Document Level Novelty Detection : Textual Entailment Lends a Helping Hand

Tanik Saikh, Tirthankar Ghosal, Asif Ekbal, Pushpak Bhattacharyya
Department of Computer Science and Engineering, Indian Institute of Technology Patna
{tanik.srf17,tirthankar.pcs16,asif,pb}@iitp.ac.in

Abstract

In this paper we present a way of detecting novelty of a document with respect to the relevant source documents with the aid of methods used in detecting Textual Entailment (TE). The proposed TE system is based on supervised machine learning approach that makes use of different similarity metrics. The TE system is further interpreted to detect the novelty of an incoming document with respect to some source document(s) already seen by the system. We design a corpus to support this foundation of novelty at the document level and coin it as the *Document Level Novelty Detection (DLND)* corpus. We employ standard supervised classification algorithms such as Support Vector Machine (SVM), Multilayer Perceptron (MLP) and Random Forest (RF) and investigate their performance on DLND. Evaluation results show the accuracies of 78.78%, 77.27% and 74.24% for SVM, MLP and RF, respectively on DLND. To establish the efficacy of our methods we evaluate our model on the benchmark datasets released in the shared task of *Recognizing Textual Entailment - 6 (RTE-6)* and *Recognizing Textual Entailment - 7 (RTE-7)*. Experiments show the accuracies of 94.91% and 96.72% on RTE-6 and RTE-7 dataset, respectively.

1 Introduction

Novelty detection from texts is an age-old problem in text mining and have found significance in various applications of Natural Language Processing (NLP) such as Text Summarization (Bysani, 2010). Novelty detection from texts implies figuring out new information from a given piece of text and subsequently arriving to the judgment that whether a given piece of text could be termed as novel or not. The decision should always be with respect to some relevant pieces of source texts. The problem of novelty detection has been studied via various NLP and machine learning (ML) paradigms ranging from classification to clustering. On the other hand TE is a NLP problem which is defined as a directional relationship between the two text fragments, termed as *Text (T)* and *Hypothesis (H)*. It is said that:

***T** entails **H** if, typically, a human reading **T** would infer that **H** is most likely to be true* (Dagan et al., 2006)

i.e. to judge that whether *H* could be inferred from *T*. This inference is not only based on understanding of T but also on some prior domain knowledge. Novelty detection finds it's relevance with TE in the sense that, a certain hypothesis *H* entailed from a certain piece of source text *T* could be considered as non-novel with respect to *T* if a human reading the hypothesis *H* after reading *T* would find redundant information in *H*. Whereas if *H* is not entailed from *T* then a human reading *H* after *T* would find new piece of information in *H* and hence *H* could be considered as novel with respect to *T*. The basis of our work also proceeds with this intuition and is grounded with the very basic relationships of textual entailment with textual similarity. Textual similarity is bi-directional relationship between two text fragments whereas textual entailment is an uni-directional relationship between the hypothesis and source text where the former could be derived from the latter but not the reverse. Similarity, it can be manifested in a scale that

S Bandyopadhyay, D S Sharma and R Sangal. Proc. of the 14th Intl. Conference on Natural Language Processing, pages 131–140,
Kolkata, India. December 2017.

ranges from semantic equivalence to complete unrelatedness, whereas TE can be either *Yes* or *No*. The implication of novelty with TE was first attempted in the TAC RTE-6 Novelty Detection Subtask (Bentivogli, 2010) and also being carried out in RTE-7 (Bentivogli, 2011). In these tracks also they defined those piece of Hypotheses as *Novel* which are *Not Entailed* by Texts. On the basis of this intuition we carry out the experiments described henceforth. These tasks were rendered at the sentence-level and they established this view of TE as an opposite characteristic to novelty. In this work we take forward this view to investigate novelty detection at the document level via TE with emphasis to textual similarity measures. The contributions of the present work could be enumerated as follows:

- Investigating the role of TE to detect novelty of a document.
- Creating our own benchmark corpora for novelty detection at the document level.

1.1 Motivation

The motivation behind the current work stemmed from the following:

- Exponential dump of redundant information across the web which hinders user quest of new meaningful pieces of information.
- Explore the implication of TE to detect novelty at the document level.

We make use of lexical level similarity features to build the TE system. The studies (Saikh et al., 2015; Saikh et al., 2016) showed that the use of similarity measures such as *Cosine Similarity, Jaccard, Dice, Overlap etc.* as features can effectively be used in taking entailment decision between a pair of texts (RTEs datasets) and these were also used in detecting paraphrase relations between a pair of texts written in Indian languages (Tamil, Malayalam, Hindi and Punjabi) as in FIRE-2016 shared task, namely *Shared Task on Detecting Paraphrases in Indian languages (DPIL)*. This straightforward relationship between textual similarity and TE encouraged us to explore various similarity measures to detect entailment at the document level. Entailment criteria lead us to investigate the novelty of the target text with respect to a set of source text(s). Our understanding and survey reveal that in spite of having great potential in various applications, novelty detection at the document level did not garner required attention. Thus investigating textual similarity measures to infer document level entailment formed the very basis of our work for detection of novelty at the document level. To the best of our knowledge our approach in viewing document level novelty detection task is novel and has not been tried before. We believe that our method towards detecting novelty of a document correlating with textual entailment would provide a strong baseline and instigate further research along this line.

1.2 Related works

Research in novelty detection could be traced back to the Topic Detection and Tracking (TDT) (Wayne, 1997) evaluation campaigns where the concern was First Story Detection (FSD) or to detect new events with respect to online news streams, notable being the UMass approach (Allan et al., 2000). The task gained popularity in the tracks of Text Retrieval Conferences (TREC) of the year of 2002, 2003 and 2004 (Voorhees, 2002; Voorhees, 2003; Clarke et al., 2004) although the focus was at sentence level novelty detection. 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), 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. At the document level, (Karkali et al., 2013) computed novelty score based on the inverse document frequency scoring function. More recently (Dasgupta and Dey, 2016) conducted experiments with information entropy measure to calculate innovativeness of a document. Novelty detection with the help of TE was first introduced as a subtask of RTE-6 (Bentivogli, 2010) challenge organized by Text Analysis Conference in the year of 2010. Several participants took part in this shared task and reported various interesting results which opened a new avenue of determining novelty with the help of TE. The best result was obtained by (Houping Jia and Xiao, 2010) with an

F-Score of 82.91%. The authors made use of Syntactic method (MINIPAR parser relationship) and semantic knowledge (Wordnet, Verb Ocean and LingPipe) to achieve the accuracy. The novelty detection subtask was again organized as a part of RTE-7 (Bentivogli, 2011). In this track the best F-Score of 90.95% was obtained by (Tsuchida and Ishikawa, 2011). Their machine learning based approach employed lexical level matching measures as features. Other participating system's results in this track were very promising and revealed that detecting novelty using entailment could be a good direction. We leverage this idea of TE for detecting novelty but at the document level. Due to the non-availability of a proper, dedicated document level novelty detection corpus, we create a dataset for the purpose. We use supervised machine learning algorithms : SVM (Vapnik, 1995; Chang and Lin, 2011), RF (Breiman, 2001) and MLP (Becerra R., 2013; Costa et al., 2015) on features extracted from our as well as RTE datasets. Evaluation shows encouraging performance on both the datasets as reported in Section 4.

2 Proposed Method for Novelty Detection

We propose a supervised scheme for detecting document level novelty using the features for detecting TE. The proposed method aims at developing a machine learning based TE system where different similarity measures were employed as features. The features include vector based similarity measures (i.e. cosine, Dice), set based similarity measures (i.e. Jaccard, Overlap and harmonic), lexical level similarity measures (i.e. unigram similarity with respect to novel/non-novel, unigram similarity with respect to source), entailment trigger polarity based similarity (based on negation), the length difference between text and hypothesis, the number of overlapping keywords and the number of overlapping Named Entities (NEs). Given a pair of documents (i.e. target-source) the system has to decide whether the target document can be entailed from any of the source(s). A document is treated as non-novel if it is fully entailed from any or all of the source documents. Else if there is sufficient new information in the target document which is not derived from the source(s), the document is viewed as novel. Paucity of a dedicated document level novelty detection corpus led us to create the corpus and we term the resource as the *Document Level Novelty Detection (DLND)* corpus. It consists of 202 different topics mostly taken from the *politics* and *business* domains. In each topic there exists at least one novel and non-novel documents and three source documents. Each target (novel or non-novel) document is compared with three source documents on the same topic. We calculate similarity scores between a target document and three on-topic source documents with the help of above mentioned measures. So for each target document pitched against the three source documents, we obtain three scores for each feature. Hence we rely on two methods, namely *Maximum* and *Average* to arrive upon the final measure.

1. *Maximum:* For each topic, each target document is compared with all the three source documents. This yields three scores for each similarity measure. We take the maximum of the three values with the intuition that a *non-novel* document would have a high similarity score with all or any one of the source document(s). Whereas a *novel* document would contain new information and would be lexically distant from all the three source documents. Hence even if we take the maximum of the similarity values, it would yield low score as compared to that of the *non-novel* documents. Let us consider there is a *novel/non-novel* target document d_t which is to be compared with three source documents d_{s1}, d_{s2} and d_{s3}. For each feature, we thus compute three scores sc_1, sc_2 and sc_3. We take the maximum of these three scores as the feature value for the respective feature.

2. *Averaging:* In this approach we take the average of the three scores obtained against the three source documents. This we do assuming that reference information is distributed in the source documents. So for a target document d_t with three source documents, d_{s1}, d_{s2} and d_{s3}, we hence obtain three scores (for each feature) sc_1, sc_2 and sc_3. We take the average of these three scores as feature value for the respective features.

For each instance we generate the feature vector consisting of all the features as mentioned above. We assign the class label as *Not Entailed*, when

we compare with a novel document and as *Entailed* when the comparison is performed with a non-novel document. We assume a piece of text as *Novel* which is *Not Entailed* with respect to the set of repositories (source documents). Such relation between novelty and TE was established in the subtask, namely novelty detection using textual entailment in (Bentivogli, 2010; Bentivogli, 2011). We develop models using three popular supervised machine learning algorithms, namely SVM with linear Kernel (Vapnik, 1995; Chang and Lin, 2011), MLP (Becerra R., 2013; Costa et al., 2015), and RF (Breiman, 2001). SVM is known to be one of the very promising classifiers for binary classification. MLP makes use of back-propagation to classify instances and random forest combines the output of multiple decision tree which is a tree based classifier. We make use of Weka [1] implementation of these classification algorithms.

2.1 Features used for Novelty Detection

Features play very crucial role in any machine learning assisted experiment. Hence, use of proper features for solving the problem is an important part of such a particular system. We use the following set of features for training and testing of classifiers:

1. *Cosine Similarity:* Cosine similarity (Nguyen H.V., 2011) is a vector based similarity metric. It calculates similarity between the two vectors of A and B by the following formula. This is a well known similarity metric and perhaps the most widely used one.

$$Cos\theta = A.B/||A||.||B|| \quad (1)$$

   where, A and B are two vector representations of two texts. The similarity score lies between 0 to -1, where, -1 indicates exactly opposite, 1 indicates exactly same, and 0 indicates the independence. It is to be assumed that higher the similarity score obtained more is the chance that the pair of text snippets become textually entailed, so it could be a good predictor of TE.

2. *Jaccard Similarity:* Jaccard similarity (Jaccard, 1901) is a set based similarity metric. It is defined as follows:

$$Jaccard(A, B) = |A \cap B|/|A \cup B| \quad (2)$$

   where A and B represent two sets of documents. A similar pair is expected to share more words and hence the entailment relation holds (Almarwani and Diab, 2017). Following this intuition we make use of set based similarity metric in our work. This is very well established similarity metric and measure the similarity between the two finite sets.

3. *Dice Similarity:* Dice Similarity (Dice., 1945) is also a vector based similarity metric. It's value lies within the range of 0 to 1. It can be calculated using the following formula.

$$Dice(A, B) = 2|A \cap B|/(|A| + |B|) \quad (3)$$

   Here, A and B represent the first and second set of documents, respectively. The mathematical derivation of this measure is same as the derivation of F-measure, where precision and recall both are taken into account. So this measure also captures both precision and recall.

4. *Overlap:* Overlap (Jayapal, 2012) is another set based similarity metric, where a discourse can be represented by a set. Elements of the set are words. It's value lies between 0 to 1. It can be calculated as per the following equation:

$$Overlap(A, B) = |A \cap B|/min(|A|, |B|) \quad (4)$$

   Here, A and B correspond to the Bag-of-Words (BoW) representation of two comparing documents.

5. *Harmonic:* Harmonic (Joshi et al., 2007) is a set based similarity metric. It can measure the similarity between two pairs of documents by the following equation

$$Harmonic(A, B) = |A \cap B|(|A + |B|)/2.|A|.|B| \quad (5)$$

   Here A and B representing two comparing documents in terms of set.

6. *Unigram similarity with respect to target document:* Here we measure the similarity between two documents by calculating the number of common unigrams between a pair of comparing documents normalized by the number of unigrams present in novel/non-novel (target) document to which the comparison is being performed. This can be illustrated by the following equation, where *nuc:*

[1] http://www.cs.waikato.ac.nz/ml/weka/

Number of common unigrams in two documents and *nut: Number of unigrams in the target document.*

$$US_t = \frac{nuc}{nut}$$

More is the overlapping of unigrams between the two documents higher is the chance of entailment between these.

7. *Unigram Similarity with respect to source document:* Unigram similarity with respect to source document is computed following the same way as the previous approach, *except* the normalization is done by the number of unigrams present in the source document. This can be represented by the following formula, where *nuc: Number of unigrams common between two documents* and *nus: Number of unigrams in source document*

$$U.S_s = \frac{nuc}{nus}$$

8. *Length difference:* The length difference between the two comparing documents is used as a feature. Our analysis to the datasets released as part of RTE-1 to RTE-5 show that length of *"Text (T)" -the entailing text* is always larger than the length of *"Hypothesis (H)" - the entailed hypothesis* as shown in Table 1, where, THP : number of T-H pairs, ATL: average text length in words and AHL : average hypothesis length in words for the development and the test set belonging to each dataset. These statistics, therefore, shows that the length difference can be used as a feature in the experiment.

Datasets	Development set			Test Set		
	THP	ATL	AHL	THP	ATL	THP
RTE-1	567	23	9	800	25	10
RTE-2	800	26	9	800	27	8
RTE-3	800	34	8	800	29	7
RTE-4	0	0	0	1000	39	7
RTE-5	600	97	7	600	96	7

Table 1: Statistics of the RTEs datasets

9. *Number of overlapping keywords:* The meaning of a textual document is often represented by a set of keywords. We extract the keywords present in each source and target document. we make use of Rapid Automatic Keyword Extractor (RAKE) [2] (Rose S. and W., 2010) for this purpose. We count the number of overlapping keywords between the two (source and target) comparing documents. This count is set as the feature value in our experiment.

10. *Number of overlapping Named Entities (NEs):* Named entities (NEs) provide important evidence in taking the entailment decision between a pair of texts. We use Stanford NER[3] for recognizing the NEs. We extract NEs present in novel, non-novel and source document and find the number of overlapping NEs between the two (source and target) comparing documents. We use this count as the feature value in our experiment.

11. *Polarity feature:* Most of the features used in our work are based on lexical matching. Presence of negation might cause a problem in the entailment decision if we rely solely on the lexical matches. As an example, let us consider the following two sentences: *T: Puja lives in Delhi.* and *H: Puja does not live in Delhi*, If we compare these two sentences using lexical matching it will produce a considerably high similarity score. Thus the system will decide these as textually entailed, but actually they are not so. In order to handle this situation we define the feature as following. A document might contain more than one negation words. In order to handle negation at the document level we make use of stanford NER tagger and RAKE key phrase extractor to identify NEs and keywords present in a particular document. In every sentence in a document we search for the keyword or NE. If any of these or both are present in a sentence, we pick up those sentences. We count the number of negation words like *"no/not"* present in those sentences. We take those count as the feature value. This is a very trivial approach and needs further investigation.

3 Dataset Description

We evaluate the efficacy of our approach on the RTE-6 and RTE-7 datasets for novelty detection subtask. It is to be noted that these two datasets were created aiming sentence-level novelty detection. However in the present work we focus on detecting document level novelty. To investigate

[2]https://github.com/aneesha/RAKE

[3]https://nlp.stanford.edu/software/CRF-NER.html

the implication of our methods for detecting novelty of a document we create the *Document Level Novelty Detection (DLND)* corpus.

3.1 Benchmark Datasets (RTE-6/7)

The novelty detection subtask was organized in conjunction with the main tasks of RTE-6 (Bentivogli, 2010) and RTE-7 (Bentivogli, 2011) tracks. In these tracks, organizers released a benchmark dataset for novelty detection using TE. We make use of this corpus to evaluate our system. In RTE-6 the novelty detection dataset consists both development and test sets. Each set contains 10 different topics. Statistics of development and test sets are shown in Table 2. There exists multiple texts for each hypothesis in both development and test datasets. The entailment decisions are either **Yes, i.e Non-novel** or **No i.e Novel** for each hypothesis and text pair.

		Development Set	Test Set
RTE-6	Topics	10	10
	Hypotheses	183	199
RTE-7	Topics	10	10
	Hypotheses	284	269

Table 2: RTE-6 and RTE-7 Novelty Subtask Dataset Statistics

3.2 DLND Corpus

We prepare the *Document Level Novelty Detection (DLND)* corpus by **unbiased** topic-wise crawling of newspaper articles belonging mostly to politics and business genre for a period of five months (from November 2016 - March 2017). The objective was to investigate, that for a given set of on-topic relevant documents already seen/read by the user, what is the novelty of an incoming on-topic document to him/her? We follow the heuristics that, on a given date, different newspapers would report similar contents regarding a specific event, and hence be content-wise non-novel to a reader once s/he had already read one of them. Reporting on subsequent dates on the same event would contain some new information, hence could be considered as novel. For this we keep three on-event reporting by different agencies as the **Source** documents usually chosen from the initial dates of reporting. Having read the source documents we ask the annotators to annotate the on-event other crawled documents from different dates as **non-novel** or **novel** with respect to the source collection based on the information coverage and human judgment. The final structure of DLND corpus looks like as shown in Figure 1. Three annotators with post-graduate level of knowledge in English were employed to use their expertise for labeling an incoming target document as *novel* if the target document has minimum semantic/lexical overlap with the source documents. A certain target document was labeled as *non-novel* if there was maximum lexical/semantic overlap with the source documents. We left out the indecisive cases for our experiments. We found the inter-rater agreement to be **0.82** in terms of **Kappa co-efficient** (Cohen, 1960) which is considered to be good as per (Landis and Koch, 1977). Intuitively, we perceive

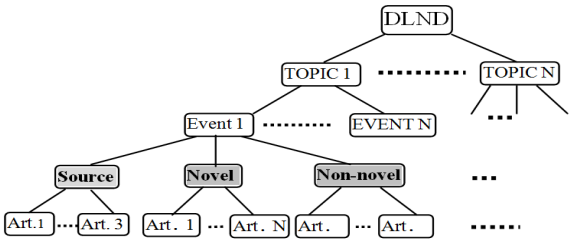

Figure 1: DLND corpus structure

the source collection of each event as the memory of the reader whereas novel and non-novel are the test instances against the knowledge of the reader. The datasets consists of 202 different topics. For each topic there exist at least one novel/non-novel document and three source documents. We partition the whole corpus into train and test sets following the ratio of 7:3. Statistics of the datasets for training and test set in terms of average document length in three categories, namely Novel, Non-Novel and Source documents are shown in Table 3.

	Novel	Non-Novel	Source
Training Set	3057	2337	2908
Test Set	1310	1001	1246

Table 3: Statistics of the DLND Datasets

4 Experiments, Results and Discussion

In this section we discuss the pre-processing done on the datasets, results obtained through experimentations and thereby analyze the errors. As the documents were collected from the various web sources, these were not well structured. We pre-processed the documents by removing white spaces.

4.1 Experiments

We calculate similarity scores between a target (novel/non-novel) and source document using various similarity measures, and use these as features in our classifiers. These scores are used to generate the feature vectors for classifier's training and/or testing. As already mentioned we used RF, MLP and SVM as our classification algorithms. These models are used to assign a class label (*Entailed or Not Entailed*) to each instance in the test set. These predicted classes are compared to the gold label to compute the final results.

Novelty and TE are highly co-related. In the context of similarity, a target document is said to be novel with respect to a collection of source document(s) if it has very less similarity to the sources. Otherwise, it is termed as novel. On the other hand similarity and TE are directly proportional if we keep aside the presence of negations in the comparing texts. TE between two texts can be judged by measuring the similarity between those two particular texts. We can conclude that novelty and TE are opposed to each other. Entailment can be a way of judging the non-novelty of a document. We report the results on test set of different classifiers in Table 4. Results reported in

	Accuracy (Percentage)	
Classifiers	**Maximum**	**Averaging**
SVM (Best Performing Classifier)	**78.78**	**78.55**
MLP	77.27	75.61
RF	74.24	69.73

Table 4: Results on DLND test datasets

Table 4 demonstrate that SVM in both the cases performs best amongst all. This is not unexpected keeping in mind the success of SVM in solving a wide range of text classification problems with features which are overlapping in nature. MLP makes use of back-propagation technique to classify instances. In our setting we use 5 layers that might have caused better accuracy. Random Forest also seems to suit well to our task.

4.2 Results on benchmark datasets

We evaluate our model on the benchmark datasets of RTE-6 and RTE-7 for novelty detection. The task was to detect those hypotheses which are novel (not-entailed) with respect to the corpus. We show the results in Table 5, where P: Precision, R: Recall and F1: F-score. We also compare the performance with the best systems reported in RTE-6 (Houping Jia and Xiao, 2010) and also in RTE-7(Tsuchida and Ishikawa, 2011). The best result obtained in RTE-6 novelty detection subtask is with the F-score of 82.91% by (Houping Jia and Xiao, 2010). Syntactic (output of MINIPAR parser, nodes matching texts and hypotheses) and semantic (WordNet, Verb Ocean, and LingPipe) matching between texts and hypotheses were employed for that purpose. An F-score of 90.95% was obtained as the best score by (Tsuchida and Ishikawa, 2011) on RTE-7 novelty detection dataset, with mostly lexical matching features in a machine learning framework. As is evident, our proposed system successfully outperforms those state-of-the-art techniques of RTE-6 and RTE-7 by a significant margin.

		P(%)	R(%)	F1(%)
RTE-6	(Houping Jia and Xiao, 2010)	72.39	97	82.91
	Proposed Method	**95.74**	**99.08**	**96.86**
RTE-7	(Tsuchida and Ishikawa, 2011)	86.92	95.38	90.95
	Proposed Method	**96.97**	**99.73**	**98.33**

Table 5: comparison of results obtained with the best system's results on RTE-6 and RTE-7

4.3 Sensitivity Analysis of Features

In order to illustrate the contribution of each feature to our predicting class, we perform an ablation study. Table 6 below reports the accuracy figures on training set (based on 10-fold cross validation) by removing one feature after another, where the acronyms *U.S.N, U.S.S, L.D, Keyword and NE stands for Unigram similarity with respect to target (Novel/Non-Novel) document, Unigram similarity with respect to source, Length Difference, number of overlapping keywords and number of overlapping Named Entities respectively*. Table 6 shows that 'unigram similarity with respect to target document', # of keywords match, # of NE match, and Cosine similarity are the most contributing features to our experiments.

Feature Removed	Accuracy (%)
None	85.38
Cosine Similarity	84.85
Jaccard Similarity	85.10
Dice	85.17
Overlapping	85.13
Harmonic	84.85
U.S.N	83.60
U.S.S	84.06
L.D	85.03
Keyword	82.70
NE	83.12
Polarity	84.96

Table 6: Feature sensitivity analysis

4.4 Error Analysis

Below we analyze the output of the system and summarize the causes of the errors committed.

1. In our current work we assumed that more the similarity at the lexical level, higher is the chance that the document pair is entailed to each other. The intuition behind this lexical matching based experiment was grounded with a very basic assumption that more the overlapping tokens between two comparing documents higher is the chance of holding TE relation between that pair of text snippets. Although this assumption works up to a certain extent, but fails when semantics is to be considered.

2. Presence of negation words often creates problem in entailment decision. To overcome this we make use of polarity based feature (i.e presence/absence of negation words). This intuition works well for the single occurrence of negation word, but as we deal with documents there might be multiple negation words in a particular document. Dealing with multiple occurrences of negation words at the document level is very challenging. We will investigate this in more details in the future.

3. Although the proposed system considers the *NEs and keywords*, but it does not take *Multiword Expressions (MWEs)* into account. Dealing with multi-word expressions may be useful in taking entailment decision.

4. One of the major drawbacks of this system is the sparsity problem. The system represents a text with lexical-level sparse vectors. So, there might be some instances (having different vocabulary) for which similarity measure can produce zero score, even though they are highly entailed.

4.5 Comparisons with the state-of-the-art

In order to compare our method with state-of-the-art systems we evaluate a recent method proposed in (Dasgupta and Dey, 2016) on our DLND corpus. This particular entropy-based approach produced novelty score (NS) of a document **d** with respect to a collection **c**. We adapt the respective threshold criteria and infer that documents with novelty score above *(average+standard deviation)* are *Novel* and that with novelty score below *(average-standard deviation)* are *Non-Novel*. We left out the remaining *average novelty* class cases.

System	Accuracy (%)	F_1 (%)
(Dasgupta and Dey, 2016)	67.94	70.34
Proposed Approach (Maximum-SVM)	**78.78**	**93.49**

Table 7: Comparison with the state-of-art

From Table 7 we could see that our proposed *Maximum* method based on SVM classifier performs better compared to the approach of (Dasgupta and Dey, 2016) by a margin of almost 11 points in terms accuracy.

4.6 Tests of Significance

To analyze if the improvement obtained in our system is statistically significant over the state-of-the art, we perform *t-test* at 5% significance level. The *p-values* for F-measures produced by 20 runs of our system against the best performing systems of RTE-6 was 5.30e-85 and for RTE-7 was 1.60e-74. We also pitched our system's F-measure against that obtained by the approach of (Dasgupta and Dey, 2016) on DLND for 20 runs and the p-value was 2.27e-91. All the p-values thus are less than 0.05 and hence the improvement is statistically significant and unlikely to be observed by chance in 95% confidence interval.

5 Conclusion and Future Works

In this work we addressed the problem of detection of novelty of a document with respect to on-topic source document(s) using the concept of TE. We built an entailment model based on supervised approaches that make use of features extracted from the different lexical level similarity metrics. We also created a dedicated resource for document level novelty detection which may pave the way for further research in this topic. Our evaluation on DLND shows promising results to serve as a strong baseline for further research. Evaluation on the RTE-6 and RTE-7 datasets demonstrate the effectiveness of our approach over the existing literature methods on novelty detection. The research carried out in these experiments opens up a new avenue for detecting novelty of text at document level using textual entailment.
In future, we would like:

1. To employ deep semantic features so that the system can capture ambiguous sentences contained in a particular document.

2. To investigate semantic textual similarity to detect novelty of a document with deep learning techniques.

3. To address the sparsity problem, we intend to incorporate WordNet based similarity measures and explicit semantic analysis that will use bag-of-word representation retrieved from the Wikipedia text. Also distributional representation of words(word2vec) may prove effective to capture semantics.

4. To see the performance of the best performing systems of RTE-6 (Houping Jia and Xiao, 2010) and in RTE-7 (Tsuchida and Ishikawa, 2011) applied to our DLND dataset.

Acknowledgments

We would like to acknowledge "Elsevier Centre of Excellence for Natural Language Processing" at IIT Patna for supporting the research work furnished here in this paper.

References

James Allan, Victor Lavrenko, Daniella Malin, and Russell Swan. 2000. Detections, Bounds, and Timelines: Umass and tdt-3. In *Proceedings of topic detection and tracking workshop*, pages 167–174.

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.

Nada Almarwani and Mona Diab. 2017. Arabic Textual Entailment with Word Embeddings. In *Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP)*, pages 185–190, Valencia, Spain.

Garca Bermdez R.V. Velzquez L. Rodrguez R. Pino C. Becerra R., Joya G. 2013. Saccadic Points Classification Using Multilayer Perceptron and Random Forest Classifiers in EOG Recordings of Patients with Ataxia SCA2. *(eds) Advances in Computational Intelligence. IWANN. Lecture Notes in Computer Science*, 7903(3).

Magnini B. Dagan I. Dang H.T. Giampiccolo D. Bentivogli, L. 2010. The Sixth PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the Text Analysis Conference (TAC 2010), November 15-16, 2010 National Institute of Standards and Technology Gaithersburg, Maryland, USA*.

Clark P. Dagan I. Dang H. T. Giampiccolo D. Bentivogli, L. 2011. The Seventh PASCAL Recognizing Textual Entailment Challenge. In *In TAC 2011 Notebook Proceedings, November 14-15, 2011, Gaithersburg, Maryland, USA*.

Leo Breiman. 2001. Random Forests. *Mach. Learn.*, 45(1):5–32.

Praveen Bysani. 2010. Detecting Novelty in the Context of Progressive Summarization. In *Proceedings of the NAACL HLT 2010 Student Research Workshop*, pages 13–18, Los Angeles, CA, June. Association for Computational Linguistics.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27.

Charles LA Clarke, Nick Craswell, and Ian Soboroff. 2004. Overview of the TREC 2004 Terabyte Track. In *TREC*, volume 4, page 74, National Institute of Standards and Technology Gaithersburg, MD.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and psychological measurement*, 20(1):37–46.

Kevyn Collins-Thompson, Paul Ogilvie, Yi Zhang, and Jamie Callan. 2002. Information Filtering, Novelty Detection, and Named-Page Finding. In *TREC*, Gaithersburg,MD.

Wanderson Costa, Leila Maria Garcia Fonseca, and Thales Sehn Körting. 2015. Classifying Grasslands and Cultivated Pastures in the Brazilian Cerrado Using Support Vector Machines, Multilayer Perceptrons and Autoencoders. In *Machine Learning and Data Mining in Pattern Recognition - 11th International Conference, MLDM 2015, Hamburg, Germany, July 20-21, 2015, Proceedings*, pages 187–198.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, pages 177–190, Berlin, Heidelberg. Springer-Verlag.

Tirthankar Dasgupta and Lipika Dey. 2016. Automatic Scoring for Innovativeness of Textual Ideas. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.

Lee R. Dice. 1945. Measures of the Amount of Ecologic Association between Species. *Ecology*, 26(3):297302.

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.

Tengfei Ma Xiaojun Wan Houping Jia, Xiaojiang Huang and Jianguo Xiao. 2010. PKUTM Participation at TAC 2010 RTE and Summarization Track. National Institute of Standards and Technology Gaithersburg, Maryland, USA.

Paul Jaccard. 1901. Étude Comparative de la Distribution Florale dans une Portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579.

Arun Jayapal. 2012. Similarity Overlap Metric and Greedy String Tiling for Plagiarism Detection at PAN 2012. In Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors, *CLEF (Online Working Notes/Labs/Workshop)*, volume 1178 of *CEUR Workshop Proceedings*, Rome, Italy. CEUR-WS.org.

Pushkar Joshi, Mark Meyer, Tony DeRose, Brian Green, and Tom Sanocki. 2007. Harmonic Coordinates for Character Articulation. *ACM Trans. Graph.*, 26(3), july.

Margarita Karkali, François Rousseau, Alexandros Ntoulas, and Michalis Vazirgiannis. 2013. Efficient Online Novelty Detection in News Streams. In *International Conference on Web Information Systems Engineering*, pages 57–71. Springer.

J Richard Landis and Gary G Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *biometrics*, pages 159–174.

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.

Bai L. Nguyen H.V. 2011. Cosine Similarity Metric Learning for Face Verification. In *Kimmel R., Klette R., Sugimoto A. (eds) Computer Vision ACCV 2010. ACCV 2010.*, volume 6493 of *Lecture Notes in Computer Science*, pages 709–720, Berlin, Heidelberg. Springer.

Cramer N. Rose S., Engel D. and Cowley W. 2010. Automatic keyword extraction from individual documents. *In M. W. Berry and J. Kogan (Eds.), Text Mining: Theory and Applications: John Wiley and Sons.*

Liyun Ru, Le Zhao, Min Zhang, and Shaoping Ma. 2004. Improved Feature Selection and Redundance Computing-THUIR at TREC 2004 Novelty Track. In *TREC*, Gaithersburg,MD.

Tanik Saikh, Sudip Kumar Naskar, Chandan Giri, and Sivaji Bandyopadhyay. 2015. Textual Entailment Using Different Similarity Metrics. In *Computational Linguistics and Intelligent Text Processing - 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, Proceedings, Part I*, pages 491–501.

Tanik Saikh, Sudip Kumar Naskar, and Sivaji Bandyopadhyay. 2016. JU_NLP@DPIL-FIRE 2016: Paraphrase Detection in Indian Languages - A machine Learning Approach. In *Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation, Kolkata, India.*, pages 275–278.

Barry Schiffman and Kathleen McKeown. 2005. Context and Learning in Novelty Detection. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 716–723, Vancouver, British Columbia, Canada, October. 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*.

M. Tsuchida and K. Ishikawa. 2011. IKOMA at TAC2011: A Method for Recognizing Textual Entailment using Lexical-level and Sentence Structure-level Features. National Institute of Standards and Technology Gaithersburg, Maryland, USA.

Vladimir N. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA.

Ellen M Voorhees. 2002. Overview of TREC 2002. In *Trec*, National Institute of Standards and Technology Gaithersburg, MD.

Ellen M Voorhees. 2003. Overview of TREC 2003. In *TREC*, pages 1–13, National Institute of Standards and Technology Gaithersburg, MD.

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

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 *TREC*, pages 556–567, Gaithersburg,MD.

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.