Novelty Goes Deep. A Deep Neural Solution To Document Level Novelty Detection

Tirthankar Ghosal†, Vignesh Edithal†, Asif Ekbal†, Pushpak Bhattacharyya†,
George Tsatsaronis‡, Srinivasa Satya Sameer Kumar Chivukula‡
†Indian Institute of Technology Patna, Bihar, India
‡Elsevier, Amsterdam, Netherlands
†(tirthankar.pcs16, edithal.cs14, asif, pb)@iitp.ac.in
‡(g.tsatsaronis, s.chivukula)@elsevier.com

Abstract

The rapid growth of documents across the web has necessitated finding means of discarding redundant documents and retaining novel ones. Capturing redundancy is challenging as it may involve investigating at a deep semantic level. Techniques for detecting such semantic redundancy at the document level are scarce. In this work we propose a deep Convolutional Neural Network (CNN) based model to classify a document as novel or redundant with respect to a set of relevant documents already seen by the system. The system is simple and does not require manual feature engineering. Our novel scheme encodes relevant and relative information from both source and target texts to generate an intermediate representation for which we coin the name Relative Document Vector (RDV). The proposed method outperforms the existing benchmark on two document-level novelty detection datasets by a margin of $\sim$5% in terms of accuracy. We further demonstrate the effectiveness of our approach on a standard paraphrase detection dataset where the paraphrased passages closely resembles semantically redundant documents.

1 Introduction

Document-level novelty detection implies the categorization of a document as *novel* if it contains sufficient *new* information with respect to whatever *relevant* is previously known or seen. Existing literature on novelty detection from texts mostly followed the detection of novelty at the sentence level, i.e., to extract the sentences that carry new information with respect to some relevant source texts (c.f. Section 2). *The current work is aimed at automatically classifying an incoming document as* ***novel*** *or* ***non-novel*** *on the basis of documents already seen by the system*. We view the problem as a two-class classification problem in machine learning 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. The source document set could be seen as the memory of the reader which stores known information. Document-level novelty detection is a crucial problem and finds application in diverse domains of language processing such as information retrieval, document summarization, predicting impact of scholarly articles, etc. It is a complex problem that comprehends *lexical, syntactic, semantic and pragmatic* levels of texts in conjunction with certain characteristics like *relevance, diversity, relativity, and temporality* (Ghosal et al., 2018). Literature methods for novelty detection based on lexical similarity, divergence features, and information retrieval measures, although proved effective for sentence level but could not address the document-level comprehension needs. Our understanding of the problem led us to an assertion that a deep neural network might be able to extract the text subtleties involved in understanding a document's *novelty*. To the best of our knowledge this is the very first attempt to address document-level novelty detection without the involvement of any hand-crafted features. Our approach achieves significant performance improvement over the existing measures on novelty detection from texts on three different datasets, thus establishing our contention.

Proceedings of the 27th International Conference on Computational Linguistics, pages 2802–2813
Santa Fe, New Mexico, USA, August 20-26, 2018.

1.1 Motivation

Motivation behind the current work stems from *the eternal quest of the inquisitive mind* : **new information**.

- The exponential dump of redundant information in documents across the web, which does not add anything new to the knowledge, yet consumes valuable space and time (of readers) accentuates identifying novel documents and discarding non-novel documents in the current context.
- Plagiarism detection in scholarly articles at the semantic level is another ambitious vision that demands research attention. Methods for document-level novelty detection can aid in the investigation.
- Our survey reveals that in spite of having potential in various applications, document-level novelty detection has not garnered deserved attention from the community. Also it was intriguing to investigate how deep neural architectures could distinguish between novel and non-novel texts.

1.2 Contributions

The principal offerings of the present work could be outlined as :

1. *Proposing an effective deep learning strategy for document-level novelty detection: a first of its kind*
2. *Presenting an effective semantic vector representation to model information from both source and target documents.*

2 Related Works

Research in novelty mining could be traced back to the Topic Detection and Tracking (TDT) (Wayne, 1997) evaluation campaigns where the concern was to detect new events or First Story Detection (FSD) with respect to online news streams. 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. Some notable contributions from TDT are (Allan et al., 1998; Yang et al., 2002; Allan et al., 2000; Yang et al., 1998).

The task gained prominence in the novelty tracks of Text Retrieval Conferences (TREC) from 2002 to 2004 (Soboroff and Harman, 2005; Harman, 2002; Soboroff and Harman, 2003; Clarke et al., 2004) although the focus was sentence-level novelty detection. The goal of these tracks was to highlight the relevant sentences that contain novel information, given a topic and an ordered list of relevant documents. Some interesting works in TREC were based on the sets of terms (Zhang et al., 2003a; Zhang et al., 2003b), term translations (Collins-Thompson et al., 2002), Principal Component Analysis (PCA) vectors (Ru et al., 2004), Support Vector Machine (SVM) classification (Tomiyama et al., 2004) etc. Similar works relied on named entities (Gabrilovich et al., 2004; Li and Croft, 2005; Zhang and Tsai, 2009), language models (Zhang et al., 2002; Allan et al., 2003), contexts (Schiffman and McKeown, 2005), etc. Next came the novelty sub tracks of Recognizing Textual Entailment-Text Analytics Conference (RTE-TAC) 6 and 7 (Bentivogli et al., 2011) where Textual Entailment was viewed as one close neighbor to sentence level novelty detection.

At the document level an interesting work was carried out by (Yang et al., 2002) via topical classification of on-line document streams and then detecting novelty of documents in each topic exploiting the named entities. 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 in order 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 innovativeness of a document. Each of these works evaluated their methods on separate datasets and to the best of our knowledge except APWSJ none of these is publicly available.

Novelty detection is also studied in works related to diversity in information retrieval literature. Idea is to retrieve relevant yet diverse documents in response to user query. The work on Maximal Marginal Relevance (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; Clarke et al., 2008).

The work that we present here significantly differs from the existing literature as we provide a deep neural network solution to the problem which does not depend on hand crafted features, and learns the notion of novelty and non-novelty from the data itself.

3 Document Novelty : Challenges and Observations

Deciding the novelty of an entire document or what amount of new information makes a document appear novel are very challenging tasks. Even more challenging since the task inherently exhibits several text characteristics (*relevance, diversity, relativity, temporality*) to be addressed simultaneously. A novel document should be *relevant* to context, should exhibit *diversity* in information content, should have *relatively* more new relevant yet diverse information and is usually deemed as a *temporal* update over the existing knowledge. The difficulty associated with the task is further exaggerated when there are a large number of history documents against which a target document is to be judged for novelty. We propose a possible solution that could address most of these challenges and also reduce the need of manual feature engineering to an extent. We analyze the available benchmark sentence-level novelty detection datasets like TREC/RTE-TAC and come up with certain observations :
(1) Named Entities (NEs) play a vital role in adjudging the *relevance* of a test document against its source document(s).
(2) Semantic features play a crucial role in identifying *redundancy* or *non-novelty*. Lexical level features (such as *n-gram* match) do well when there is a word-by-word match or when two texts share a fair amount of tokens. But it fails in cases where surface form is very different yet conveys the same meaning (*paraphrases*).
(3) Lexical features are sometimes efficient to detect document level new information content. For e.g., if the source document(s) consists of information regarding the facts of a certain accident and the target document reports the police investigation carried out afterwards-although the two documents would share a fair amount of NEs, yet would differ in contextual information content which could be captured to a large extent via lexical level features.

4 Dataset Description

Since our focus is on document-level novelty detection we consider only document level datasets. Existing benchmark datasets on sentence-level novelty detection are not suited to our experimental framework. We also evaluate our approach on the standard Webis Crowd Paraphrase Corpus (Webis CPC) (Burrows et al., 2013) whose inherent semantic richness sets a competitive benchmark for our method.

4.1 APWSJ Corpus

The APWSJ data consists of news articles from Associated Press (AP) and Wall Street Journal (WSJ) corpus covering the same time period (1988 to 1990) and many on the same topics, guaranteeing some redundancy in the document stream. Similar to (Zhang et al., 2002) we use the documents within the designated 33 topics[1] with redundancy judgments by the assessors. The dataset was meant to filter superfluous documents in a retrieval scenario and deliver only the documents having *redundancy score* below a calculated threshold. Documents for each topic were delivered in a chronological manner and

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

the assessors provided two degrees of judgments on the non-novel documents: *absolute redundant* or *somewhat redundant* based on prior documents. The unmarked documents were treated as *novel.*

4.2 Webis CPC

The Webis Crowd Paraphrase Corpus 2011 (Webis-CPC-11) contains 7,859 candidate paraphrases obtained from the Mechanical Turk crowdsourcing. The corpus[2] is made up of 4,067 accepted paraphrases, 3,792 rejected non-paraphrases, and the original texts. For our experiment we assume the original texts as the source document and the corresponding candidate paraphrase/non-paraphrase as the target document. We hypothesize that a paraphrased document would not contain any new information and hence could be treated as *non-novel* whereas a non-paraphrased candidate could be taken as *novel*. Basically we seek to investigate the viability of our proposed method to discover the semantically redundant documents (paraphrases) from this dataset.

4.3 TAP-DLND 1.0

We experiment with this recent benchmark resource for document-level novelty detection (DLND) (Ghosal et al., 2018). The dataset[3] is balanced and consists of 2,736 novel documents and 2,704 non-novel documents. For each novel/non-novel document there are three source documents against which the target document is annotated. The state of *novelty* for each target document is to be measured against those source documents i.e. 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. The semantic nature of the dataset makes it an ideal candidate for our experiments.

5 Proposed Model

Having an acceptable benchmark dataset avaialable at our end, instead of handcrafting the similarity/divergence based features,we try to learn feature representations from a target document with respect to the source document(s) using a Convolutional Neural Network (CNN). Our proposed model is based on a recent sentence embedding paradigm proposed by (Conneau et al., 2017). We leverage their idea and create a representation of the *relevant* target document *relative* to the designated source document(s) and call it as the *Relative Document Vector (RDV)*[4]. We then train a Convolutional Neural Network (CNN) with the RDV of the target documents in the similar line of (Kim, 2014), and finally classify a document as *novel* or *non-novel* with respect to its source documents (Figure 1). Although there are document embedding models available, our method is specifically tailored to address the *relativity* and *diversity* criteria which is fundamental to the definition of *novelty*. Here T_1 is the *target* document whose state of *novelty* is to be determined against the source document(s) $S_1, S_2, ...S_M$ i.e. to say the objective is to automatically figure out whether T_1 is *novel* or not once the machine has already seen/scanned $S_1, S_2, ...S_M$. Our model assumes that the documents are relevant to a context. We reserve *Temporality* to be explored in a later work.

5.1 Embedding and Sentence Encoder

The task of *Novelty Detection* requires high-level understanding and reasoning about semantic relationships within texts. Textual Entailment or Natural Language Inference is one such task which exhibits such complex semantic interactions. Following from (Conneau et al., 2017) we therefore employ a sentence encoder based on a bi-directional Long Short Term Memory (LSTM) architecture with max pooling, trained on the large-scale Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015). SNLI entries supposedly captures rich semantic associations between the text pairs (entailment/inference relationships between premise and hypothesis). The output of our sentence encoder is

[2]https://www.uni-weimar.de/en/media/chairs/computer-science-department/webis/data/corpus

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

[4]We call our document vector *relative*, as we desire to encode the relative *new* information of a target document w.r.t. its relevant source document(s)

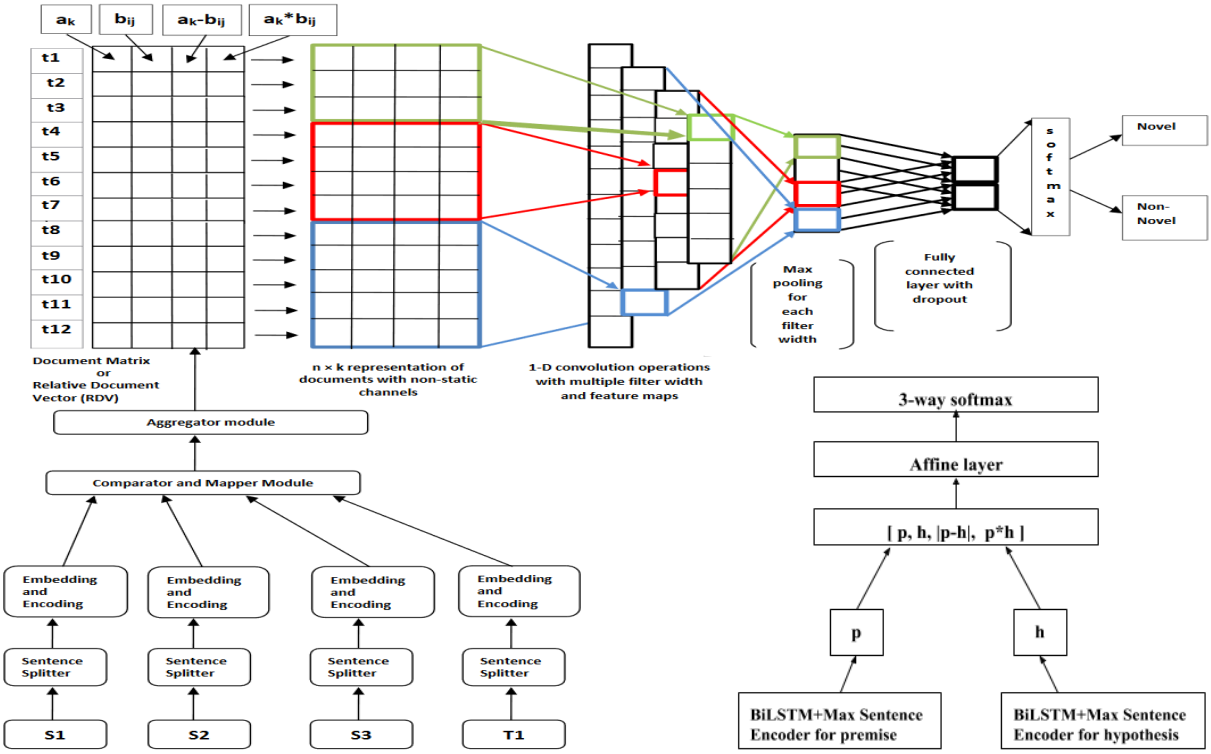


Figure 1: RDV-CNN framework for Novelty Detection. Generic SNLI Training (Conneau et al., 2017). The sentence encoder is trained on SNLI. The RDV-CNN is trained with the respective novelty datasets.

a fixed sized (2048 dimension) sentence embedding of each of the sentences[5] of the input document (source or target).

5.1.1 BiLSTM with max pooling

For a sequence of T words $\{w_t\}_{t-1,...,T}$, a bidirectional 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. We combine the varying number of $\{h_t\}_t$ to form a fixed-size vector by selecting the maximum value over each dimension of the hidden units (max pooling).

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

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

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

5.1.2 Training the sentence encoder

We follow the generic SNLI training scheme of (Conneau et al., 2017). The semantic nature of SNLI corpus makes it a good candidate for learning universal sentence embeddings in a supervised way to capture universally useful features. The training[6] on SNLI is done using the shared sentence encoder that outputs a representation for the premise p and hypothesis h (Figure 1). The representations are further concatenated as:

$$[p, h, |p - h|, p * h]$$

to form a resulting vector, which captures information from both the premise and hypothesis, and is fed into a 3-class classifier[7] (multi-layer perceptron with 1 hidden layer of 512 units) consisting of multiple fully-connected layers culminating into a softmax layer.

[5] split using NLTK

[6] SGD with a learning rate of 0.1 and a weight decay of 0.99 are used. At each epoch, the learning rate is divided by 5 if the development accuracy decreases. Training is done with a mini-batch size of 64. GloVe vectors trained on Common Crawl 840B (https://nlp.stanford.edu/projects/glove/) with 300 dimensions are used as fixed word embeddings.

[7] SNLI has 3 classes of sentence-pair judgments: neutral, contradiction, entailment

5.2 Comparator Module

This module finds the closest sentence in the source document(s) with respect to each of the target sentences. Thus each sentence (t_k) in a target document (T) is mapped to its nearest source sentence (s_{ij}) where i denotes the source document and j signifies the sentence position within that document. The encoder module outputs a fixed-length vector representation a_k for the target sentence t_k and b_{ij} for any source sentence s_{ij}. We define closeness as the maximum of cosine similarity between the vectors a_k and b_{ij}.

$$s_{cosine}(\overrightarrow{a_k}, \overrightarrow{b_{ij}}) = \frac{\overrightarrow{a_k}.\overrightarrow{b_{ij}}}{|\overrightarrow{a_k}||\overrightarrow{b_{ij}}|}$$

Thus we have a mapping relationship for each sentence in the target document with one of the source sentences.

$$b_{ij} \rightarrow a_k$$

where a_k has the max. cosine similarity with b_{ij}.

5.3 Aggregator module

This module aggregates the mappings produced in the comparator module to generate a document matrix. The mapping of a target sentence t_k to its closest source sentence s_{ij} is rendered by constructing a feature vector that captures the relation between the source and the target. This feature vector consists of the concatenation of the two sentence embeddings corresponding to t_k and s_{ij}, their absolute element-wise difference, and their element-wise product (Mou et al., 2016). The first heuristic follows the most standard procedure of the Siamese architecture, while the latter two are certain measures of "similarity" or "closeness". Thus, the *Relative Sentence Vector* (RSV) corresponding to a target sentence t_k is represented as :

$$RSV_k = [a_k, b_{ij}, |a_k - b_{ij}|, a_k * b_{ij}]$$

where comma (,) refers to the column vector concatenation. This representation is inspired from the word embedding studies (Mikolov et al., 2013) where the linear offset of vectors is seen to capture semantic relationships between the two words. (Mou et al., 2016) successfully leveraged this idea for modeling sentence-pair relationships which we alleviate to model documents. Thus for each target sentence t_k we compute the RSV and aggregate them to form the *Relative Document Vector* (RDV) of target document T_j with respect to the source documents(s) S_i. Aggregation is realized as a *slot filling task* to shape the document matrix[8] or RDV of dimension N X $4D$ where N is the number of sentences in a target document (padded when necessary) and D is the sentence embedding dimension produced by the encoder module. *Our rationale behind the RDV-CNN is: The operators:* ***absolute element-wise difference*** *and* ***product*** *would result in such a vector composition for* ***non-novel*** *sentences which would manifest '****closeness****' whereas for* ***novel*** *sentences would manifest '****diversity****'; the aggregation of which would aid in the interpretation of document level* ***novelty*** *or* ***redundancy*** *by a deep neural network. We chose CNN due to its inherent ability to automatically extract features from distinct representations.*

5.4 CNN module

The document matrix or the RDV now becomes the input to a CNN[9] for training and subsequent classification of a target document as *novel* or *non-novel* with respect to its source document(s). The CNN component is similar to the one as used in (Kim, 2014) for sentence classification. The notable difference is in the usage of word embedding. Instead of *word2vec* embeddings we use here the relative sentence embeddings of dimension $4D$ (k^{th} sentence in the document is represented by an embedding vector $RSV_k \in \mathbb{R}^D$). We experiment with the NON-STATICTEXT channel variant (embeddings gets updated during training) of the CNN.

For each possible input channel, a given document is transformed into a tensor of fixed length N (padded

[8] each row in the document matrix is one *relative sentence vector* corresponding to each target sentence t_k

[9] *tanh* as the activation function, filter windows (h) of 3,4,5 with 100 feature maps each, dropout rate (p) of 0.5 on the penultimate layer with a constraint on l_2-norms of the weight vectors. $ADADELTA$ (Zeiler, 2012) optimizer with a learning rate set to 0.1. Training via Stochastic Gradient Descent (SGD). Input batch size : 50, number of training iterations (epochs): 250. 10% of the training data for validation.

with *zero-tensors* wherever necessary to tackle variable sentence lengths) by concatenating the relative sentence embeddings.

$$RSV_{1:N} = RSV_1 \oplus RSV_2 \oplus RSV_3 \oplus ..... \oplus RSV_N$$

where $\oplus$ is the concatenation operator. To extract *local features*[10], convolution operation is applied. Convolution operation involves a *filter*, $W \in \mathbb{R}^{HD}$, which is convolved with a window of H embeddings to produce a local feature for the H target sentences. A local feature, c_k is generated from a window of embeddings $RSV_{k:k+H-1}$ by applying a non-linear function (such as hyperbolic tangent) over the convoluted output. Mathematically,

$$c_k = f(W.RSV_{k:k+H-1} + b)$$

where $b \in \mathbb{R}$ is the *bias* and f is the non-linear function. This operation is applied to each possible window of H target sentences to produce a feature map (c) for the window size H.

$$c = [c_1, c_2, c_3, ..., c_{N-H+1}]$$

A global feature is then obtained by applying *max-pooling* operation (Collobert et al., 2011) over the feature map. The idea behind *max-pooling* is to capture the most important feature-one with the highest value -for each feature map. We describe the process by which one feature is extracted from one filter (red bordered portions in Figure 1 illustrate the case of $H = 4$). The model uses multiple filters for each filter size to obtain multiple features representing the text. These features form the penultimate layer and are passed to a fully connected feed forward layer (with number of hidden units set to 100) followed by a *SoftMax* layer whose output is the probability distribution over the labels (*novel* or *non-novel*).

6 Results and Discussion

We proceed with the intuition: *redundancy* recognition would eventually lead us to *novelty* detection. The three datasets we use represent the most comprehensive resources that are publicly available and could be effectively used for document-level novelty detection.

6.1 On APWSJ: a document-level novelty detection dataset

We take upon the results reported by (Zhang et al., 2002) on APWSJ and use similar metrics to report the performance of our approach. Table 1 exhibits the effect of different redundancy measures (taken from (Zhang et al., 2002)) on APWSJ considering both *absolutely redundant* and *somewhat redundant* documents as redundant or non-novel. Our RDV-CNN approach delivers comparable performance in a 10-*fold* cross-validation classification setting. Although the system suffers in identifying the *redundant* documents but succeeds to minimize the errors committed. It signifies the affinity of our architecture towards detecting *novel* documents. It is to be noted here that (Zhang et al., 2002) conducted their experiments in an information filtering scenario using some thresholding scheme, where the **redundant or non-novel** documents were to be filtered from being delivered (*Not Delivered*). Only the *novel* documents were to be *Delivered* by the retrieval system. No learning was involved. Even APWSJ corpus was developed to support this Information Retrieval (IR) perspective of retrieving novel documents. Hence, there is a huge class imbalance. The presence of a relatively less number of *non-novel* documents in APWSJ (only 9.07%) and also that *somewhat redundant* documents are considered as *redundant*[11] in our experiment may have hindered the learning of *redundant* patterns by our system. However, the superiority of our approach is established when we experiment with two other balanced datasets that closely approximates the semantic level redundancy we aim to capture.

6.2 On Webis-CPC-11: a corpus for paraphrase detection (simulating non-novelty)

As stated earlier, we deem paraphrase detection as one close simulation of semantic level redundancy (non-novelty) detection. With this view we subject our model to detect passage-level paraphrase pairs

[10] features specific to a region in case of images or window of target sentences

[11] as originally considered in (Zhang et al., 2002). We replicate the same experimental conditions for fair comparison.

Measure	Recall	Precision	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%
Proposed Approach (RDV-CNN)	0.58	**0.76**	**22.9%**

Table 1: Results for Redundant class on APWSJ, $LM \rightarrow$ Language Model, $Mistake \rightarrow$100-Accuracy. Except for RDV-CNN, all other numbers are taken from (Zhang et al., 2002)

from Webis-CPC-11. We investigate similar evaluation systems for novelty detection as carried out in (Ghosal et al., 2018). The three novelty detection measures (Set Difference, Geometric Distance, Language Model), originally formulated by (Zhang et al., 2002) and another based on *Inverse Document Frequency (IDF)* by (Karkali et al., 2013) are our benchmarks for evaluation. Instead of setting a fixed threshold[12] as in these works we train a Logistic Regression (LR) classifier based on those measures to automatically determine the decision boundary.

Baseline 1: As a baseline we take *state-of-the-art* document embedding (*Paragraph Vector*) technique by (Le and Mikolov, 2014) for document representation; concatenate the source and target document vectors and pass it to LR.

Baseline 2: Both the target and source sentence encodings[13] are passed to separate BiLSTM layers; the two resultant vectors are concatenated and passed to a Multi Layered Perceptron (MLP) with hidden dim of 2048 followed by classification via softmax.

Table 2 clearly shows that our RDV-CNN outperforms all the measures and baselines and maintains a significant supremacy to detect semantic-level redundant passages.

Evaluation System	Description	P	R	F_1	A
Baseline 1	Paragraph Vector+LR	0.72	0.58	0.64	66.94%
Baseline 2	BiLSTM+MLP	0.71	0.73	0.72	70.91%
Novelty Measure 1(Zhang et al., 2002)	Set Difference + LR	0.71	0.52	0.60	64.75%
Novelty Measure 2(Zhang et al., 2002)	Geometric Distance + LR	0.69	0.75	0.71	70.23%
Novelty Measure 3(Zhang et al., 2002)	LM: Dirichlet Prior + LR	0.74	0.77	0.75	74.34%
Novelty Measure 4(Karkali et al., 2013)	IDF + LR	0.65	0.55	0.59	61.72%
Proposed Approach	**RDV-CNN**	**0.75**	**0.84**	**0.80**	78.02%

Table 2: Results for Paraphrase class on Webis-CPC (in %), IDF$\rightarrow$ Inverse Document Frequency, LR$\rightarrow$ Logistic Regression

6.3 On TAP-DLND 1.0: a corpus for document-level novelty detection

We experiment with the recently published TAP-DLND 1.0 corpus for our *document level novelty detection* experiments. TAP-DLND 1.0 is well balanced and consists of a fair share of different levels (lexical as well as semantic) of textual views for both novel and non-novel documents. Since our objective here is to identify both novel and non-novel documents we report results for each class. We take the same baseline as we use for Webis-CPC in the previous section. Our approach outperforms the popular semantic document embedding technique *Paragraph Vector* and another deep neural baseline (BiLSTM+MLP) by a considerable margin(12% and 6% in terms of accuracy, respectively). We also re-execute the other novelty detection measures by (Zhang et al., 2002) (Set Difference, Geometric Distance, Language Model) and (Karkali et al., 2013) (IDF) on TAP-DLND 1.0 and report the results. Our RDV-CNN even surpasses

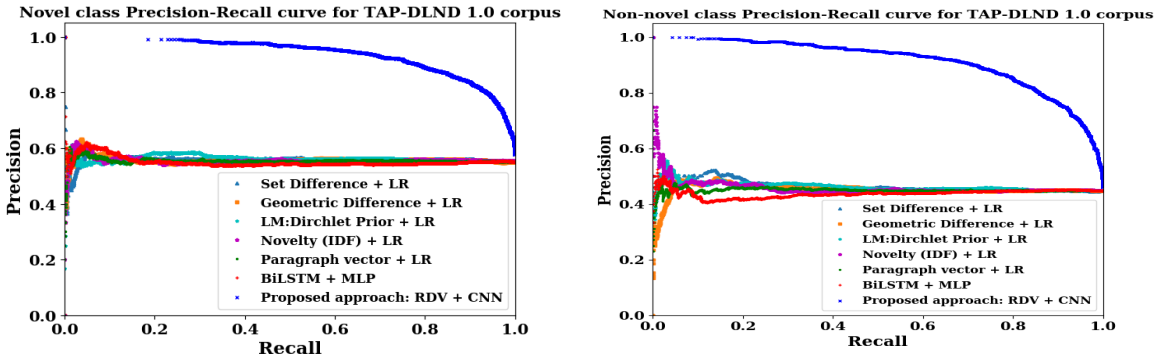
[12] the weak thresholding algorithm reported in these works yield poor results

[13] trained on SNLI as in section 5.1.2

Evaluation System	Description	P(N)	R(N)	F_1(N)	P(NN)	R(NN)	F_1(NN)	A
Baseline 1	Paragraph Vector+LR	0.75	0.75	0.75	0.69	0.69	0.69	72.81%
Baseline 2	BiLSTM+MLP	0.78	0.84	0.80	0.78	0.71	0.74	78.57%
Novelty Measure 1 (Zhang et al., 2002)	Set Difference+LR	0.74	0.71	0.72	0.72	0.74	0.73	73.21%
Novelty Measure 2 (Zhang et al., 2002)	Geometric Distance+LR	0.65	0.84	0.73	0.84	0.55	0.66	69.84%
Novelty Measure 3 (Zhang et al., 2002)	LM:Dirichlet Prior+LR	0.73	0.74	0.74	0.74	0.72	0.73	73.62%
Novelty Measure 4 (Karkali et al., 2013)	Novelty (IDF)+LR	0.52	0.92	0.66	0.66	0.16	0.25	54.26%
(Ghosal et al., 2018)	Supervised features	0.77	0.82	0.79	0.80	0.76	0.78	79.27%
Proposed Approach	**RDV-CNN**	**0.86**	0.87	**0.86**	**0.84**	**0.83**	**0.83**	**84.53%**

Table 3: Results on TAP-DLND 1.0 , P→ Precision, R→ Recall, A→ Accuracy, R→ Recall, MLP→ Multi Layer Perceptron, N→ Novel, NN→ Non-Novel, IDF→ Inverse Document Frequency

the current benchmark (Ghosal et al., 2018) on TAP-DLND 1.0 by a considerable margin in identifying both novel and non-novel documents. Figure 2 shows the consistency and edge of our approach over other measures. This behavior we attribute to our customized architecture of mapping each target sen-



(a) PRC for Novel class

(b) PRC for Non-Novel Class

Figure 2: Precision-Recall Curve for TAP-DLND 1.0

tence to the nearest source sentence to produce a *relative* document representation being input to a CNN. The CNN then extracts the features from the *relative* document representation and finally classifies the incoming document. Results signify that our network is able to learn the complex semantic interactions between source and target information necessary to conclude upon the state of novelty of a document. It is to be noted that except (Ghosal et al., 2018) and our baselines, all other measures that we consider for comparison were developed with an information retrieval perspective. The *p-values* for F_1 score produced by 20 runs of our system against the baseline are less than 0.05 and hence the improvement is statistically significant and unlikely to be observed by chance in 95% confidence interval.

6.4 Observations and Analysis

We scrutinize the results and arrive to the following observations:
(1) The RDV-CNN customized architecture of mapping each target sentence to the nearest source sentence for producing a *relative* document representation tackles the *relativity* criterion required for the problem. It facilitates CNN to extract the relevant features from the *relative* document representation which accounts for the better performance across the two datasets.
(2) Lexical approaches perform closer to our approach in detecting paraphrase pairs. This is due to the higher number of Named Entities (NEs) shared between those literary texts in Webis-CPC-11.
(3) Poor recall for non-novel class by IDF measure (Karkali et al., 2013) is due to the existence of many

The associated codes of this paper is available at: https://github.com/edithal-14/A-Deep-Neural-Solution-To-Document-Level-Novelty-Detection-COLING-2018-

new entity terms in the target documents for TAP-DLND 1.0.
(4) Lexical overlap based measures performs poorly in identifying *non-novel* documents in TAP-DLND 1.0. This behavior approves our discussion in Section 3 that we need semantic flair to address *non-novelty*. RDV-CNN bridges the gap.
(5) The deep baselines did not appear as promising to capture the semantic interactions between source and target sentences as did RDV-CNN (Section 5.3).
(6) We thoroughly examined our gold and predicted class labels. Errors committed by our system are mostly due to presence of multiple source premise sentence for a target sentence. Our RDV-CNN could capture only one premise for a target sentence and map them. We intend to accommodate multiple premises in RDV definition in our future work.
(7) The universal sentence encoding generated by BiLSTM trained on SNLI are sometimes unable to capture the complex semantic interactions between source and target sentences. This mostly occurred for new NEs and out-of-vocabulary (OOV) words.

7 Conclusion and Future work

In this work we put forward a deep learning solution for document-level novelty detection. Our deep neural network succeeds in extracting the text subtleties essential for this complex task. The system displays comparable potential on an existing resource, significant performance gain over the high-level *paraphrase detection* task and outperforms the state-of-the-art on an apt dataset. Analysis of results validates our claim that semantic features play a vital role in identifying *semantic level redundancy* aka *non-novelty* whereas relevant lexical divergence is a good indicator of *novelty*. However, we agree that given a large number of documents as source our architecture would be computationally expensive which we intend to mitigate. Also we desire to incorporate *relevance* judgment as an essential component our architecture. Our future agenda includes investigating the role of attention mechanism in creating the *Relative Document Vector* as an intermediate representation of target documents for document-level novelty detection.

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. 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). We thank the anonymous reviewers for their valuable feedback and Prof. Donia Scott, University of Sussex for her advice in the Writing Mentoring Program as part of COLING 2018. We also thank Elsevier Center of Excellence for Natural Language Processing, Indian Institute of Technology Patna for adequate help and support to carry out this research.

References

James Allan, Ron Papka, and Victor Lavrenko. 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*, pages 37–45. ACM.

James Allan, Victor Lavrenko, and Hubert Jin. 2000. First story detection in tdt is hard. In *Proceedings of the ninth international conference on Information and knowledge management*, pages 374–381. ACM.

James Allan, Courtney Wade, and Alvaro Bolivar. 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*, pages 314–321. ACM.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2011. The seventh PASCAL recognizing textual entailment challenge. In *Proceedings of the Fourth Text Analysis Conference, TAC 2011, Gaithersburg, Maryland, USA, November 14-15, 2011*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 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*, pages 632–642.

Steven Burrows, Martin Potthast, and Benno Stein. 2013. Paraphrase Acquisition via Crowdsourcing and Machine Learning. *Transactions on Intelligent Systems and Technology (ACM TIST)*, 4(3):43:1–43:21, June.

Jaime Carbonell and Jade Goldstein. 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*, pages 335–336. ACM.

Praveen Chandar and Ben Carterette. 2013. Preference based evaluation measures for novelty and diversity. In *SIGIR*.

Charles LA Clarke, Nick Craswell, and Ian Soboroff. 2004. Overview of the trec 2004 terabyte track. In *TREC*, volume 4, page 74.

Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *SIGIR*.

Charles L. A. Clarke, Nick Craswell, Ian Soboroff, and Azin Ashkan. 2011. A comparative analysis of cascade measures for novelty and diversity. In *WSDM*.

Kevyn Collins-Thompson, Paul Ogilvie, Yi Zhang, and Jamie Callan. 2002. Information filtering, novelty detection, and named-page finding. In *TREC*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Tirthankar Dasgupta and Lipika Dey. 2016. Automatic scoring for innovativeness of textual ideas. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.

Evgeniy Gabrilovich, Susan Dumais, and Eric Horvitz. 2004. Newsjunkie: providing personalized newsfeeds via analysis of information novelty. In *Proceedings of the 13th international conference on World Wide Web*, pages 482–490. ACM.

Tirthankar Ghosal, Amitra Salam, Swati Tiwary, Asif Ekbal, and Pushpak Bhattacharyya. 2018. 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.*

Donna Harman. 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*.

Margarita Karkali, François Rousseau, Alexandros Ntoulas, and Michalis Vazirgiannis. 2013. Efficient online novelty detection in news streams. In *WISE (1)*, pages 57–71.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751.

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

Xiaoyan Li and W Bruce Croft. 2005. Novelty detection based on sentence level patterns. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 744–751. ACM.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, volume 13, pages 746–751.

Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2016. Natural language inference by tree-based convolution and heuristic matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*.

Liyun Ru, Le Zhao, Min Zhang, and Shaoping Ma. 2004. Improved feature selection and redundance computing-thuir at trec 2004 novelty track. In *TREC*.

Barry Schiffman and Kathleen R McKeown. 2005. Context and learning in novelty detection. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 716–723. Association for Computational Linguistics.

Ian Soboroff and Donna Harman. 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*, pages 38–53.

Ian Soboroff and Donna Harman. 2005. Novelty detection: the trec experience. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 105–112. Association for Computational Linguistics.

Tomoe Tomiyama, Kosuke Karoji, Takeshi Kondo, Yuichi Kakuta, Tomohiro Takagi, Akiko Aizawa, and Teruhito Kanazawa. 2004. Meiji university web, novelty and genomic track experiments. In *TREC*.

Flora S Tsai and Yi Zhang. 2011. D2s: Document-to-sentence framework for novelty detection. *Knowledge and information systems*, 29(2):419–433.

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

Charles L Wayne. 1997. Topic detection and tracking (tdt). In *Workshop held at the University of Maryland on*, volume 27, page 28. Citeseer.

Yiming Yang, Tom Pierce, and Jaime Carbonell. 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*, pages 28–36. ACM.

Yiming Yang, Jian Zhang, Jaime Carbonell, and Chun Jin. 2002. Topic-conditioned novelty detection. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 688–693. ACM.

Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701.

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

Yi Zhang, Jamie Callan, and Thomas Minka. 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*, pages 81–88. ACM.

Min Zhang, Chuan Lin, Yiqun Liu, Leo Zhao, and Shaoping Ma. 2003a. THUIR at TREC 2003: Novelty, robust and web. In *Proceedings of The Twelfth Text REtrieval Conference, TREC 2003, Gaithersburg, Maryland, USA, November 18-21, 2003*, pages 556–567.

Min Zhang, Ruihua Song, Chuan Lin, Shaoping Ma, Zhe Jiang, Yijiang Jin, Yiqun Liu, Le Zhao, and S Ma. 2003b. Expansion-based technologies in finding relevant and new information: Thu trec 2002: Novelty track experiments. *NIST SPECIAL PUBLICATION SP*, (251):586–590.