Politecnico di Torino

Hybrid Text Summarization through Reinforcement Learning

Sicilian Team

Francesco Di Salvo
francesco.disalvo@studenti.polito.it

Gianluca La Malfa
gianluca.lamalfa@studenti.polito.it

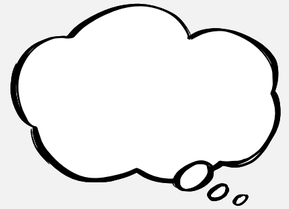
Key points

- Explore hybrid neural summarization
- Pipeline tested on FNS dataset and CNN/Daily Mail dataset
- Performance evaluation with Rouge score
- Comparison between Rouge and BERT score.

Introduction

Extractive TS

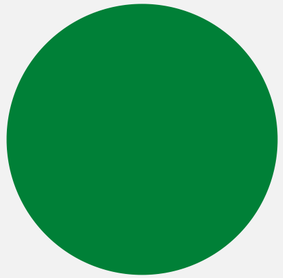
- Extract and concatenate relevant sentences
- Need for ranking functions



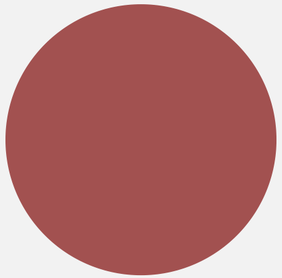
Abstractive TS

- Paraphrasing salient text
- Leverages NLU and NLG

Methodology



Reinforcement Learning

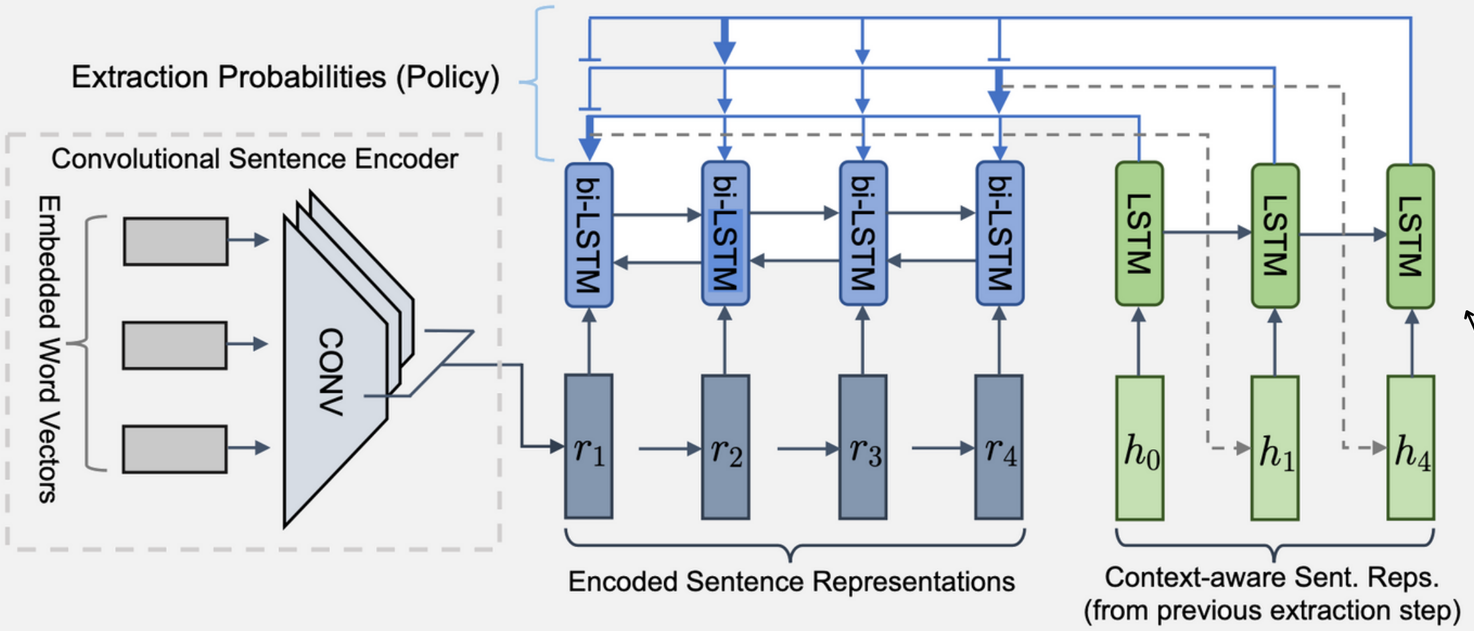


Pre-trained Extractor

Pre-trained Abstractor

Extractor

Captures context aware representation



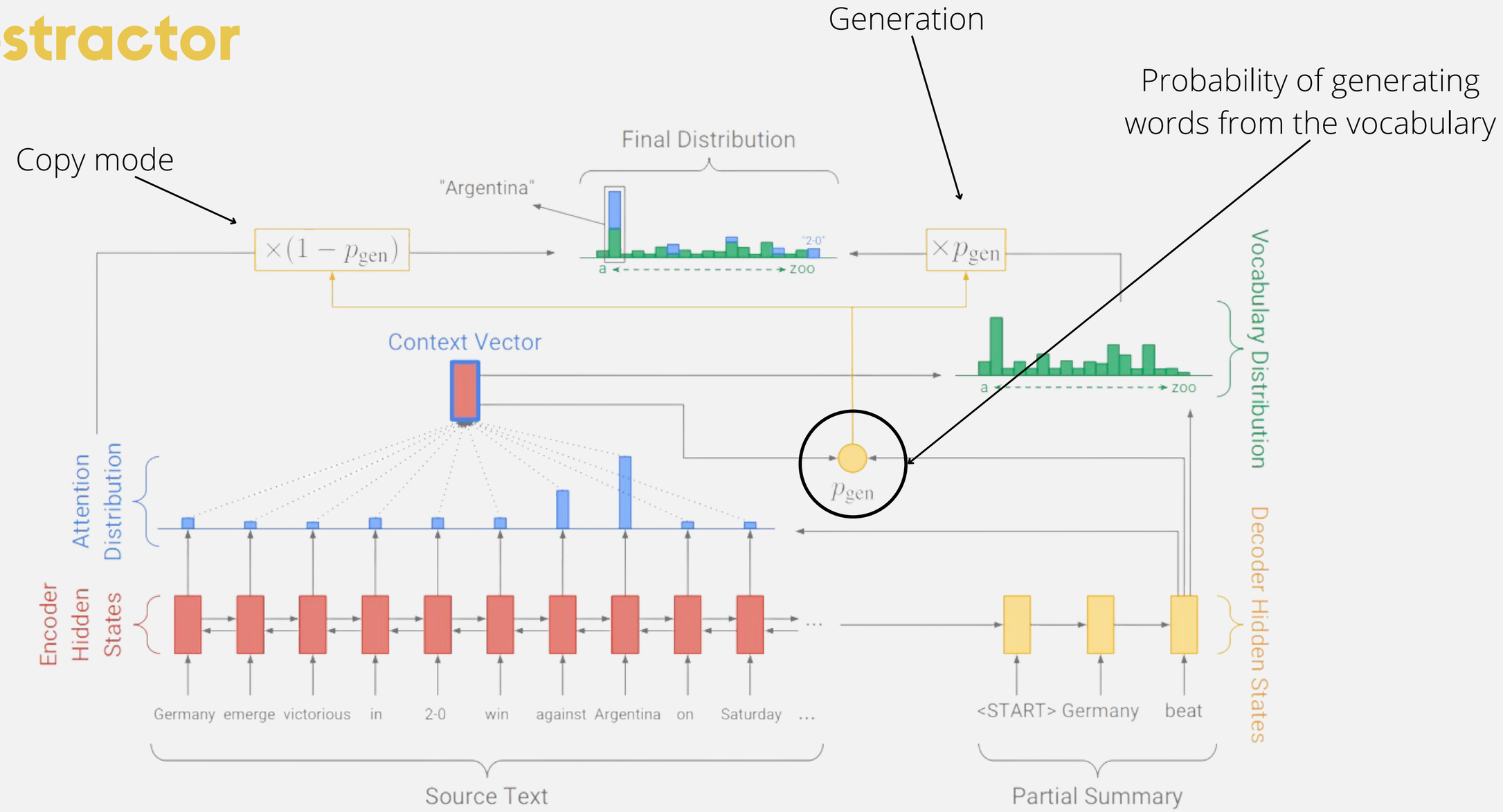
Converts W2V into sentences

2 Hop Attention Mechanism

1. Context vector from HS
2. Extraction probabilities from CV

Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting, Chen et al.

Abstractor



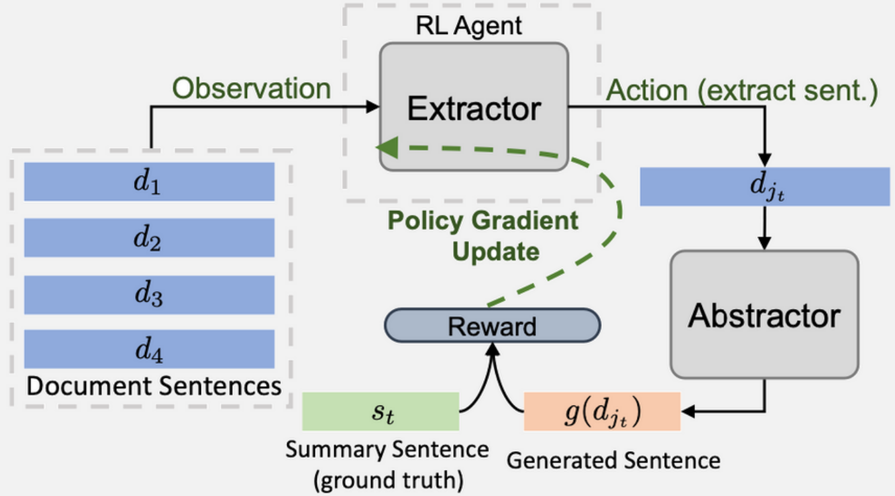
Get To The Point: Summarization with Pointer-Generator Networks, See et al.

Reinforcement Learning

Reinforcement Learning (RL) frameworks use an agent that interacts with a stochastic environment, based on the Markov Decision Process.

A2C policy, based on Rouge-L F1 rewarding function.

- Actor
- Critic



$$r(t+1) = ROUGE_L F1(g(d_{j_t}), s_t)$$

Encourage

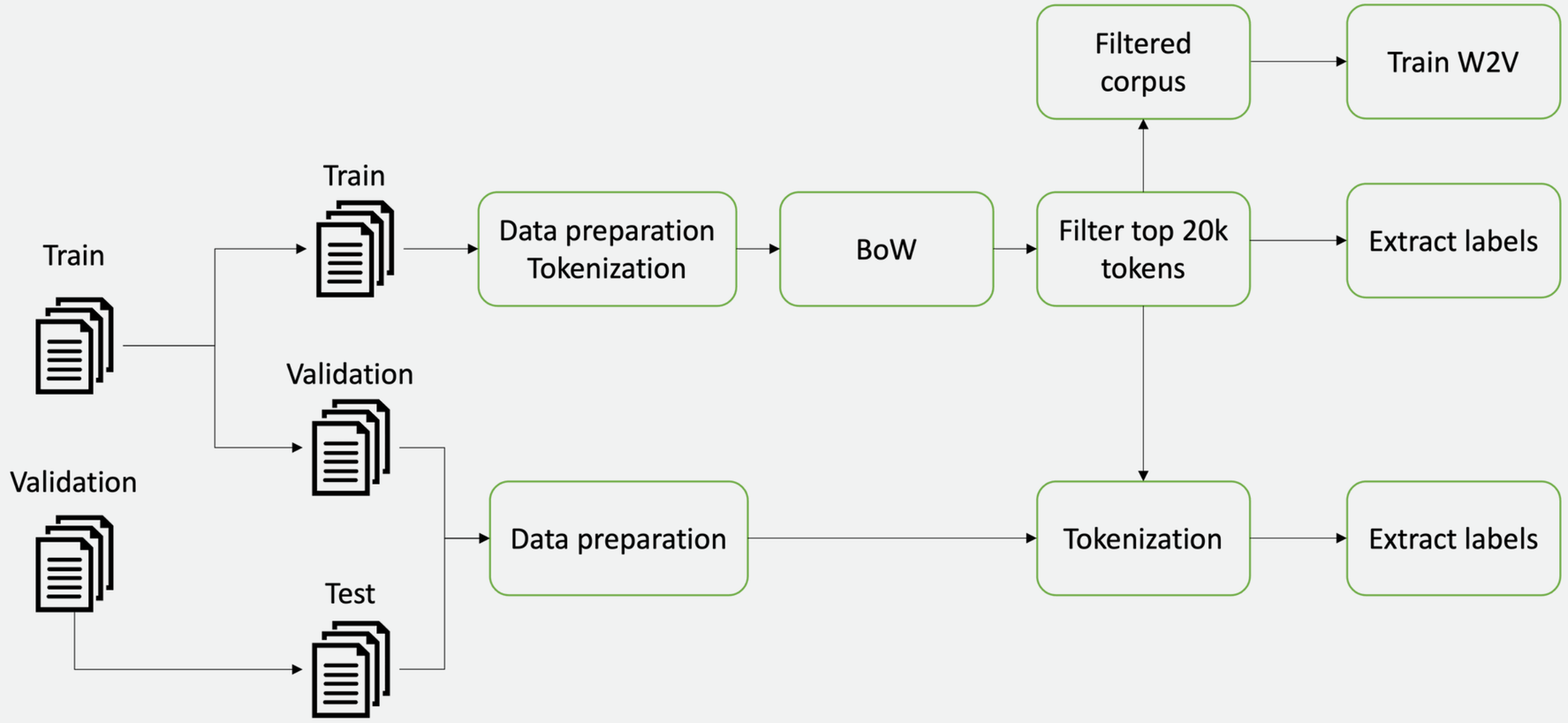
Discourage

Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting, Chen et al.

Datasets

FNS

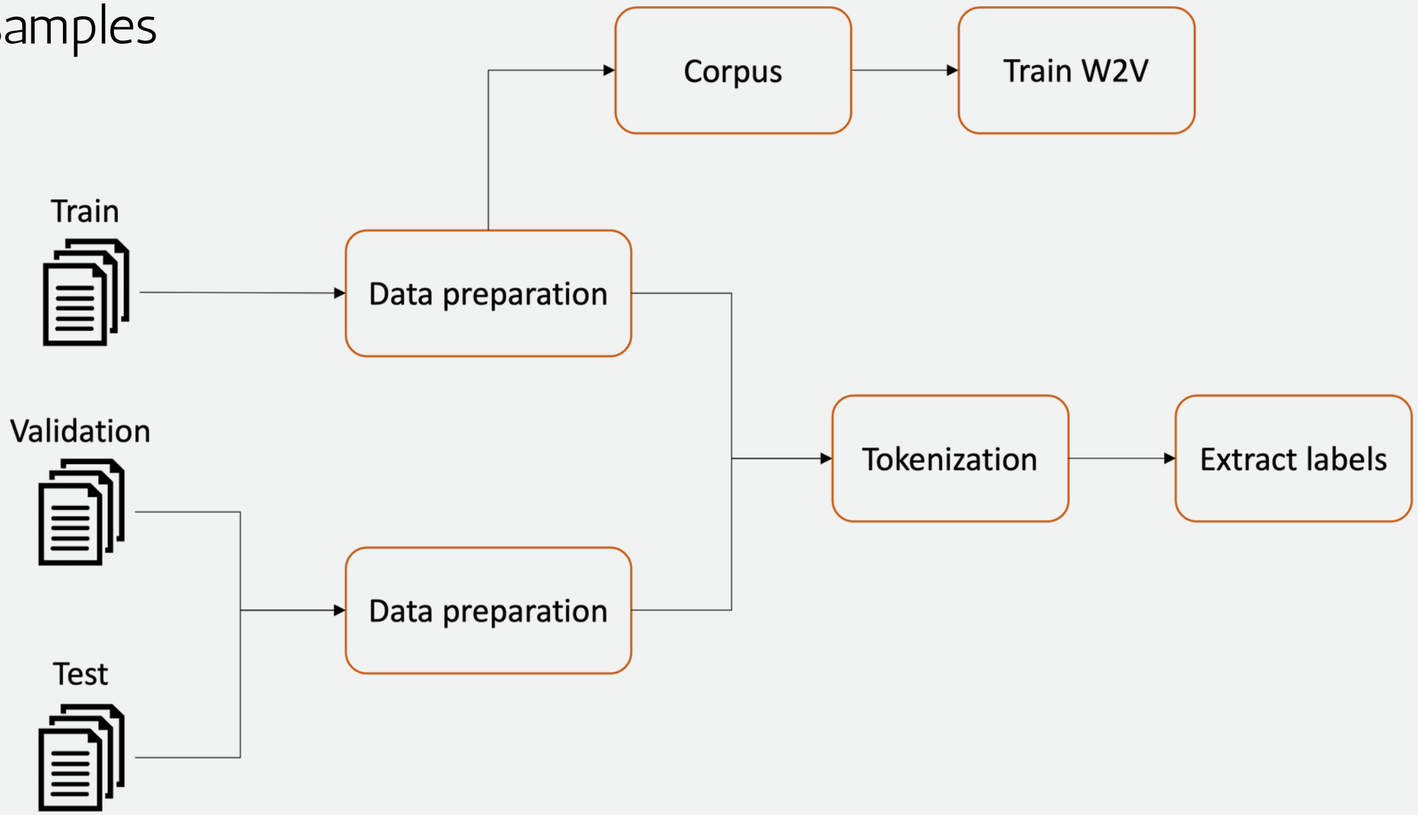
- Financial Narrative Summarization from UK firms listed on London Stock Exchange
- 3,000 training samples and
- 363 validation samples.



Datasets

CNN - DAILY MAIL

- Human generated highlights from news stories in CNN and Daily Mail websites.
- 10,000 training subsamples
- 1,000 validation subsamples
- 1,000 test subsamples



ROUGE score

Compare consequent tokens between human-generated and machine-generated summaries

Rouge-L: longest common subsequence

👎 Main weak points

- It only relies on syntactical matches
- It is not semantic aware

BERT score

Compute semantic similarity between tokens of reference and hypothesis.

Pros

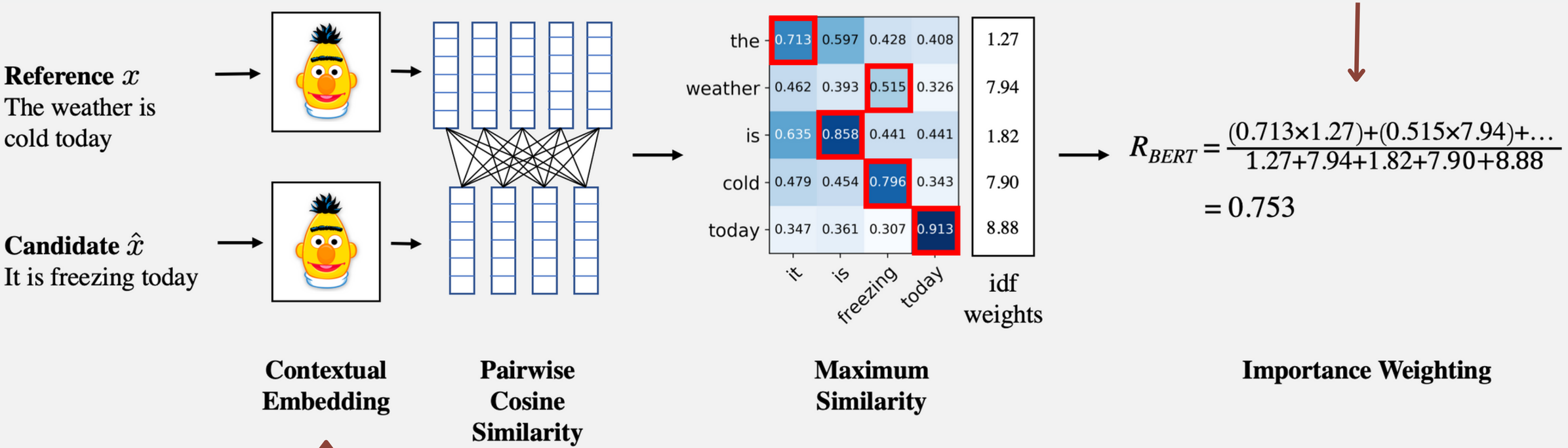
- High performances
- More accurate evaluation of context

Cons

- Computational expensive
- Not all languages are supported

BERT score

Rare words can be more indicative other than common words



BERTScore: Evaluating Text Generation with BERT, Zhang et al.

Results

	FNS		DailyCNN	
Split	Large	Small	Large	Small
Train	2,550	300	10,000	1,200
Val	450	50	1,000	100
Test	363	50	1,000	150

Large pipeline

- Full datasets
- Assess model performances on different domains

Small pipeline

- Random subsample
- Asses model performances with different metrics

Large pipeline

Rouge-L F1 score

FNS		DailyCNN	
Only Extractor	Full pipeline	Only Extractor	Full pipeline
0.36	0.38	0.20	0.23

- Comparable results between the extractor only and the full pipeline
- Drop in performances between datasets
  - Less summary sentences in CNN/DailyMail
  - Structure of the summaries
  - Broader number of topics

Small pipeline

	FNS		DailyCNN	
Extracted labels & RL Policy	**Rouge-L**	**BERT score**	**Rouge-L**	**BERT score**
Rouge-L	0.27	0.80	0.09	0.78
BERT score	0.26	0.81	0.10	0.78

- Comparable differences between the chosen metrics
- Comparable BERTScore between datasets

Conclusions

- Rouge should be used only for having a measure of overlap

- BERTScore demonstrated to be highly flexible and robust
  - Different domains
  - Different types of provided summaries

- Future works
  - Evaluate BERTScore on the entire datasets
  - Test the pipeline on new domains
  - Include new languages other than English
  - Evaluate different RL policies

Thank you for your attention !