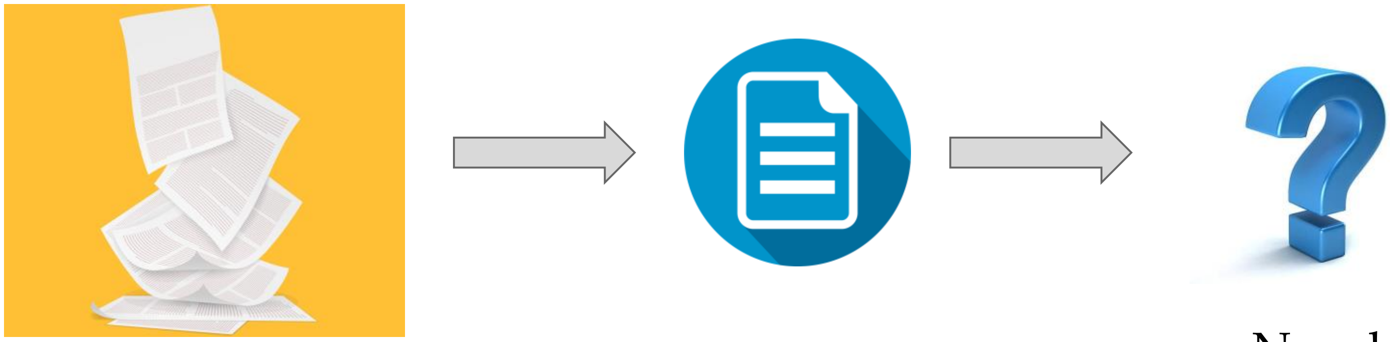
Novelty Goes Deep. A Deep Neural Solution To Document-Level Novelty Detection

Subhadeep Koner
MABE, UTK
COSC 521, Spring 2019.

Novelty Detection

- ***Novelty:*** *The search of new; eternal quest of the inquisitive mind*



Novel

Motivation and Contribution

- Exponential rise of redundant/duplicate information/documents across the web
- Redundancy at the semantic level (text reuse,rewrite,paraphrase, etc.)
- Mostly IR oriented rule-based existing methods, at the sentence-level
- Plagiarism Detection at the semantic level

★ No handcrafted rules, only from the data

★ Leveraging the semantic power of natural language inference towards detection of redundancy/non-novelty

★ Encapsulating source and target information within an effective document representation for learning via a deep neural network

Textual Novelty Detection

- ❖ Novelty Mining: elicit new information from texts
- ❖ An IR task for long: retrieve novel sentences
- ❖ Document-Level Novelty Detection: A frontier less explored
- ❖ Properties (Ghosal et. al, 2018):
  - Relevance
  - Relativity
  - Diversity
  - Temporality

{ w.r.t. a set of seed documents called as the source *or* information already known/memory of the reader }

- ❖ Applications in diverse domains of information processing :
  - Extractive text summarization
  - News Tracking
  - Predicting scholarly articles impact

Problem Definition

- Categorize a document as novel or non-novel based on sufficient relevant new information
- For e.g., :
  - *d1 : Singapore is an island city-state located at the southern tip of the Malay Peninsula. It lies 137 kilometers north of the equator.*
  - *d2 : Singapore's territory consists of one main island along with 62 other islets. The population in Singapore is 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.*
- If we consider source as d1 and d2; d3 is novel, d4 is non-novel
- We take a very objective and simplistic view considering only the new information content.
- We investigate whether a deep network can be trained to perceive novelty at the document-level and also identify semantically redundant/non-novel documents

Datasets: APWSJ

- We select datasets that serve as a suitable testbeds for our investigation.
- APWSJ (Associated Press-Wall Street Journal) Novelty Detection Corpus (Zhang et al., 2002)
    - Developed using the TREC 2002 and 2004 datasets
    - Developed from an IR perspective
    - Topicwise stream of documents labelled as redundant, somewhat redundant and completely redundant
    - Unmarked ones are Novel as per the dataset definition.
    - 33 topics are used for the experiment as in the original paper
    - Only 9.07% documents are absolute redundant
    - Hence we take both somewhat redundant and completely redundant documents as redundant, as also is reported in one experiment in the original paper

Datasets: Webis-CPC-11

- Webis-CPC-11 (The Webis Crowd Paraphrase Corpus 2011)
  - Simulates a higher form of semantic redundancy at the paragraph-level
  - Upon literary text
  - Binary class: Paraphrase and Not Paraphrase
  - Paraphrases simulate non-novelty
  - Non-Paraphrases does not necessarily mean Novel. Relevance, Relativity are not always preserved. However Diversity is there.
  - Investigation interest is in the detection of paraphrases aka non-novelty
  - 4,067 paraphrases; 3,792 non-paraphrases

Datasets: TAP-DLND 1.0

- TAP-DLND 1.0 (Tirthankar-Asif-Pushpak Document-Level Novelty Detection Corpus)
  - ➢ A balanced document-level novelty detection dataset
  - ➢ Consists of events belonging to different categories
  - ➢ Satisfying Relevance, Relativity, Diversity, Temporality criteria for Novelty
  - ➢ 3 source documents per event; target documents are annotated against the information contained in the source documents
  - ➢ Binary Classification: Novel or Non-Novel
  - ➢ 2736 novel and 2704 non-novel documents
  - ➢ Inter-annotator agreement is 0.82

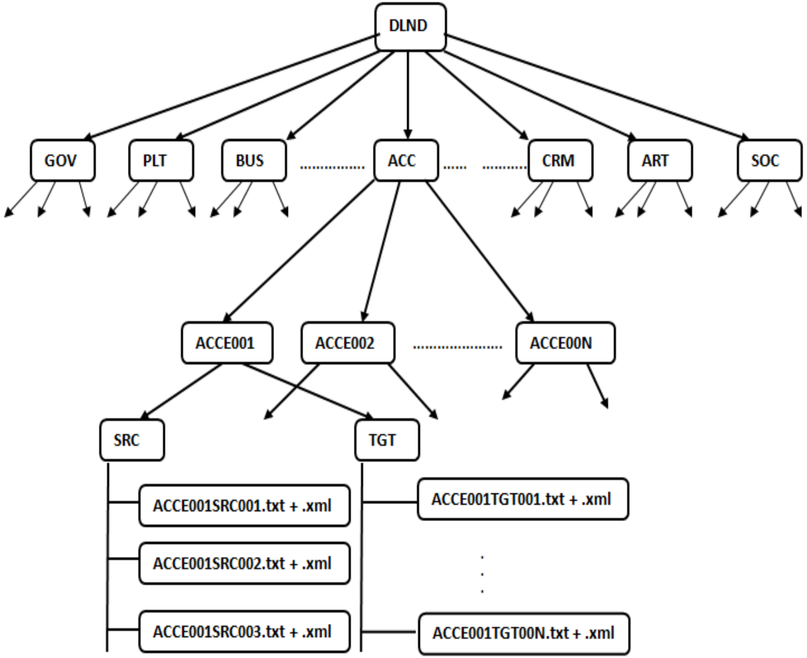


Fig 1: TAP-DLND 1.0 Structure (Ghosal et. al, 2018)

Proposed Model

- *Objective: Given a set of relevant documents to a topic, can a neural network learn the state of novelty of an incoming document? Can it identify semantically redundant documents?*
- Building upon the idea of Textual Entailment and borrowing the semantics involved in natural language inference from the large scale Stanford Natural Language Inference (SNLI) corpus
- Each document is splitted into sentences. The sentences are encoded using the representation of a Bidirectional LSTM + max pooling trained on SNLI
- Idea is to create an effective target document representation that could encapsulate both source and target information in it. (Novelty of a text cannot be universally defined, unless the source is identified; with respect to what? is the natural argument)

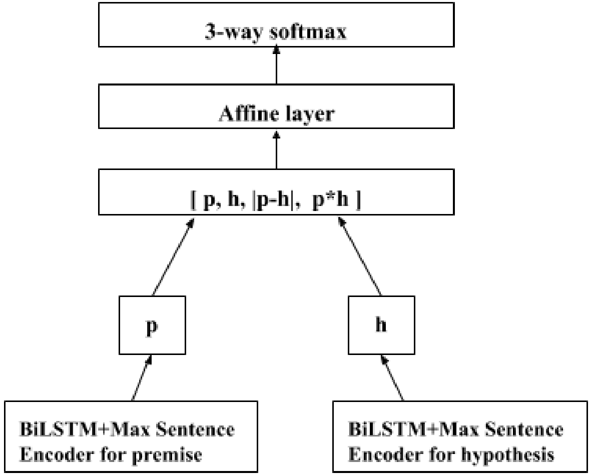
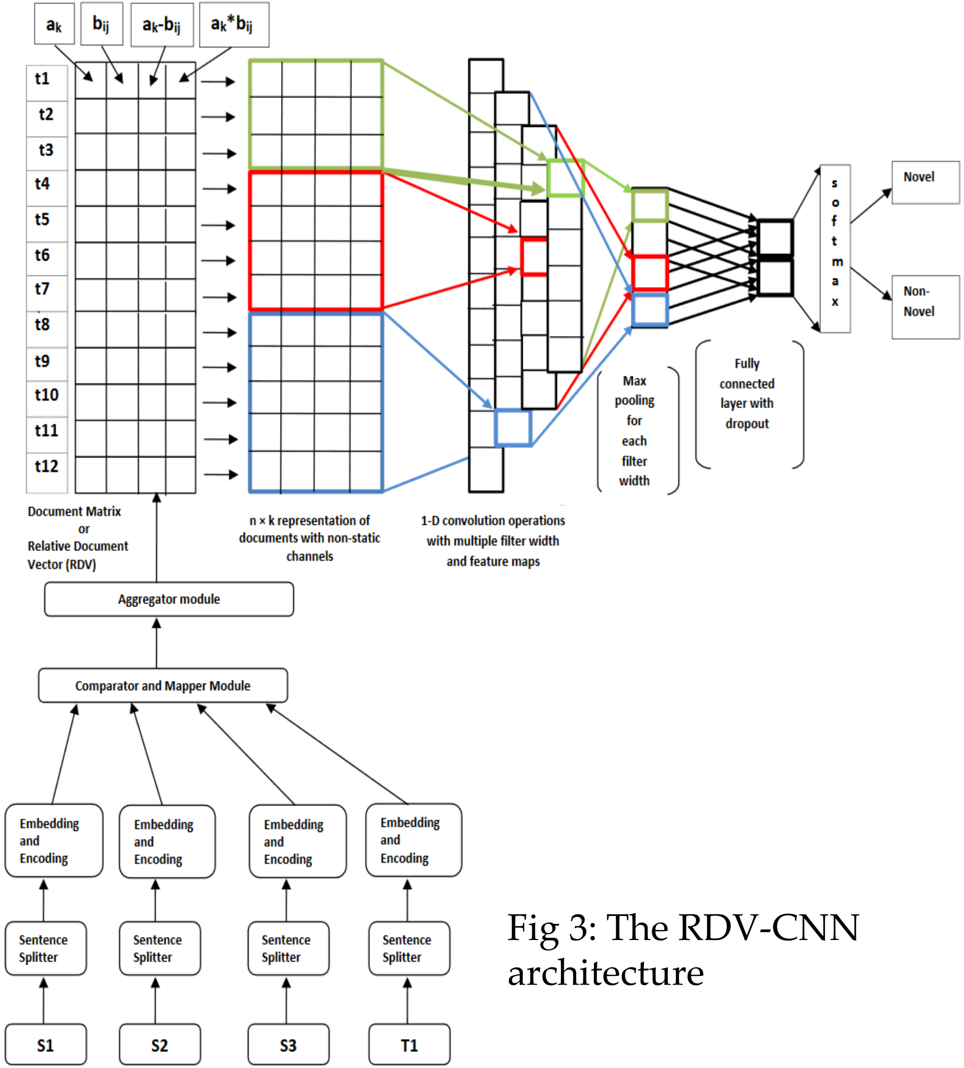

3-way softmax
Affine layer
[p, h, |p-h|, p*h]
p
h
BiLSTM+Max Sentence Encoder for premise
BiLSTM+Max Sentence Encoder for hypothesis

Fig 2: SNLI training of the sentence encoders (Conneau et al., 2017)

RDV-CNN

- For each target sentence, the nearest source sentence is pulled using cosine similarity between the sentence vectors.
- Each of the source encapsulated target sentence representation is stacked to form the Relative Document Vector (RDV)
- A target sentence with its nearest source sentence is encoded as:
  - $a_k | b_{ij} | a_k\text{-}b_{ij} | a_k{*}b_{ij}$ (Mou et al., 2016)

Where k is the number of sentences in the target document and b_{ij} is the semantic representation of the j-th sentence of the i-th source document.

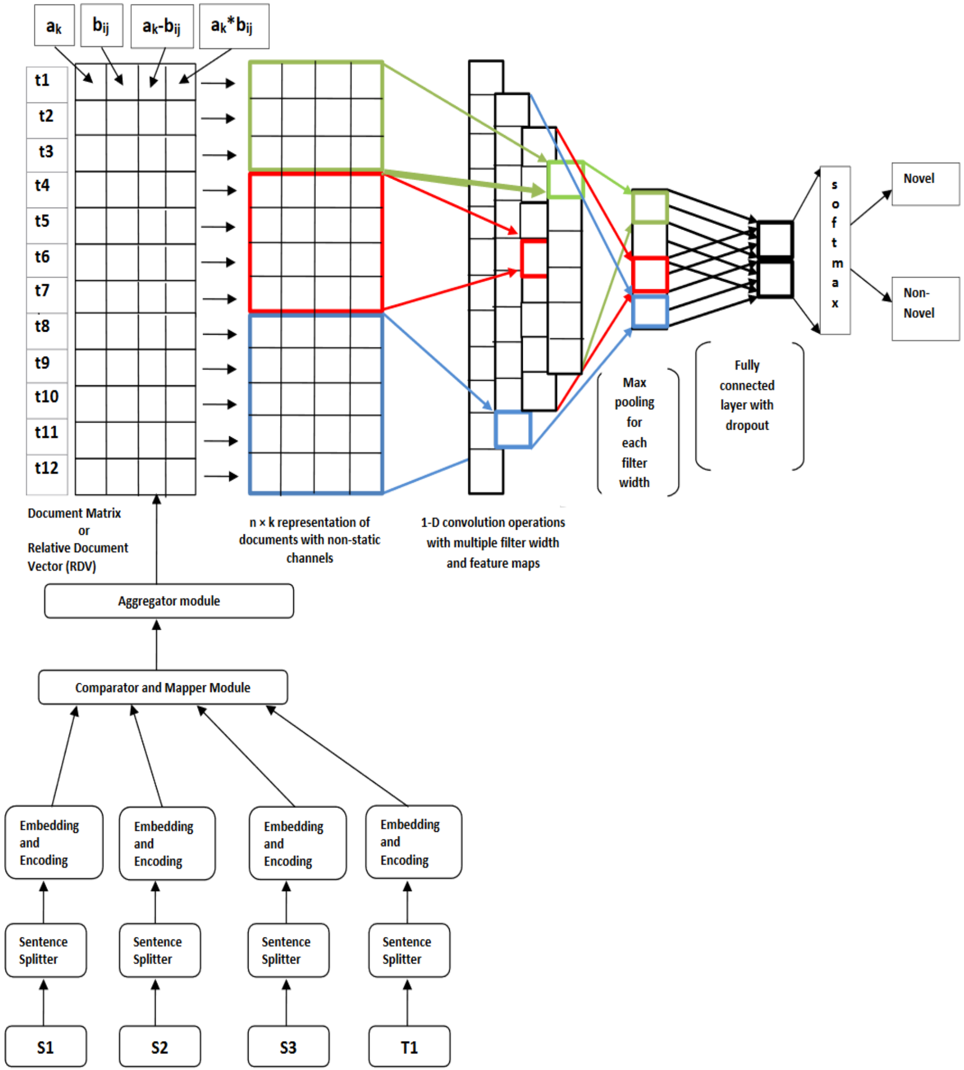


a_k
b_{ij}
a_k-b_{ij}
a_k*b_{ij}
t1
t2
t3
t4
t5
t6
t7
t8
t9
t10
t11
t12
Document Matrix or Relative Document Vector (RDV)
n × k representation of documents with non-static channels
1-D convolution operations with multiple filter width and feature maps
Max pooling for each filter width
Fully connected layer with dropout
softmax
Novel
Non-Novel
Aggregator module
Comparator and Mapper Module
Embedding and Encoding
Embedding and Encoding
Embedding and Encoding
Embedding and Encoding
Sentence Splitter
Sentence Splitter
Sentence Splitter
Sentence Splitter
S1
S2
S3
T1

Fig 3: The RDV-CNN architecture

Why RDV?

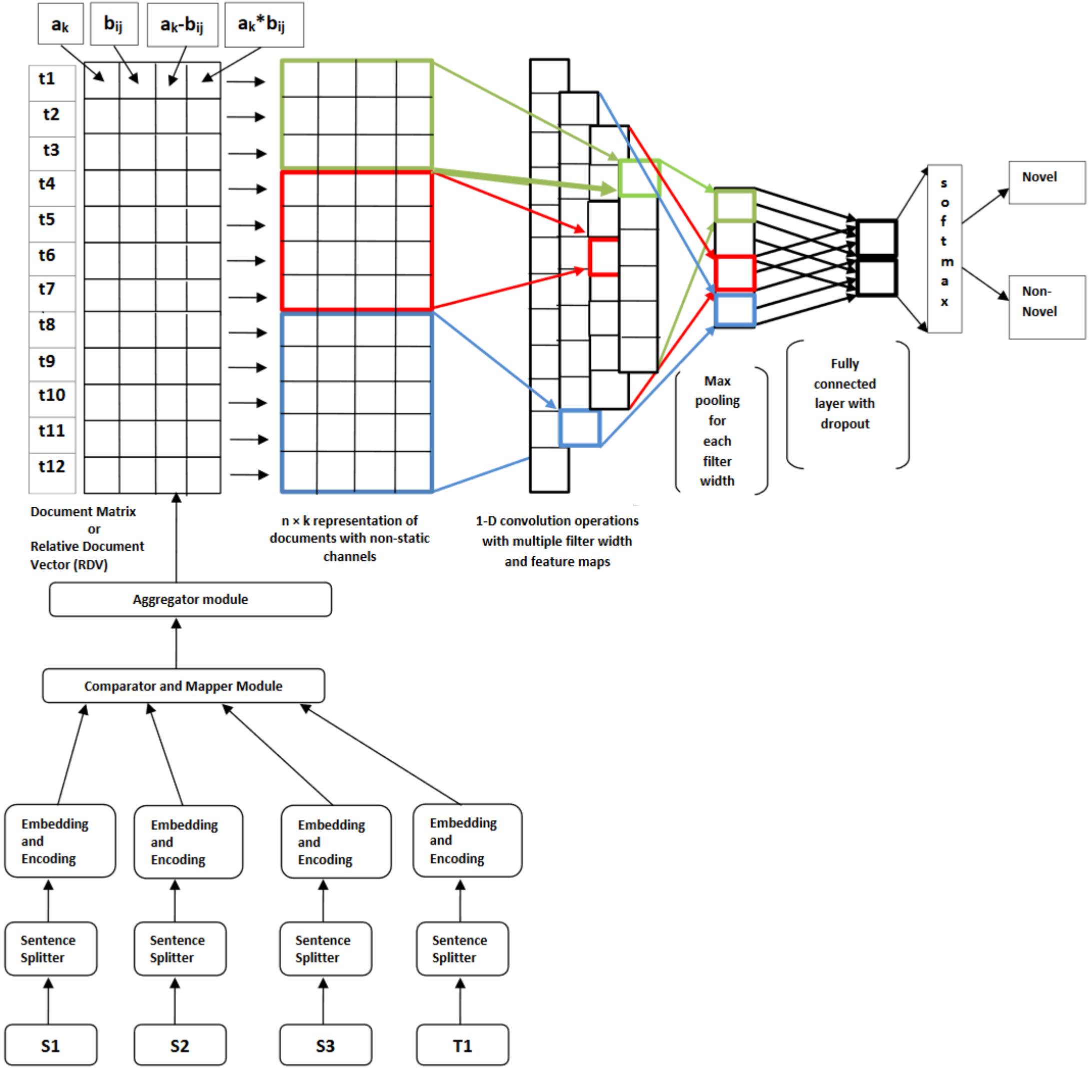
- Intuition: A non-novel document would contain many redundant sentences. Hence cosine similarity will pull that particular source sentence which contributes more towards making the target sentence redundant. Hence a joint encoding of source+redundant sentence would be different from that of a source+novel sentence.
- Thus the RDV of a non-novel document would be different from that of a novel document.
- A Convolutional Neural Network (CNN) is then trained with the RDV of the target documents



a_k
b_{ij}
a_k-b_{ij}
a_k*b_{ij}
t1
t2
t3
t4
t5
t6
t7
t8
t9
t10
t11
t12
Document Matrix or Relative Document Vector (RDV)
n × k representation of documents with non-static channels
1-D convolution operations with multiple filter width and feature maps
Max pooling for each filter width
Fully connected layer with dropout
softmax
Novel
Non-Novel
Aggregator module
Comparator and Mapper Module
Embedding and Encoding
Embedding and Encoding
Embedding and Encoding
Embedding and Encoding
Sentence Splitter
Sentence Splitter
Sentence Splitter
Sentence Splitter
S1
S2
S3
T1

Setup

- Sentence Encoder trained on SNLI (word vectors: GloVe 800B)
- S1, S2,... are source documents
- T1 is the target document
- Relevance Detection is not taken care of here (inherently manifested within the datasets taken)
- We report the 10-fold cross validation performance



Rationale

- *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*.
- Relevance criteria is inherently manifested within the datasets we work with.
- The proposed architecture looks for relative, diverse new information of a target with respect to corresponding sources and learns the notion of a novel or non-novel document.
- Learning of novel vs. non-novel patterns via the relative representation

Results on APWSJ

- On the APWSJ dataset. Except the proposed method we take all other numbers from (Zhang et al., 2002)
- Mistake=100-Accuracy as is there in the original paper.

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: Mixed	0.56	0.67	27.4%
Proposed Method	**0.58**	**0.76**	**22.9%**

Table 1: Comparison with (Zhang et al., 2002) on APWSJ

Results on the Paraphrase Detection Task

- On the Webis-CPC-11 dataset
- Interest is on to detect the semantically redundant paraphrases: non-novelty

Evaluation System	**Description**	**Precision**	**Recall**	**F-measure**	**Accuracy**
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	Set Difference+LR (Zhang et al., 2002)	0.71	0.52	0.60	64.75%
Novelty Measure 2	Geometric Distance+LR (Zhang et al., 2002)	0.69	0.75	0.71	70.23%
Novelty Measure 3	Language Model (KLD) +LR (Zhang et al., 2002)	0.74	0.77	0.75	74.34%
Novelty Measure 4	IDF+LR (Karkali et al., 2013)	0.65	0.55	0.59	61.72%
Proposed Approach	**RDV-CNN**	**0.75**	**0.84**	**0.80**	**78.02%**

Table 2: Performance on Webis-CPC-11 (Non-Novelty Detection)

Results on TAP-DLND 1.0

Evaluation System	Description	Precision (Novel)	Recall (Novel)	F-measure (Novel)	Precision (Non-Novel)	Recall (Non-Novel)	F-measure (Non-Novel)	Accuracy
Baseline 1 (without SNLI pre-training))	Paragraph Vector+LR	0.75	0.75	0.75	0.69	0.69	0.69	72.81%
Baseline 2 (without RDV-CNN)	BiLSTM+MLP	0.78	0.84	0.80	0.78	0.71	0.74	78.57%
Novelty Measure 1	Set Difference+LR (Zhang et al., 2002)	0.74	0.71	0.72	0.72	0.74	0.73	73.21%
Novelty Measure 1	Geometric Distance+LR (Zhang et al., 2002)	0.65	0.84	0.73	0.84	0.55	0.66	69.84%
Novelty Measure 1	LM:(KLD)+LR (Zhang et al., 2002)	0.73	0.74	0.74	0.74	0.72	0.73	73.62%
Novelty Measure 1	IDF+LR (Karkali et al., 2013)	0.52	0.92	0.66	0.66	0.16	0.25	54.26%
(Ghosal et al., 2018)	Supervised Method (Feature-Based)	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: RDV-CNN on TAP-DLND 1.0

Observations & Analysis

- Named Entities (NEs) are important to establish relevance between texts. Lexical measures performs close to ours in Webis-CPC-11, due to large number of NEs in those literary texts
- Lexical approaches do not fare well to identify non-novel content in TAP-DLND 1.0. Our RDV-CNN based on semantic knowledge from SNLI is able to identify semantic level redundancy to some extent.
- Novel texts are mostly lexically different, but non-novel texts exhibit semantic-level redundancy; hence harder to identify
- We need semantic flair to address non-novelty; hence our system fares well
- Encapsulation of the nearest source information with the target provided better means for feature discovery by the CNN. Also rich sentence embeddings from the SNLI corpus contributed to the better performance than the baselines.
- Errors committed by our system is due to:
    - Multiple premises contributing to one target sentence
    - Presence of new NEs or OOVs in the test data

Conclusions

- A first hand deep neural attempt towards document-level novelty detection leveraging the knowledge from textual entailment
- Outperformed existing methods
- Relevance detection is a prelude to the task which is to be addressed. Identifying appropriate source documents out of a pool.
- To tackle multi-premise scenarios. Highly unlikely that one sentence would only contribute towards a target sentence.

THANK YOU !!

QUESTIONS ??