

# Text Summarization

CPE 393: Text Analytics

***Dr. Sansiri Tarnpradab***

*Department of Computer Engineering  
King Mongkut's University of Technology Thonburi*

*Intro*

*Pattern  
Matching*

*Text  
Visualization*

*Web Scraping*

*Text  
Preparation*

*Text Feature  
Representation*

*Text  
Classification*

*Text  
Summarization*

*Text  
Clustering*

*Topic  
Modeling*

*TBA  
Advanced Topic*

???



# *Summary*

(noun.)

A short, clear description that gives the main facts or ideas about something.

# *Outline*

- Significance & Benefits
- Types of summaries
- Types of summarization
  - Extractive
  - Abstractive
- Methods
- Evaluation

# Is *Summarization* important?

## -----*Necessities*

- To understand a long document
- To find an answer to a question
- To engage in the discussion

## -----*Benefits*

- Grasp main ideas quickly
- Better understand the so-far discussion
- Save time

# Types of Summaries

## -----Single Document Summary

- Summary of one document
- Summary of a single product review [[ACL ref](#)]

## -----Multi-Document Summary

- Summary from multiple documents
- Written about the same topic
- [Newsblaster](#), Customer reviews, Forum threads

## -----Query Focused Summary

- Summary based on a specific query
- More user control and personalization
- [AnswerSumm](#) (StackExchange forums),  
[WikiHowQA](#) (WikiHow)

# Types of Summarization

## -----Extractive

**Extract existing words and phrases as-is**

- Take a page of text
- Mark the most important sentences
- Select sentences with higher ranks
- Put together



## -----Abstractive

**Paraphrase to generate a shorter version**

- Take a page of text
- Understand context
- Creates words and phrases
- Puts together in a meaningful way



# Extractive Summarization



# Methods

Representation, Representation, Representation ...

## -----Centrality-based

- PageRank
- [MEAD](#)
- More...

## -----Graph-based

- LexRank
- TextRank
- More...

## -----ML/DL

- Traditional methods
- Modern methods

## -----Others

- Lead-3: produces the leading 3 sentences of the document as its summary
- [NLP progress](#)
- [What have we achieved on Text Summarization?](#)

# Ranking

## LexRank & TextRank

### LexRank

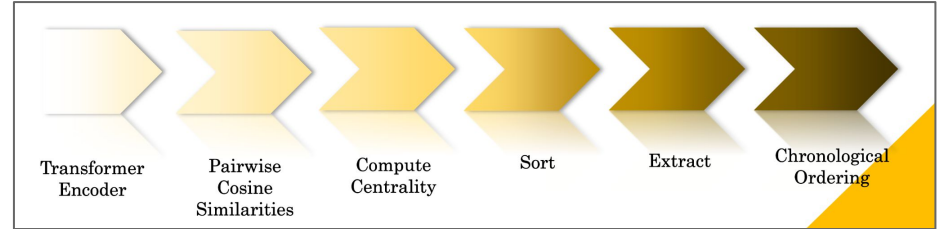
- Unsupervised approach based on graph-based centrality scoring of sentences
- **Idea:** if one sentence is very similar to many others, it will likely be an important sentence
  - High rank → Summary-worthy
- [Python library](#) (`lexrank`)

### TextRank

- Unsupervised key text units selection method
- Graph-based ranking model
- Co-similarity between sentences is computed
- Result = weight graph
  - High weighted sentence → Summary worthy
- [Python library](#) (`summa`)

# DL - Transformer

- Transformer applied to encode sentence
  - sentence embedding
  - [SBERT](#): sentence transformers



- Incorporate LexRank
  - Compute pair-wise cosine similarities
  - Compute centrality for each sentence
  - Sort sentences based on their scores
  - Extract only 20% [compression ratio](#)
  - Order summary sentences based on their chronological order

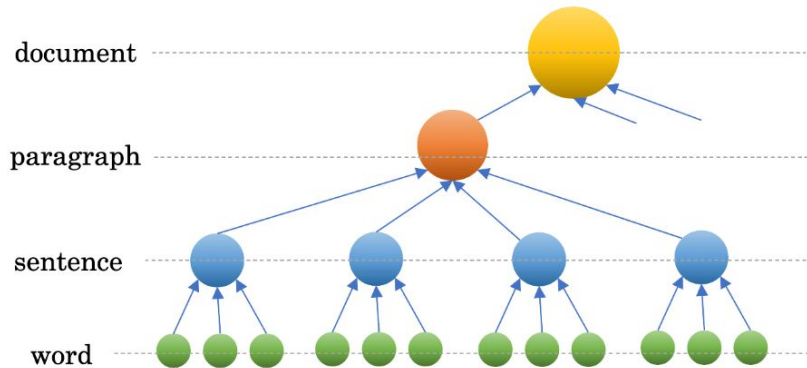
# Online Forum Summarization

The Excerpt

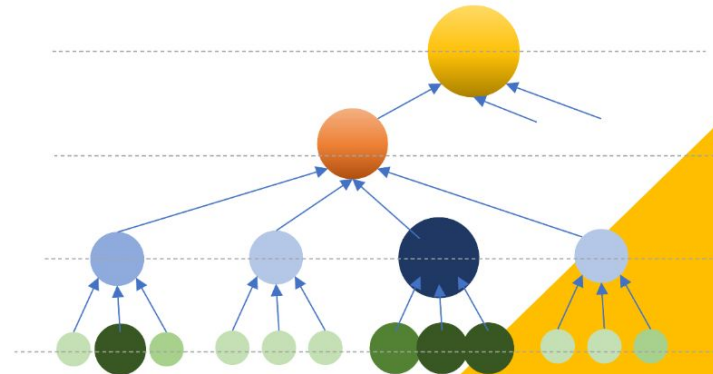
Full paper can be downloaded [here](#)

# Key Components

## Hierarchy of Data



## Attention Mechanism



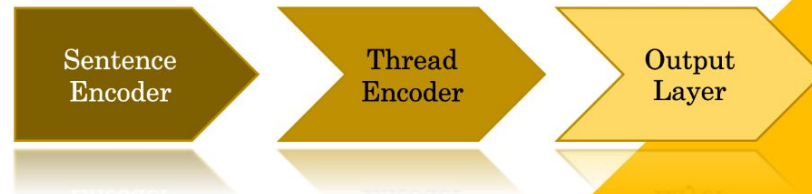
# Proposed Approach

- Hierarchical Attention Networks\*

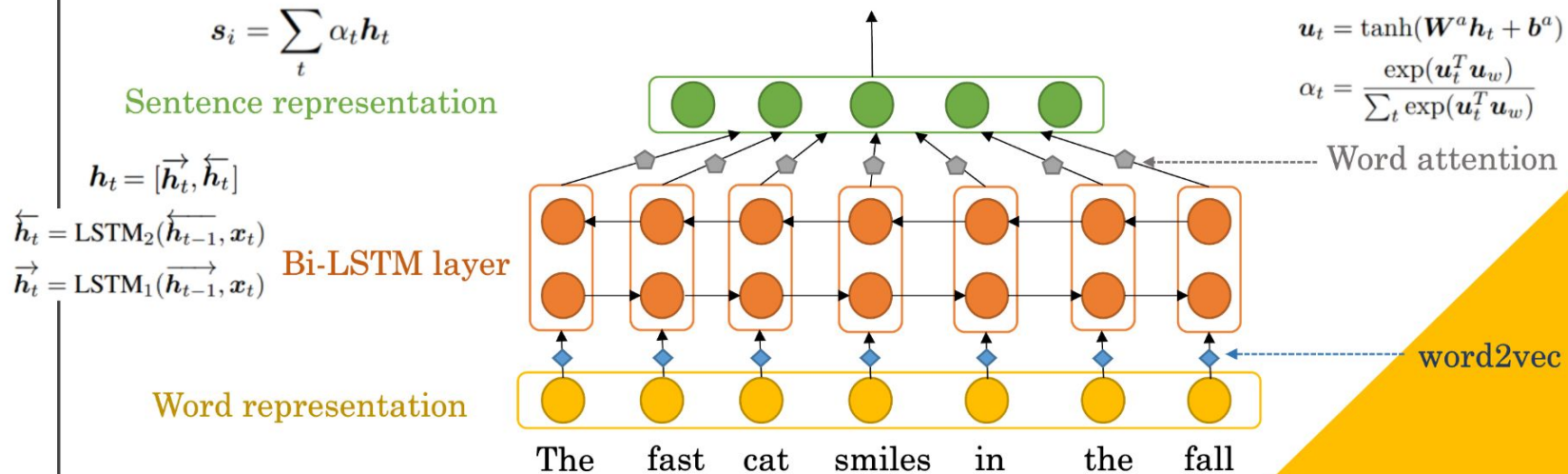
- Context information
- Attention mechanism

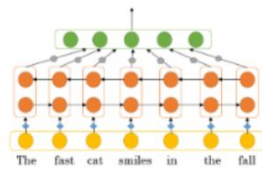
- Components

- Sentence Encoder
- Thread Encoder
- Output Layer



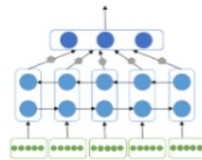
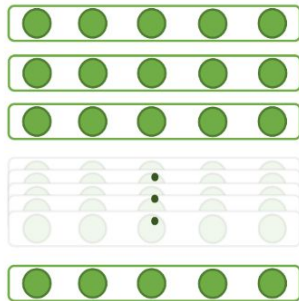
# Sentence Encoder





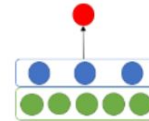
Sentence  
Encoder

Sentence representations



Thread  
Encoder

Thread representation



Output  
Layer

Binary Classification





# Abstractive Summarization

# Comparison of Output

China's Huawei overtook Samsung Electronics as the world's biggest seller of mobile phones in the second quarter of 2020, shipping 55.8 million devices compared to Samsung's 53.7 million, according to data from research firm Canalys. While Huawei's sales fell 5 per cent from the same quarter a year earlier, South Korea's Samsung posted a bigger drop of 30 per cent, owing to disruption from the coronavirus in key markets such as Brazil, the United States and Europe, Canalys said. Huawei's overseas shipments fell 27 per cent in Q2 from a year earlier, but the company increased its dominance of the China market which has been faster to recover from COVID-19 and where it now sells over 70 per cent of its phones. "Our business has demonstrated exceptional resilience in these difficult times," a Huawei spokesman said. "Amidst a period of unprecedented global economic slowdown and challenges, we've continued to grow and further our leadership position." Nevertheless, Huawei's position as number one seller may prove short-lived once other markets recover given it is mainly due to economic disruption, a senior Huawei employee with knowledge of the matter told Reuters. Apple is due to release its Q2 iPhone shipment data on Friday.

Input

While Huawei's sales fell 5 per cent from the same quarter a year earlier, South Korea's Samsung posted a bigger drop of 30 per cent, owing to disruption from the coronavirus in key markets such as Brazil, the United States and Europe, Canalys said. Huawei's overseas shipments fell 27 per cent in Q2 from a year earlier, but the company increased its dominance of the China market which has been faster to recover from COVID-19 and where it now sells over 70 per cent of its phones.

Extractive Output

Huawei overtakes Samsung as world's biggest seller of mobile phones in the second quarter of 2020. Sales of Huawei's 55.8 million devices compared to 53.7 million for south Korea's Samsung. Shipments overseas fell 27 per cent in Q2 from a year earlier, but company increased its dominance of the china market. Position as number one seller may prove short-lived once other markets recover, a senior Huawei employee says.

Abstractive Output



Advice/opinions?

An outstanding way to understand the European immigrants experiences as they arrived.

The ferry will stop at the Statue of Liberty, if you want to see that.

Basically, the 'exhibit' consists of some of the original buildings with displays giving the history of the Ellis Island facility.

Frommers is really good.

You can poke around hotels, attractions, and restaurant summaries from there.

<http://www.frommers.com/destinations/newyorkcity/A23955.html>"

Go take a photo of the statue of Liberty, I would not bother going inside, you can't go all the way up now and the security queue is a nightmare, it is tighter than some airports - seriously.

I guess if I was American I would be very keen to see Ellis but I am now thinking of giving this one a miss.

Am I able to grab a photo of Liberty with all of her in the photo or is this impossible - due to extreme size?

Kangaroo!

We have just come back from visiting New York, and went on the circle line sightseeing cruise which was an excellent way of viewing the Statue of Liberty.

It's possible.

Sold!

Thanks!

Just a quick dumb question :.

Is SOL on Ellis Island or a different island?

The island the Statue of Liberty is on and Ellis Island are different islands.

Just to be clear the sightseeing boat will take you from Manhattan Island to Liberty Island (SOL) and Ellis Island and costs a fee.

# Flaws of Extractive Summarization

An outstanding way to understand the European immigrants experiences as they arrived. The ferry will stop at the Statue of Liberty, if you want to see that. Basically, the 'exhibit' consists of some of the original buildings with displays giving the history of the Ellis Island facility. We have just come back from visiting New York, and went on the circle line sightseeing cruise which was an excellent way of viewing the Statue of Liberty. The island the Statue of Liberty is on and Ellis Island are different islands.

## Lack cohesion

- Merely a concatenation of sentences
- Does not “flow”

## Readability

- A quality of being easy or enjoyable to read
- Sentences written by different users

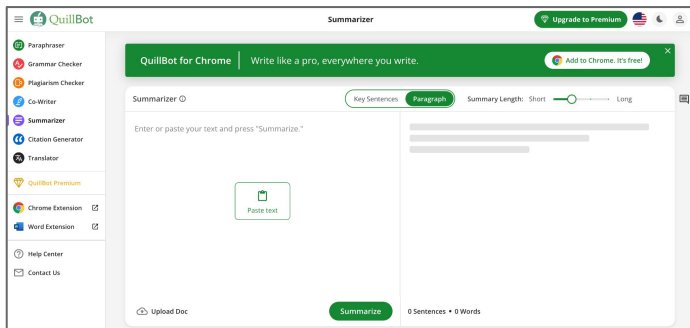
## Dangling anaphora problem

- Sentences that contain pronouns while missing their referents when extracted out of context
- “We” refers to whom?

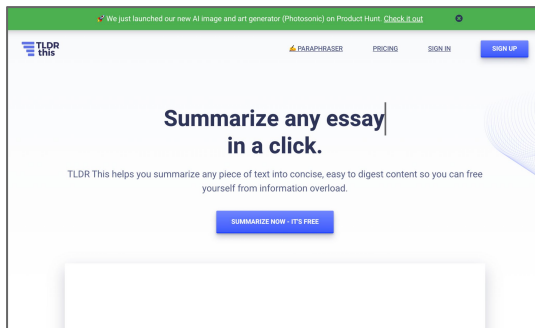
## Very different from human-written summaries

- Less overlapping when compared with reference summaries
- Low ROUGE scores

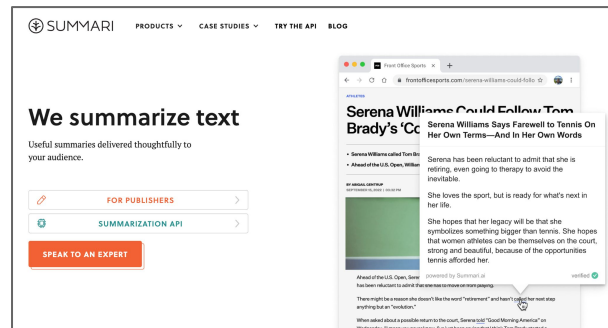
# Applications



<https://quillbot.com/>



<https://tldrthis.com/>



<https://www.summari.com/>



English (US)

German

Modes: Standard Fluency Formal Simple Creative Expand Shorten

Synonyms:



Advice/opinions?

An outstanding way to understand the European immigrants experiences as they arrived.

The ferry will stop at the Statue of Liberty, if you want to see that.

Basically, the 'exhibit' consists of some of the original buildings with displays giving the history of the Ellis Island facility.

Frommers is really good.

You can poke around hotels, attractions, and restaurant summaries from there.

<http://www.frommers.com/destinations/newyorkcity/A23955.html>"

Go take a photo of the staute of Liberty, I would not bother going inside, you cant go all the way up now and the security queue is a nightmare, it is tighter than some airports - seriously.

I guess if I was American I would be very keen to see Ellis but I am now thinking of giving this one a miss.

Am I able to grab a photo of Liberty with all of her in the photo or is this impossible - dut to extreme size?

Kangaroo!

We have just come back from visiting New York and went on the circle line

Get Premium for unlimited words. X

235/125 Words

Rephrase

Advice/opinions?

An excellent approach to comprehending the experiences of European immigrants upon arrival.

If you want to see the Statue of Liberty, the ferry will make a stop there.

The "exhibit" is actually a collection of some of the old Ellis Island buildings that have historical informational exhibits inside of them.

The travel guidebook Frommers is excellent.

From there, you can look through summaries of hotels, attractions, and restaurants.

<http://www.frommers.com/destinations/newyorkcity/A23955.html>"Go take a photo of the Statue of Liberty; I would not bother going inside; you can't go all the way up now, and the security line is a nightmare; it is tighter than some airports, truly.

I suppose if I were an American, I would be eager to see Ellis, but I'm now considering giving this one a try.



1/9 Sentences • 129 Words



# Classical Approaches

## Early Work

Full Abstractive Summarization

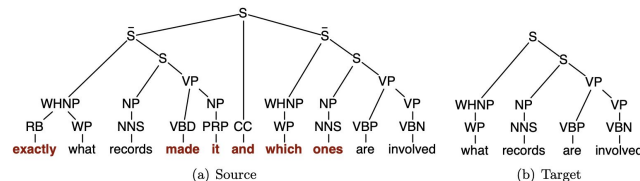
Graph-based

Template-based

## Early work

### Sentence Compression

- Tree-to-tree transduction method (rewriting)
- Given a source sentence of words, a target compression is formed by removing any subset of these words



### Sentence Fusion

- Given a set of similar sentences, produces a new sentence containing the information common to most sentences.
- Generation is performed by reusing and altering phrases from input sentences.

### Sentence Revision

- Generates sentences not found in the input and synthesizes information across sentences
- Insertion and substitution of the phrases

# Classical Approaches

Early Work

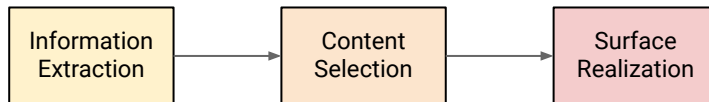
Fully Abstractive Summarization

Graph-based

Template-based

## Fully Abstractive Summarization

- Methods in the early work offer little improvement over extractive methods
- Contains three pipeline subtasks



- **Information extraction**
  - Extract important information from the input text.
  - Phrasal-level information such as noun phrases (NPs) and verb phrases (VPs) together with their contextual information
- **Content selection**
  - Select a subset of the candidate phrases extracted from the information extraction step
- **Surface realization**
  - Combine the candidates selected in content selection using grammatical/syntactic rules to generate a summary.



# Classical Approaches

Early Work

Full Abstractive Summarization

Graph-based

Template-based

## Graph-based Methods

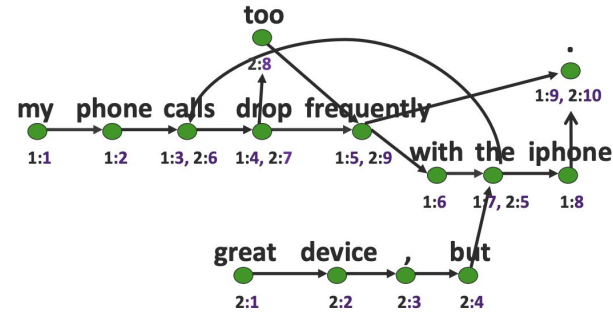
### Word graph

- Encode sentences into a graph (weighted directed graph), where each node is a word
- Incrementally add sentences to the graph
- Words that share similarities are mapped onto the same existing node
- Summary = the best path in the graph (e.g. ranking, scoring function, etc)

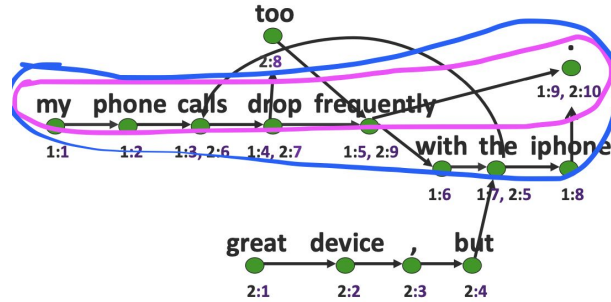
### Opinosis

- Goal: to generate concise abstractive summaries of *highly redundant* opinions
- Presentation shared by the author
- Step 1: build a graph
- Step 2: Generate candidate summaries
- Step 3: Obtain final summary sentence

1. My phone calls drop frequently with the iPhone.
2. Great device, but the calls drop too frequently.

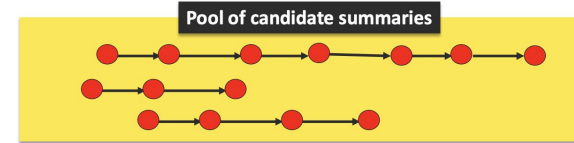


Step 1: Build a graph



Step 2: Generate candidate summaries

- Explore a pool of candidate summaries
- Each must have valid start & end node
- Score candidate summaries



Step 3: Final summary

Select top 2 scoring candidates that are most dissimilar

# Neural Approaches

## Encoder-Decoder Framework

### Improvements to the Encoder-Decoder Framework

## Encoder-Decoder Framework

- seq2seq models employ the encoder-decoder architecture
- **Encoder** encodes source sentences as a list of fixed-length vector representations
- **Decoder** outputs a summary based on the encoded vectors
- Jointly trained on document-summary pairs
- Examples: (Encoder-Decoder)
  - CNN → Feed-forward neural network (2015)
  - RNN → Feed-forward neural network (2016)
  - LSTM → Feed-forward neural network (2016, 2017)
  - GRU → Feed-forward neural network (2016, 2017)
  - [...] → RNN (2015, 2016, 2017)

# Neural Approaches

Encoder-Decoder Framework

Improvements to the Encoder-Decoder  
Framework

## Attention

- Some words/phrases are more important than others
- Attention is used to compute a weight for each element in each timestep indicating its importance
- The resulting weight distribution over the elements can be used to compute the **context vector**

## Distraction

- Some words are overly highlighted, leading to redundancy in the summary
- Subtract the history context vector from the current context vector to distract the network from content that has been attended to previously

## Pointer networks

- Favor rare words and OOV words
- Can be seen as an extension of attention that allows us to focus on those rare or OOV words that are important.

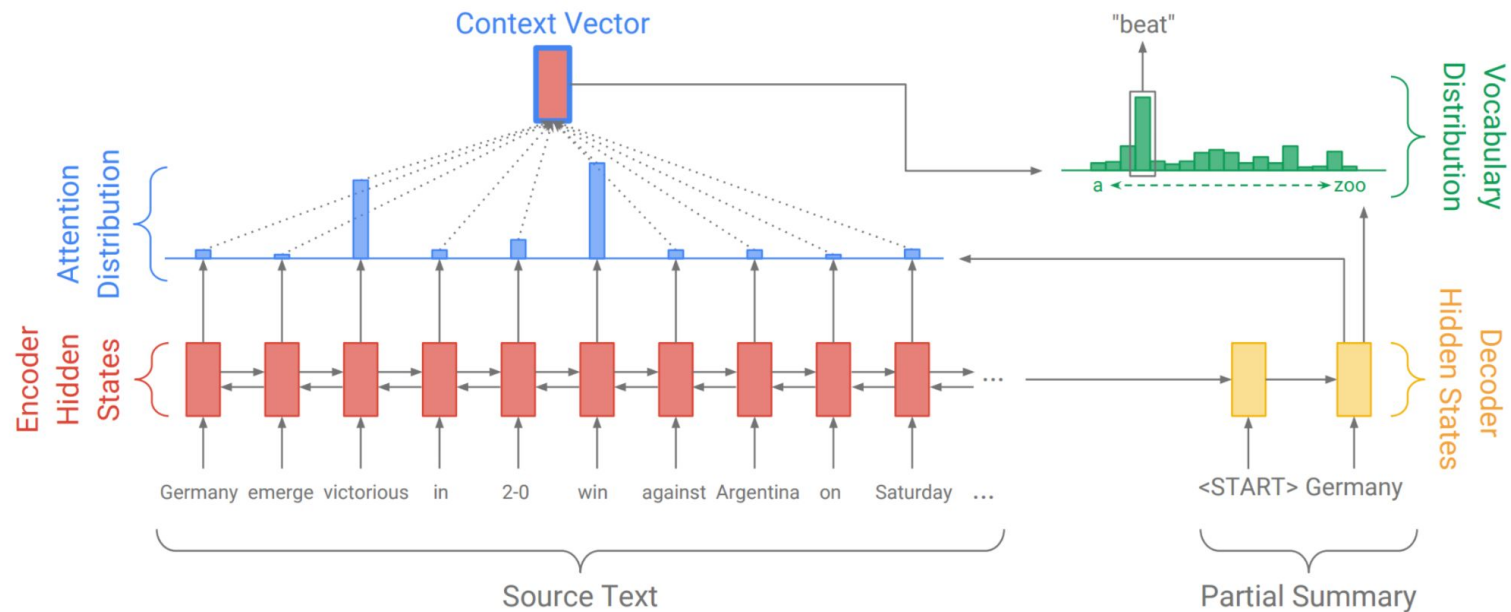


Figure from: See, Abigail, Peter J. Liu, and Christopher D. Manning. "[Get to the point: Summarization with pointer-generator networks.](#)"

### Generation Probability [0,1]

$P_{\text{gen}}$  = probability of generating a word from the vocabulary, versus copying a word from the source.

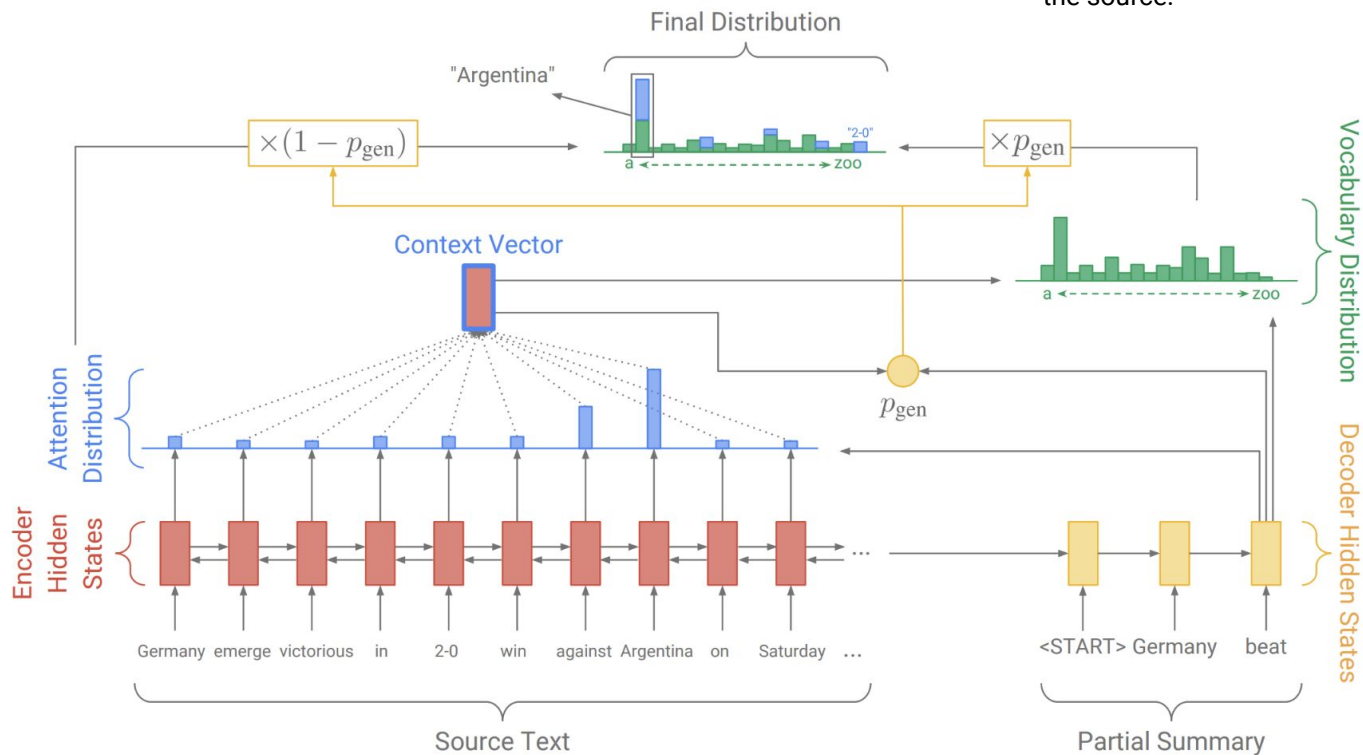
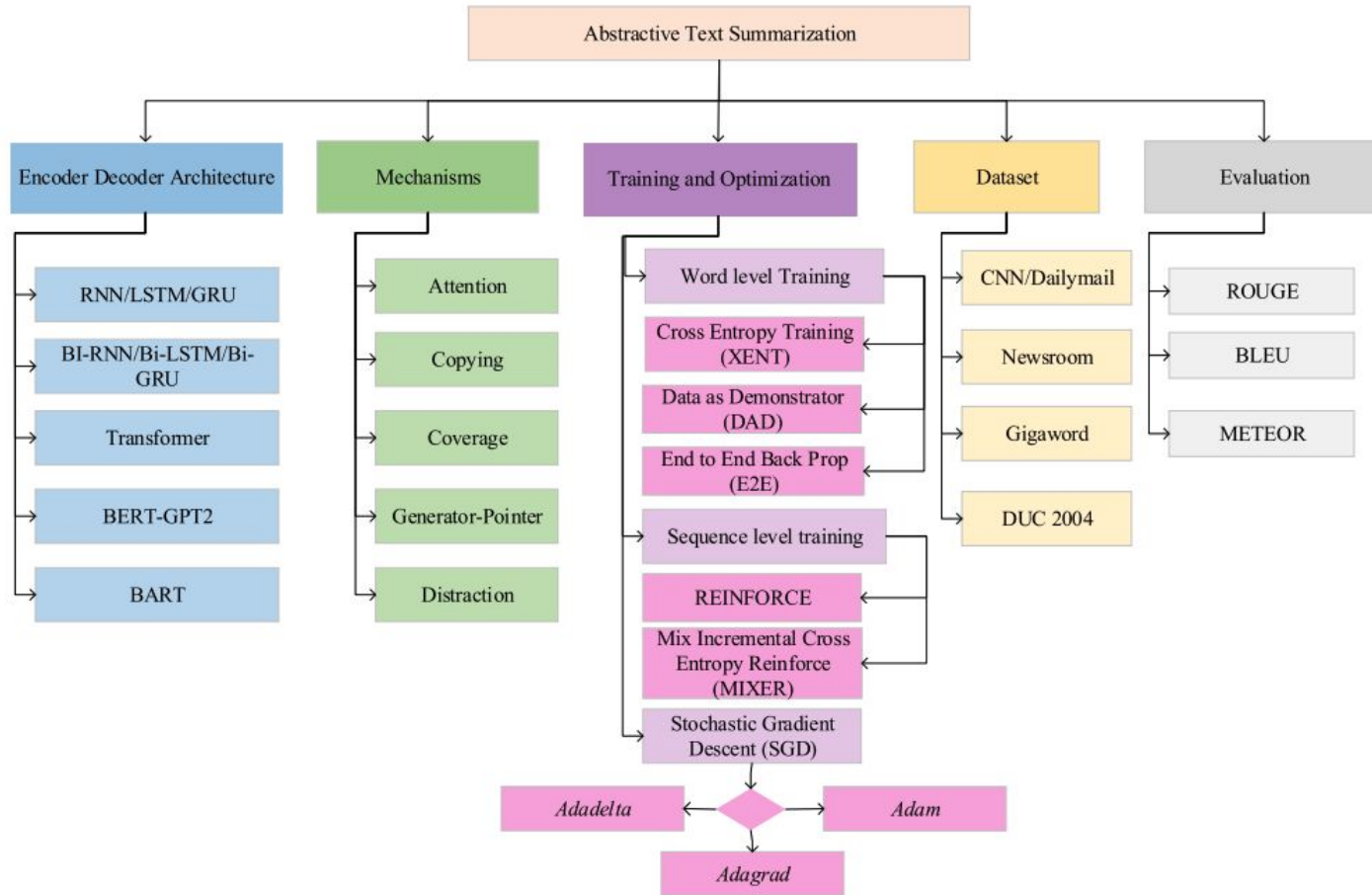


Figure from: See, Abigail, Peter J. Liu, and Christopher D. Manning. "[Get to the point: Summarization with pointer-generator networks.](#)"

# Text *Generation* via Transformer

## Transformer

- Language → Sequential
- Recurrent Neural Networks
  - Not performing well with long document
  - It forgets...
  - Cannot parallel well
  - Throw in more GPUs
- (2017) **Google x U Toronto**
  - A model that *does not forget*
  - A model that supports *parallelization*
  - Attention is all you need
- Positional encoding
- Self-attention
- Masked Language Modeling
- Next Sentence Prediction objectives
- Examples:
  - BERT
  - GPT-3
  - T5



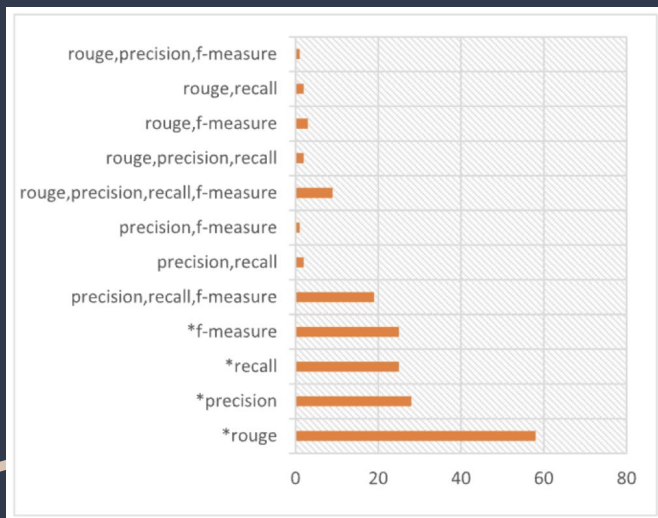
**FIGURE 7. Abstractive text summarization research framework position.**

Figure from: [A. Ayub Syed et al.: Survey of the State-of-the-Art Models in NATS](#)



# Evaluation

# Evaluate a summary



## -----Ground Truth Creation

- Annotators
- Summary creation
  - Direct extraction or in their own words

## -----Automatic Evaluation

In terms of **extracting sentences** → how many ideal sentences are in an automatic machine summary

- Precision
- Recall
- F-score/F-measure

In terms of **content** → compare the actual words in a sentence, not the whole sentence

- **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation)
- BLEU (BiLingual Evaluation Understudy)
- METEOR (Metric for Evaluation of Translation with Explicit ORdering)
- CR (Compression Ratio)
- Copyrate

# ROUGE Metrics



the cat was found under the bed



the cat was under the bed

- Compare **machine-generated** summaries against **human-produced** summaries
  - Machine-generated → system summaries
  - Human-produced → reference summaries
- **Overlapping**
  - **n-grams**: ROUGE-N where  $N = \{1, 2, 3\}$
  - **Subsequences**: ROUGE-L
  - **skip-bigrams**: ROUGE-S
  - Precision, Recall, F-scores
- Download [ROUGE paper](#)

## Precision

$$\frac{\text{No. overlapping words}}{\text{Total words in system summary}}$$

## Recall

$$\frac{\text{No. overlapping words}}{\text{Total words in reference summary}}$$

## F-scores

$$2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

# ROUGE-N

 the cat was found under the bed

 the cat was under the bed

## Precision



$$\frac{\text{No. overlapping words}}{\text{Total words in system summary}}$$

## Recall

$$\frac{\text{No. overlapping words}}{\text{Total words in reference summary}}$$

## F-scores

$$2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

|words in system summary | = 7  
|words in reference summary | = 6

## R-1 → Overlapping unigrams

{the, cat, was, under, the, bed}

- **Precision:**  $6/7 = 0.86$ 
  - 6/7 words in the system summary were relevant
- **Recall:**  $6/6 = 1.0$ 
  - All the words in the reference summary have been captured by the system summary
- **F-scores:**  $2(0.86 \cdot 1)/(0.86 + 1)$

## R-2 → Overlapping bigrams



Bigrams:

{the cat, cat was, was found, found under, under the, the bed}



Bigrams: {the cat, cat was, was under, under the, the bed}

Overlapping Bigrams: {the cat, cat was, under the, the bed}

# ROUGE-L

 the cat was found under the bed

 the cat was under the bed

## Precision

$$\frac{\text{No. overlapping words}}{\text{Total words in system summary}}$$

## Recall


$$\frac{\text{No. overlapping words}}{\text{Total words in reference summary}}$$

## F-scores

$$2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Longest common subsequence (LCS) or longest co-occurring in sequence n-grams

- Not necessarily consecutive
- The longer shared sequence, the more similarity

|words in system summary | = 7

|words in reference summary | = 6

---

  LCS: {the cat was under the bed} ∴ #overlap = 6

- **Precision:**  $6/7 = 0.86$
- **Recall:**  $6/6 = 1.0$
- **F-scores:**  $2(0.86 \cdot 1)/(0.86 + 1)$

# ROUGE

## Summary

- ROUGE tells how good the machine-generated (system) summaries are compared to one or more human-produced (reference) summaries
- Generally, at least 2 reference summaries are used
  - First, compute max of pairwise summary-level ROUGE-N
  - Then, compute average of all ROUGE-N scores
- **Pros:**
  - Correlates with human evaluation
  - Easy to compute
  - Language-independent
- **Cons:**
  - Only focus on syntactical matches, not semantics
- **Resources**
  - [ROUGE paper](#)
  - [Hugging Face library](#)
  - [Python library](#)
  - [Matlab function](#)
  - [Java package](#)

# *Conclusion*

- Significance & Benefits
- Types of summaries
- Types of summarization
  - Extractive
  - Abstractive
- Methods
- Evaluation

Q & A