed by Bishop Theodore Mascarenhas, CBCI secretary general in New Delhi, said.

13: The Catholic church reiterated its commitment to nation building, empowerment of tribals, Dalits and the poor, he said.

14: “However, it will not accept any infringement on the rights of the minorities and the constitutionally guaranteed freedom of conscience and the right freely to profess, practice and propagate religion,” the strongly worded media release stressed.

15: “The just aspirations of the tribals are being threatened by amendments to the Chhotanagpur Tenancy Act and the Santhal Pargana Tenancy Act.

16: We request the chief minister to protect the rights of the tribals. . .” Mascarenhas went on to add, apparently referring to reports in the media where Das is learnt to have allegedly claimed those involved in “tribal conversions” were leading the protest against the state’s proposed amendments to the Acts.

17: The CBCI secretary general pointed out that according to official data, 35 lakh out of 69 families in Jharkhand lived below the poverty line and that the state fell far short of the national average in almost all development indicators.

18: “The chief minister should refrain from using divisive tactics and interfering in the legitimate rights given to us under the Constitution and devote himself to serious issues like malnutrition, illiteracy, unemployment,” Mascarenhas added.

19: In a news meet in Ranchi today, senior Congress leader and former Union minister Subodh Kant Sahay also said the Raghubar Das government was trying to create a rift in tribal society on the name of Christian and Sarna religions.

20: “But we won’t allow them to do it.

21: We will expose their ill-intent,” Sahay said.

22:.

23: RANCHI: Chief minister Raghubar Das on Sunday came under fire from tribal Christians of the state for his recent remarks on conversion of tribals. The Rashtriya Isai Mahasangh (RIM), a socio-religious body of the Christian community spread across several states, on Sunday accused Das of demeaning the Christian missionaries and their ongoing social service going on in the state for decades.

24: “Do they (the government) want us to stop serving the society?

25: Relating social service that we do for conversion of religion is not only shameful, but also an effort to polarize the tribals in the name of Christians and Sarna,” said Prabhakar Tirkey, national general secretary of RIM at a congregation in Ranchi’s Xavier’s Institute of Social Service.

26: Tirkey also accused Das of trying to create rift between tribals following Sarna code and those who follow Christianity. Last week, Das became vocal on the religious conversion of tribals.

27: During his Gram Sabha meetings in Dumka and Pakur and later at BJP’s state working committee meeting in Ranchi, Das said those involved in converting the tribals under various temptations.

28: Approximately 4.5

---

https://doi.org/10.1017/S1351324920000194 Published online by Cambridge University Press

29: “We stand united against all divisive forces.

30: We hope that all of you will support us and stand united against such forces,” said Deepak Tirkey, a senior office bearer of RIM. Das’s comment has drawn sharp criticism from politicians from opposition.

31: Vandana Dadel, panchayati raj secretary and a 1996 batch IAS officer of Christian faith in the state government, also indirectly objected to the CM’s statement in her social media account. The meeting was also attended by former Union minister and Congress leader Subodh Kant Sahay

Instance #495

Target Document

0: RANCHI: Jharkhand chief minister Raghubar Das on Sunday came under fire from the tribal Christians of the state for his recent remarks on conversion of tribals. The Rashtriya Isai Mahasangh (RIM), a socio-religious body of the Christian community spread across several states, on Sunday accused Das on demeaning the Christian missionaries and their ongoing social service going on in the state for decades.

1: “Do they (the government) want us to stop serving the society?

2: Relating social service that we do to conversion of religion is not only shameful, but also an effort to polarize the tribals in the name of Christians and Sarna,” said Prabhakar Tirkey, national general secretary of RIM at a congregation in Ranchi’s Xavier’s Institute of Social Service.

3: Tirkey also accused Das of trying to create rift between tribal following Sarna code and those who follow Christianity. Over the past week Das became vocal on the religious conversion of tribals.

4: Approximately 4.5

5: “We stand united against all divisive forces.

6: We hope that all of you will support us and stand united against such forces,” said Deepak Tirkey, a senior office bearer of RIM. Das’s comment has drawn sharp criticism from politicians from opposition.

7: Vandana Dadel, panchayati raj secretary and a 1996 batch IAS officer of Christian faith in Jharkhand government, also indirectly objected to the CM’s statement in her social media account.

Cite this article: Ghosal T, Edithal V, Ekbal A, Bhattacharyya P, Chivukula SSSK and Tsatsaronis G (2021). Is your document novel? Let attention guide you. An attention-based model for document-level novelty detection. *Natural Language Engineering* **27**, 427–454. https://doi.org/10.1017/S1351324920000194