Natural Language Understanding, Generation, and Machine Translation

Lecture 24 and 25: Natural Language Generation: Summarization

Laura Perez-Beltrachini

(Slides by Mirella Lapata)

School of Informatics
University of Edinburgh
mlap@inf.ed.ac.uk

Outline

1. Introduction
   - What is NLG?
   - Why is it Useful?
2. Modeling Approach
   - Conditional Language Modeling
   - Task Definition
   - Sequence-to-Sequence Attentional Model
   - Summarization with Pre-Trained Encoders
3. Evaluation
   - ROUGE Metric
   - Results
4. Discussion

What is NLG?

(non-)linguistic input ⟹

⟹ text

$\underbrace{\hspace{5cm}}$

databases
news articles
log files
images

$\underbrace{\hspace{4cm}}$

reports
help messages
summaries
captions

You've been doing NLG all along with Machine Translation!

Why is it Useful?

Facilitates information access:

- A lot of data is in non-textual format
- Even textual data can be difficult to read
- People more prone to understand texts than numbers or graphs (Law et al., 2005)

Most NLP applications operate over texts:

- Search engines
- Question answering systems
- Speech synthesizers

Why is it Useful?

Stock data						
04/10/96	103	101.25	101.625	32444	-74	5485
04/09/96	104	101.5	101.625	41839	-33	5560
04/08/96	103.875	101.875	103.75	46096	-88	5594
04/05/96	Holiday					
04/04/96	104.875	103.5	104.375	18101	-6	5682

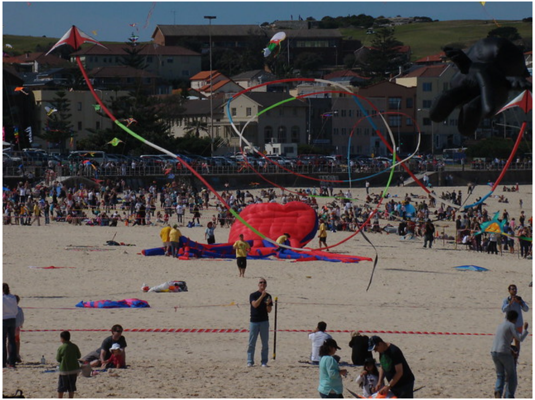
Microsoft avoided the downwards trend of the Dow Jones average today. Confined trading by all investors occurred today. After shooting to a high of $104.87, its highest price so far for the month of April, Microsoft stock eased to finish at an enormous $104.37. The Dow closed after trading at a weak 5682, down 6 points.

Why is it Useful?

Team Stat Comparison		
1st Downs	19	22
Total Yards	338	379
Passing	246	306
Rushing	92	73
Penalties	16-149	7-46
3rd Down Conversions	4-13	6-16
4th Down Conversions	0-0	0-1
Turnovers	2	0
Possession	27:40	32:20

The New England Patriots lost two linebackers and two coaches in the offseason. They still know how to win thanks in large part to two stars they didn’t lose. Tom Brady threw for 306 years and two touchdowns and Richard Seymour helped make a a game-turning defensive play as the Patriots opened their quest for an unprecedented third straight Super Bowl victory by beating Oakland 30–20 on Thursday night.

Why is it Useful?



a crowd of people on a beach flying kites.
a man flying kite in the middle of a crowded beach.
lots of people enjoying their time on the beach.

Why is it Useful?

Most blacks say MLK’s vision fulfilled, poll finds WASHINGTON (CNN) – More than two-thirds of African-Americans believe Martin Luther King Jr.’s vision for race relations has been fulfilled, a CNN poll found – a figure up sharply from a survey in early 2008.

The CNN-Opinion Research Corp. survey was released Monday, a federal holiday honoring the slain civil rights leader and a day before Barack Obama is to be sworn in as the first black U.S. president.

The poll found 69 percent of blacks said King’s vision has been fulfilled in the more than 45 years since his 1963 ’I have a dream’ speech – roughly double the 34 percent who agreed with that assessment in a similar poll taken last March.

But whites remain less optimistic, the survey found. ’Whites don’t feel the same way – a majority of them say that the country has not yet fulfilled King’s vision,’ CNN polling director Keating Holland said. However, the number of whites saying the dream has been fulfilled has also gone up since March, from 35 percent to 46 percent.

Highlights

- 69 percent of blacks polled say Martin Luther King Jr’s vision realized.
- Slim majority of whites say King’s vision not fulfilled.
- King gave his “I have a dream” speech in 1963.

Modeling Approach

A **Language Model** predicts next word, given words predicted so far:

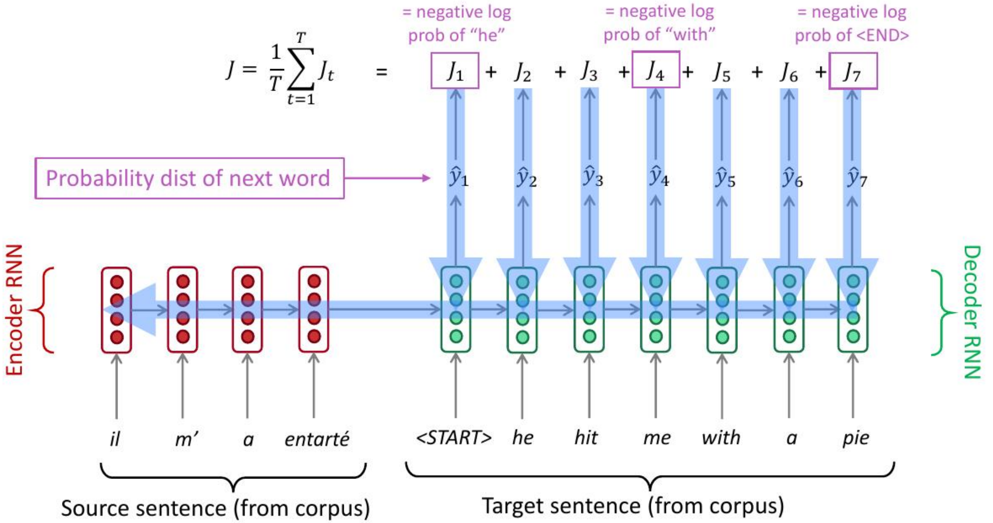
$$p(y_t | y_1, \ldots, y_{t-1})$$

A **Conditional Language Model** predicts next word, given words so far, and also some other input x:

$$p(y_t | y_1, \ldots, y_{t-1}, x)$$

- We can use an RNN to model the above probability!
- Machine Translation (x=source sentence, y=target sentence)
- Summarization (x=input text, y=summarized text)

Modeling Approach



$$J = \frac{1}{T}\sum_{t=1}^{T} J_t = \boxed{J_1} + J_2 + J_3 + \boxed{J_4} + J_5 + J_6 + \boxed{J_7}$$

= negative log prob of "he"
= negative log prob of "with"
= negative log prob of <END>
Probability dist of next word
$\hat{y}_1$ $\hat{y}_2$ $\hat{y}_3$ $\hat{y}_4$ $\hat{y}_5$ $\hat{y}_6$ $\hat{y}_7$
Encoder RNN
Decoder RNN
il
m'
a
entarté
<START>
he
hit
me
with
a
pie
Source sentence (from corpus)
Target sentence (from corpus)

Figure credit: Abigail See

Summarization: Task Definition

Given input text x, write summary y which is shorter and contains main information of x.
Summarization can be single-document or multi-document.

- **Single-document** means we summary y of single document x.
- **Multi-document** means we write a summary y of multiple documents $x_1, \ldots, x_n$

Typically $x_1, \ldots, x_n$ have overlapping content: e.g., news articles about the same event.

Summarization: Two Main Strategies

Extractive Summarization

Select parts (typically sentences) of the original text to form a summary.

- Easier
- Restrictive (no paraphrasing)

Abstractive Summarization

Generate new text using natural language generation techniques.

- More difficult
- More flexible (human-like)

The CNN/Daily Mail Dataset

- Training data consists of pairs of news articles (average 800 words) and summaries (aka story highlights), usually 3 or 4 sentences long (average 56 words).
- CNN 100K pairs; Daily Mail 200K pairs.
- Highlights have been written by journalists, in compressed telegraphic manner.
- Need not form a coherent summary — each highlight relatively stand-alone, little co-referencing.
- Download from `https://github.com/abisee/cnn-dailymail`

The CNN/Daily Mail Dataset

Most blacks say MLK's vision fulfilled, poll finds WASHINGTON (CNN) – More than two-thirds of African-Americans believe Martin Luther King Jr.'s vision for race relations has been fulfilled, a CNN poll found – a figure up sharply from a survey in early 2008.

The CNN-Opinion Research Corp. survey was released Monday, a federal holiday honoring the slain civil rights leader and a day before Barack Obama is to be sworn in as the first black U.S. president.

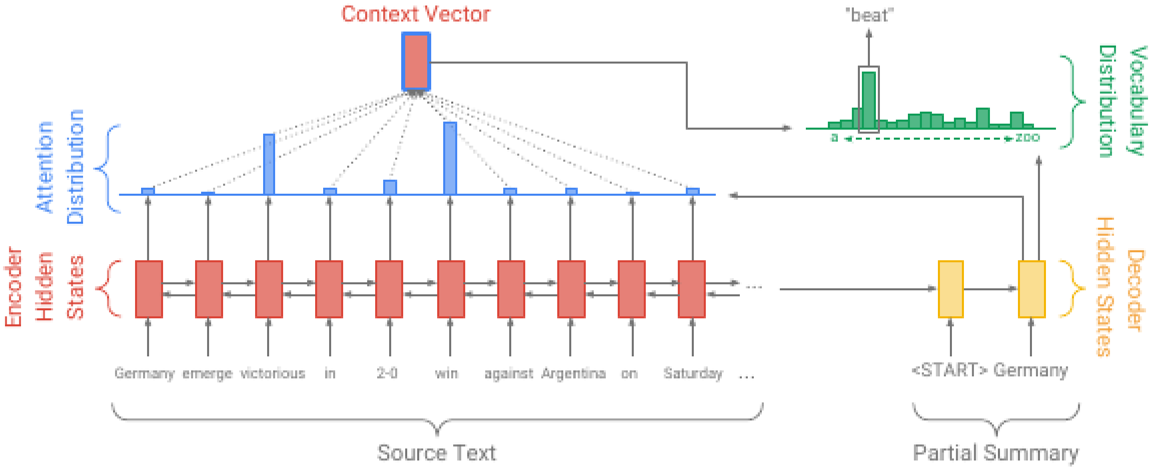
The poll found 69 percent of blacks said King's vision has been fulfilled in the more than 45 years since his 1963 'I have a dream' speech – roughly double the 34 percent who agreed with that assessment in a similar poll taken last March.

But whites remain less optimistic, the survey found. 'Whites don't feel the same way – a majority of them say that the country has not yet fulfilled King's vision,' CNN polling director Keating Holland said. However, the number of whites saying the dream has been fulfilled has also gone up since March, from 35 percent to 46 percent.

Highlights

- *69 percent of blacks polled say Martin Luther King Jr's vision realized.*
- *Slim majority of whites say King's vision not fulfilled.*
- *King gave his "I have a dream" speech in 1963.*

Sequence-to-Sequence Attentional Model



Sequence-to-Sequence Attentional Model

Encoder: single-layer bidirectional LSTM produces a sequence of hidden states h_i.

Decoder: single-layer unidirectional LSTM receives word embedding of previous word emitted by decoder and has decoder state s_t.

Attention distribution:
$$\begin{aligned} e_i^t &= v^T \tanh(W_h h_i + W_s s_t + b_{attn}) \\ a^t &= \text{softmax}(e^t) \end{aligned}$$

Context vector: weighted sum of encoder hidden states $h_i^* = \sum_i a_i^t h_i$

Vocabulary distribution: probability distribution over all words in the vocabulary, $P_{vocab} = \text{softmax}(V'(V[s_t, h_t^*] + b) + b')$

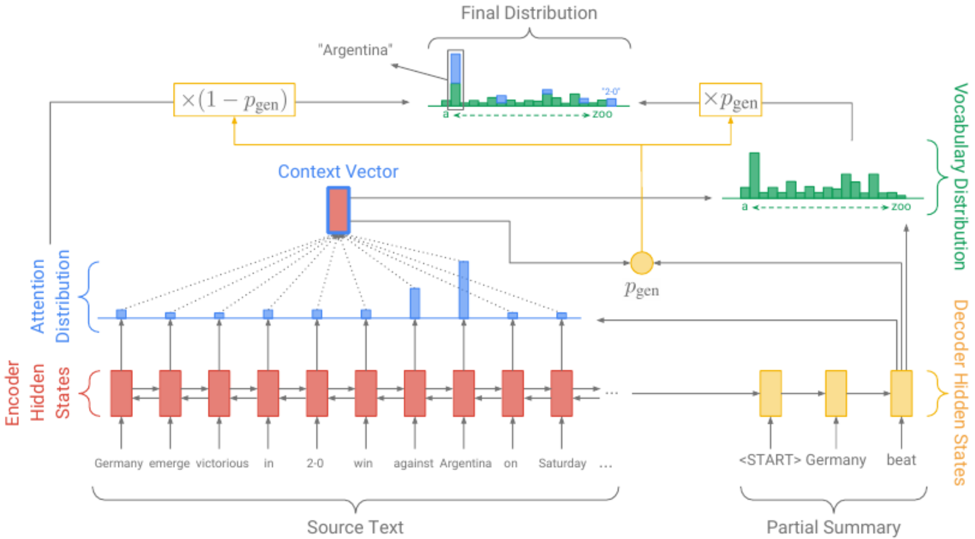
Training loss for time step t is negative loglikelihood of target word w_t^*, $\text{loss}_t = -\log P(w_t^*)$

Pointer-Generator Network

- Implements a copying mechanism (useful for rare words and phrases)
- Model allows both copying words by pointing, and generating words from a fixed vocabulary
- On each decoder step, calculate p_{gen}, probability of generating next word (rather than copying it).

$$P(w) = p_{gen} P_{vocab}(w) + (1 - p_{gen}) \sum_{i:w_i=w} a_i^t$$

Pointer-Generator Network



Coverage Mechanism

- Attempts to generate less repetitive summaries
- Penalizes repeatedly attending to same parts of the source text
- **Coverage vector** c^t tells us what has been attended so far:

$$c^t = \sum_{t'=0}^{t-1} a^{t'}$$

- Use coverage vector as extra input to attention mechanism:

$$e_i^t = v^T \tanh(W_h h_i + W_s s_t + w_c c_i^t + b_{attn})$$

- **Coverage loss** penalizes overlap between coverage vector c^t and new attention distribution a^t:

$$\text{covloss}_t = \sum_i \min(a_i^t, c_i^t)$$

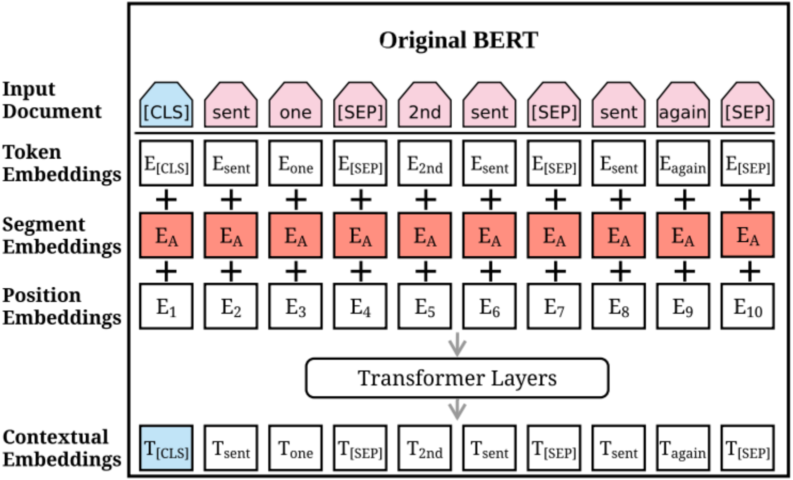
Summarization with Pre-Trained Encoders?

Pre-trained encoders like BERT very successful in many language understanding tasks:

- text classification
- textual entailment
- reading comprehension

BERT is trained on sentence level (sentences pairs), will it work on documents?

Recall BERT Model


Original BERT
Input Document
[CLS] sent one [SEP] 2nd sent [SEP] sent again [SEP]
Token Embeddings
$E_{[CLS]}$ E_{sent} E_{one} $E_{[SEP]}$ E_{2nd} E_{sent} $E_{[SEP]}$ E_{sent} E_{again} $E_{[SEP]}$
Segment Embeddings
E_A E_A E_A E_A E_A E_A E_A E_A E_A E_A
Position Embeddings
E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10}
Transformer Layers
Contextual Embeddings
$T_{[CLS]}$ T_{sent} T_{one} $T_{[SEP]}$ T_{2nd} T_{sent} $T_{[SEP]}$ T_{sent} T_{again} $T_{[SEP]}$

Challenge 1: Representation of Multiple Sentences (extractive)

Input Document

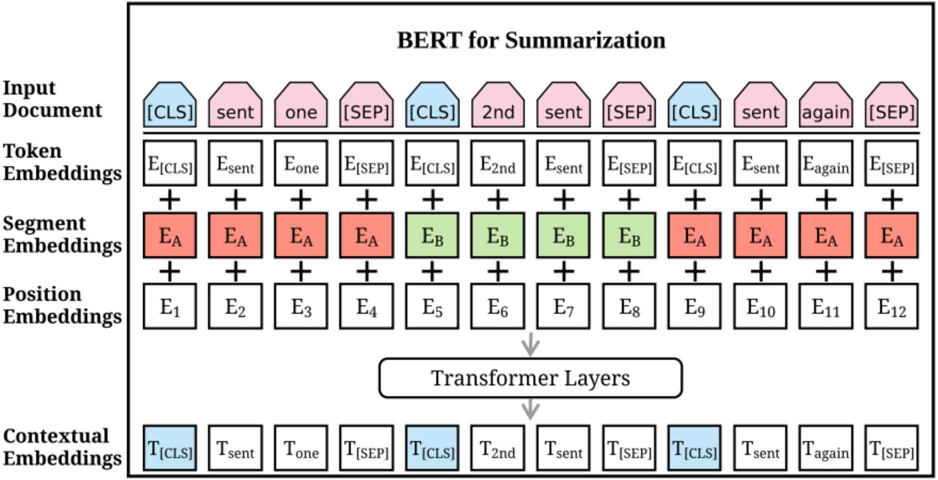
Most blacks say MLK's vision fulfilled, poll finds WASHINGTON (CNN) – More than two-thirds of African-Americans believe Martin Luther King Jr.'s vision for race relations has been fulfilled, a CNN poll found – a figure up sharply from a survey in early 2008.

The CNN-Opinion Research Corp. survey was released Monday, a federal holiday honoring the slain civil rights leader and a day before Barack Obama is to be sworn in as the first black U.S. president.

The poll found 69 percent of blacks said King's vision has been fulfilled in the more than 45 years since his 1963 'I have a dream' speech – roughly double the 34 percent who agreed with that assessment in a similar poll taken last March.

But whites remain less optimistic, the survey found. 'Whites don't feel the same way – a majority of them say that the country has not yet fulfilled King's vision,' CNN polling director Keating Holland said. However, the number of whites saying the dream has been fulfilled has also gone up since March, from 35 percent to 46 percent.

Take on Challenge 1: Modifications to the Input Format



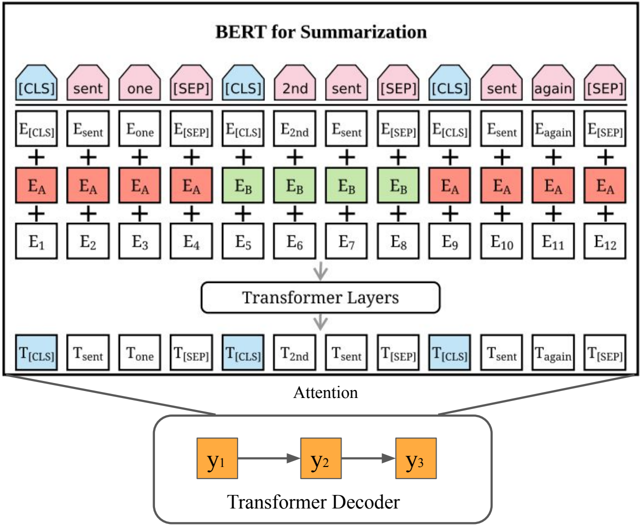
BERT for Summarization

Input Document	[CLS]	sent	one	[SEP]	[CLS]	2nd	sent	[SEP]	[CLS]	sent	again	[SEP]
Token Embeddings	$E_{[CLS]}$	E_{sent}	E_{one}	$E_{[SEP]}$	$E_{[CLS]}$	E_{2nd}	E_{sent}	$E_{[SEP]}$	$E_{[CLS]}$	E_{sent}	E_{again}	$E_{[SEP]}$
	+	+	+	+	+	+	+	+	+	+	+	+
Segment Embeddings	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_A	E_A	E_A	E_A
	+	+	+	+	+	+	+	+	+	+	+	+
Position Embeddings	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}	E_{11}	E_{12}

Transformer Layers

Contextual Embeddings	$T_{[CLS]}$	T_{sent}	T_{one}	$T_{[SEP]}$	$T_{[CLS]}$	T_{2nd}	T_{sent}	$T_{[SEP]}$	$T_{[CLS]}$	T_{sent}	T_{again}	$T_{[SEP]}$

Challenge 2: Mismatch between Encoder and Decoder (abstractive)


BERT for Summarization
[CLS] sent one [SEP] [CLS] 2nd sent [SEP] [CLS] sent again [SEP]
E[CLS] Esent Eone E[SEP] E[CLS] E2nd Esent E[SEP] E[CLS] Esent Eagain E[SEP]
EA EA EA EA EB EB EB EB EA EA EA EA
E1 E2 E3 E4 E5 E6 E7 E8 E9 E10 E11 E12
Transformer Layers
T[CLS] Tsent Tone T[SEP] T[CLS] T2nd Tsent T[SEP] T[CLS] Tsent Tagain T[SEP]
Attention
y1 y2 y3
Transformer Decoder

Leremy Stick Figure / @iconfinder.com

Take on Challenge 2: Fine-Tuning Strategy

- Learning rate schedule (Vaswani et al., 2017) :

$$lr = \tilde{lr} \cdot \min(step^{-0.5}, step \cdot warmup^{-1.5})$$

- Smaller learning rate and longer warming-up for the **encoder**:

$$\tilde{lr}_e = 2e^{-3}, warmup_e = 20,000$$

- Larger learning rate and shorter warming-up for the **decoder**:

$$\tilde{lr}_d = 0.1, warmup_d = 10,000$$

Summarization Evaluation: ROUGE

ROUGE stands for **R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation

$$\text{ROUGE-N} = \frac{\sum\limits_{S \in \{ReferenceSummaries\}} \sum\limits_{gram_n \in S} Count_{match}(gram_n)}{\sum\limits_{S \in \{ReferenceSummaries\}} \sum\limits_{gram_n \in S} Count(gram_n)}$$

- Like BLEU, it is based on n-gram overlap
- ROUGE has no brevity penalty and is based on recall
- Often F1 (combination of precision and recall) ROUGE is reported
- Most commonly-reported ROUGE scores: ROUGE-1 unigram overlap ROUGE-2 bigram overlap, and ROUGE-L longest common subsequence overlap

Results

	ROUGE		
Models	1	2	L
seq-to-seq+attn	31.33	11.81	28.83
pointer-generator	36.44	15.66	33.42
pointer-generator + coverage	39.53	17.28	36.38
lead-3 baseline	40.34	17.70	36.57
BertSumAbs	41.72	19.39	38.76

Lead-3 baseline uses first three article sentences as the summary.

Discussion

- CNN/Daily Mail dataset is rather extractive, you can go far with copy and paste operations.
- Summaries are fluent but contain factual inaccuracies!
- Do we trust ROUGE as an evaluation metric? How would you evaluate output summaries with humans?
- How would we build an extractive summarization model? How would the training data look like?

Summarization Datasets & Code

- **XWikis** (Perez-Beltrachini and Lapata, 2021), cross-lingual document-summary pairs in 4 languages derived from Wikipedia
- **MLSUM** (Scialom et al., 2020), document-summary pairs, 5 languages news outlets
- **XSum** (Narayan et al., 2018), 227K BBC articles with single-sentence summaries
- **NewsRoom** (Grusky et al., 2018), 1.3M article-summary pairs written by editors
- **arXiv**, **PubMed** (Cohan et al., 2018), abstract and paper body (113K and 215K)
- **BigPatent** (Sharma et al., 2019) 1.3 million US patents with human summaries
- **WikiHow** (Koupaee and Wang, 2018) 200K instructions with single-sentence summaries
- **Reddit TIFU** (Kim et al., 2019) 129K stories with descriptive summaries

Code: `https://paperswithcode.com/task/text-summarization/latest`