Lecture 14: Summarization

An exercise in quickly reading and experimenting

Harvard

AC295/CS287r/CSCI E-115B

Chris Tanner

SUMMARIZATION

"Imma let you finish, but lecture 13 had some of the most important info of all time" – **Kanye**

Today's slides are largely inspired from, based on, or directly from:

- **Dan Jurafsky (Stanford)**
- **Chan Young Park (CMU)**
- **Pengfei Liu (CMU)**

ANNOUNCEMENTS

- HW3 is due tonight @ 11:59pm!
- HW2 and Phase 2 are being graded
- Research Project Phase 3 due Oct 28 (Thurs) @ 11:59pm

Outline

- What is summarization?
- Data and metrics
- Workshop time
- Modern approaches

Outline

- What is summarization?
- Data and metrics
- Workshop time
- Modern approaches

What is it?

Task: produce an abridged version of a text while retaining the key, relevant information

What is it?

Useful for creating:

- outlines or abstracts of any document, article, etc
- summaries of chat and email
- action items from a meeting
- simplifying text by compressing sentences

What is it?

Input:

- single document summarization (SDS)
- multiple-document summarization (MDS)

Output:

- extractive
- abstractive

Focus:

- generic (unconditioned)
- query-focused (conditioned)

Approach:

- supervised
- unsupervised

What is it?

Single-document summarization

Given a single doc, produce:

- Abstract
- Outline
- Headline

Multiple-document summarization

Given a group of docs, produce:

- Series of news stories on the same event
- Set of web pages about some topic or question

Single-doc summarization

Document

Cambodian leader Hun Sen on Friday rejected opposition parties ' demands for talks outside the country , accusing them of trying to `` internationalize " the political crisis .
Government and opposition parties have asked King Norodom Sihanouk to host a summit meeting after a series of post-election negotiations between the two opposition groups and Hun Sen 's party to form a new government failed .
Opposition leaders Prince Norodom Ranariddh and Sam Rainsy , citing Hun Sen 's threats to arrest opposition figures after two alleged attempts on his life , said they could not negotiate freely in Cambodia and called for talks at Sihanouk 's residence in Beijing .Hun Sen , however , rejected that .``
I would like to make it clear that all meetings related to Cambodian affairs must be conducted in the Kingdom of Cambodia , " Hun Sen told reporters after a Cabinet meeting on Friday .`` No-one should internationalize Cambodian affairs .
It is detrimental to the sovereignty of Cambodia , " he said .Hun Sen 's Cambodian People 's Party won 64 of the 122 parliamentary seats in July 's elections , short of the two-thirds majority needed to form a government on its own .Ranariddh and Sam Rainsy have charged that Hun Sen 's victory in the elections was achieved through widespread fraud .They have demanded a thorough investigation into their election complaints as a precondition for their cooperation in getting the national assembly moving and a new government formed .....

Summary

Cambodian government rejects opposition's call for talks abroad

Multi-doc summarization

Documents

Fingerprints and photos of two men who boarded the doomed Malaysia Airlines passenger jet are being sent to U.S. authorities so they can be compared against records of known terrorists and criminals. The cause of the plane's disappearance has baffled investigators and they have not said that they believed that terrorism was involved, but they are also not ruling anything out. The investigation into the disappearance of the jetliner with 239 passengers and crew has centered so far around the fact that two passengers used passports stolen in Thailand from an Austrian and an Italian. The plane which left Kuala Lumpur, Malaysia, was headed for Beijing. Three of the passengers, one adult and two children, were American. ……

(CNN) -- A delegation of painters and calligraphers, a group of Buddhists returning from a religious gathering in Kuala Lumpur, a three-generation family, nine senior travelers and five toddlers. Most of the 227 passengers on board missing Malaysia Airlines Flight 370 were Chinese, according to the airline's flight manifest. The 12 missing crew members on the flight that disappeared early Saturday were Malaysian. The airline's list showed the passengers hailed from 14 countries, but later it was learned that two people named on the manifest -- an Austrian and an Italian -- whose passports had been stolen were not aboard the plane. The plane was carrying five children under 5 years old, the airline said. ……

⋮

Vietnamese aircraft spotted what they suspected was one of the doors belonging to the ill-fated Malaysia Airlines Flight MH370 on Sunday, as troubling questions emerged about how two passengers managed to board the Boeing 777 using stolen passports. The discovery comes as officials consider the possibility that the plane disintegrated mid-flight, a senior source told Reuters. The state-run Thanh Nien newspaper cited Lt. Gen. Vo Van Tuan, deputy chief of staff of Vietnam's army, as saying searchers in a low-flying plane had spotted an object suspected of being a door from the missing jet. It was found in waters about 56 miles south of Tho Chu island, in the same area where oil slicks were spotted Saturday. ……

Summary

Flight MH370, carrying 239 people vanished over the South China Sea in less than an hour after taking off from Kuala Lumpur, with two passengers boarded the Boeing 777 using stolen passports. Possible reasons could be an abrupt breakup of the plane or an act of terrorism. The government was determining the "true identities" of the passengers who used the stolen passports. Investigators were trying to determine the path of the plane by analysing civilian and military radar data while ships and aircraft from seven countries scouring the seas around Malaysia and south of Vietnam.

Output

Extractive summarization

Select from the source text spans that capture the key information

Abstractive summarization

Generate new text that encapsulates the key information from the source text

Focus

Generic summarization

- Summarize the content of the doc(s)

Query-focused summarization

- Summarize a doc w.r.t. a user's query
- Complex question-answering (answer a question by summarizing a doc that has the information to construct the answer)

Query-focused summarization: snippets

Google

how long did world war ii last

All | Images | News | Videos | Shopping | More | Tools

About 5,430,000,000 results (0.89 seconds)

https://www.history.com › news › world-war-ii-end-eve...

How Did World War II End? - HISTORY

Aug 11, 2020 — The **war** lasted six years and a day. These key moments marked the beginning of Allied victory over the Axis powers.

https://en.wikipedia.org › wiki › World_War_II

World War II - Wikipedia

World War II or the Second World War, often abbreviated as **WWII** or **WW2**, was a global war that lasted from 1939 to 1945. It involved the vast majority of the ...

Timeline of World War II (1945) · List of World War II battles · Category · War effort

Query-focused summarization: snippets

- Create answers to complex questions, summarizing multiple documents
- Instead of giving a snippet for each document, create a cohesive answer that combines information from multiple documents

Q: does this seem like a job for extractive or abstractive summarization?

Outline

- What is summarization?
- Data and metrics
- Workshop time
- Modern approaches

Outline

- What is summarization?
- Data and metrics
- Workshop time
- Modern approaches

Data and metrics

Q: How to find the list of commonly used datasets?

A: Look at the recent SOTA papers

Data and metrics

Q: How to find the list of commonly used datasets?

A: Look at the recent SOTA papers

Q: How to find the SOTA papers?

A:

- nlpprogress.com
- paperswithcode.com/sota
- connectedpapers.com

Data and metrics

Text Summarization with Pretrained Encoders

Yang Liu and **Mirella Lapata**
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
yang.liu2@ed.ac.uk, mlap@inf.ed.ac.uk

Datasets	# docs (train/val/test)	avg. doc length		avg. summary length		% novel bi-grams in gold summary
		words	sentences	words	sentences	
CNN	90,266/1,220/1,093	760.50	33.98	45.70	3.59	52.90
DailyMail	196,961/12,148/10,397	653.33	29.33	54.65	3.86	52.16
NYT	96,834/4,000/3,452	800.04	35.55	45.54	2.44	54.70
XSum	204,045/11,332/11,334	431.07	19.77	23.26	1.00	83.31

Table 1: Comparison of summarization datasets: size of training, validation, and test sets and average document and summary length (in terms of words and sentences). The proportion of novel bi-grams that do not appear in source documents but do appear in the gold summaries quantifies corpus bias towards extractive methods.

Data and metrics

SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization

Yixin Liu
Carnegie Mellon University
yixinl2@cs.cmu.edu

Pengfei Liu *
Carnegie Mellon University
pliu3@cs.cmu.edu

3 Experiments

3.1 Datasets

We use two datasets for our experiments. The dataset statistics are listed in Appendix A.

CNNDM CNN/DailyMail[2] (Hermann et al., 2015; Nallapati et al., 2016) dataset is a large scale news articles dataset.

XSum XSum[3] (Narayan et al., 2018) dataset is a highly abstractive dataset containing online articles from the British Broadcasting Corporation (BBC).

Data and metrics

4 Dataset

We use the *CNN/Daily Mail* dataset (Hermann et al., 2015; Nallapati et al., 2016), which contains online news articles (781 tokens on average) paired with multi-sentence summaries (3.75 sentences or 56 tokens on average). We used scripts supplied by Nallapati et al. (2016) to obtain the same version of the the data, which has 287,226 training pairs, 13,368 validation pairs and 11,490 test pairs. Both the dataset's published results (Nallapati et al., 2016, 2017) use the *anonymized* version of the data, which has been pre-processed to replace each named entity, e.g., *The United Nations*, with its own unique identifier for the example pair, e.g., `@entity5`. By contrast, we operate directly on the original text (or *non-anonymized* version of the data),[2] which we believe is the favorable problem to solve because it requires no pre-processing.

Get To The Point: Summarization with Pointer-Generator Networks

Abigail See
Stanford University
abisee@stanford.edu

Peter J. Liu
Google Brain
peterjliu@google.com

Christopher D. Manning
Stanford University
manning@stanford.edu

CNN/DailyMail

STORY HIGHLIGHTS

Trump will head to Texas on Tuesday

The White House has yet to say where Trump will travel

Washington (CNN) — President Donald Trump struck a unifying tone Monday as he addressed the devastation in Texas wrought by Hurricane Harvey at the top of a joint news conference with Finland's president.

"We see neighbor helping neighbor, friend helping friend and stranger helping stranger," Trump said. "We are one American family. We hurt together, we struggle together and believe me, we endure together."

Trump extended his "thoughts and prayers" to those affected by the hurricane and catastrophic flooding that ensued in Texas, and also promised Louisiana residents that the federal government is prepared to help as the tropical storm makes its way toward that state.

"To the people of Texas and Louisiana, we are 100% with you," Trump said from the East Room of the White House.

CNN/DailyMail

https://huggingface.co/datasets/cnn_dailymail

Dataset Card for CNN Dailymail Dataset

Dataset Summary

The CNN / DailyMail Dataset is an English-language dataset containing just over 300k unique news articles as written by journalists at CNN and the Daily Mail. The current version supports both extractive and abstractive summarization, though the original version was created for machine reading and comprehension and abstractive question answering.

Dataset Split	Number of Instances in Split
Train	287,113
Validation	13,368
Test	11,490

CNN/DailyMail

https://huggingface.co/datasets/cnn_dailymail

Initial Data Collection and Normalization

The data consists of news articles and highlight sentences. In the question answering setting of the data, the articles are used as the context and entities are hidden one at a time in the highlight sentences, producing Cloze style questions where the goal of the model is to correctly guess which entity in the context has been hidden in the highlight. In the summarization setting, the highlight sentences are concatenated to form a summary of the article. The CNN articles were written between April 2007 and April 2015. The Daily Mail articles were written between June 2010 and April 2015.

XSum

SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization

XSum XSum[3] (Narayan et al., 2018) dataset is a highly abstractive dataset containing online articles from the British Broadcasting Corporation (BBC).

Let's follow the trail…

XSum

Don't Give Me the Details, Just the Summary!
Topic-Aware Convolutional Neural Networks for Extreme Summarization

Shashi Narayan **Shay B. Cohen** **Mirella Lapata**
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh, EH8 9AB
shashi.narayan@ed.ac.uk, {scohen,mlap}@inf.ed.ac.uk

SUMMARY: *A man and a child have been killed after a light aircraft made an emergency landing on a beach in Portugal.*

DOCUMENT: Authorities said the incident took place on Sao Joao beach in Caparica, south-west of Lisbon.
The National Maritime Authority said a middle-aged man and a young girl died after they were unable to avoid the plane.
[*6 sentences with 139 words are abbreviated from here.*]
Other reports said the victims had been sunbathing when the plane made its emergency landing.
[*Another 4 sentences with 67 words are abbreviated from here.*]
Video footage from the scene carried by local broadcasters showed a small recreational plane parked on the sand, apparently intact and surrounded by beachgoers and emergency workers.
[*Last 2 sentences with 19 words are abbreviated.*]

Figure 1: An abridged example from our extreme summarization dataset showing the document and its one-line summary. Document content present in the summary is color-coded.

Datasets	# docs (train/val/test)	avg. document length		avg. summary length		vocabulary size	
		words	sentences	words	sentences	document	summary
CNN	90,266/1,220/1,093	760.50	33.98	45.70	3.59	343,516	89,051
DailyMail	196,961/12,148/10,397	653.33	29.33	54.65	3.86	563,663	179,966
NY Times	589,284/32,736/32,739	800.04	35.55	45.54	2.44	1,399,358	294,011
XSum	204,045/11,332/11,334	431.07	19.77	23.26	1.00	399,147	81,092

Metrics: ROUGE-N

Intrinsic metric for automatically evaluating summaries

- Based on BLEU (remember, MT)
- Not as good as human evaluation (of course)

Given a document D, and an automated summary X:

- Get N humans to produce a set of reference summaries of D
- Run model to produce X

Metrics: ROUGE-N

https://dl.acm.org/doi/pdf/10.5555/1273073.1273093

4 ROUGE

The state-of-the-art automatic summarization evaluation method is ROUGE (Recall Oriented Understudy for Gisting Evaluation, (Hovy and Lin 2002)), an n-gram based comparison that was motivated by the machine translation evaluation metric, Bleu (Papineni et. al. 2001). This system uses a variety of n-gram matching approaches, some of which allow gaps within the matches as well as more sophistcated analyses. Surprisingly, simple unigram and bigram matching works extremely well. For example, at DUC 05, ROUGE-2 (bigram match) had a Spearman correlation of 0.95 and a Pearson correlation of 0.97 when compared with human evaluation of the summaries for responsiveness (Dang 2005). ROUGE-n for matching $n-$grams of a summary X against h model human summaries is given by:

$$R_n(X) = \frac{\sum_{j=1}^{h} \sum_{i \in N_n} \min(X_n(i), M_n(i,j))}{\sum_{j=1}^{h} \sum_{i \in N_n} M_n(i,j),}$$

where $X_n(i)$ is the count of the number of times the n-gram i occurred in the summary and $M_n(i,j)$ is the number of times the n-gram i occurred in the j-th model (human) summary. (Note that for brevity of notation, we assume that lemmatization (stemming) is done *apriori* on the terms.)

Metrics: ROUGE-N

- Query: "What is water spinach?"
- Model's output: "Water spinach is a leaf vegetable commonly eaten in tropical areas of Asia."
- Human Summaries (gold truth):

Human 1: Water spinach is a green leafy vegetable grown in the tropics.

Human 2: Water spinach is a semi-aquatic tropical plant grown as a vegetable.

Human 3: Water spinach is a commonly eaten leaf vegetable of Asia.

- ROUGE-2 $= \frac{3 + 3 + 6}{10 + 9 + 9} = 12/28 = .43$

Outline

- What is summarization?
- Data and metrics
- Workshop time
- Modern approaches

Outline

- What is summarization?
- Data and metrics
- Workshop time
- Modern approaches

Workshop time

How would you attempt summarization?

Create a baseline system

Unsupervised extractive approaches

1. content selection: choose sentences to extract from the document
2. information ordering: choose an order to place them in the summary
3. sentence realization: clean up the sentences

Unsupervised extractive approaches

H. P. Luhn. 1958. The Automatic Creation of Literature Abstracts.
IBM Journal of Research and Development. 2:2, 159-165.

- Intuition dating back to Luhn (1958):
  - Choose sentences that have salient or informative words
- Two approaches to defining salient words

1. tf-idf: weigh each word w_i in document *j* by tf-idf

$$weight(w_i) = tf_{ij} \times idf_i$$

2. topic signature: choose a smaller set of salient words
   - mutual information
   - log-likelihood ratio (LLR) Dunning (1993), Lin and Hovy (2000)

$$weight(w_i) = \begin{cases} 1 & \text{if } -2\log\lambda(w_i) > 10 \\ 0 & \text{otherwise} \end{cases}$$

Signature-based content selection with queries

Conroy, Schlesinger, and O'Leary 2006

- choose words that are informative either
  - by log-likelihood ratio (LLR)
  - or by appearing in the query

$$weight(w_i) = \begin{cases} 1 & \text{if } -2\log\lambda(w_i) > 10 \\ 1 & \text{if } w_i \in question \\ 0 & \text{otherwise} \end{cases}$$

(could learn more complex weights)

- Weigh a sentence (or window) by weight of its words:

$$weight(s) = \frac{1}{|S|}\sum_{w \in S} weight(w)$$

Graph-based ranking algorithms

- unsupervised sentence extraction

$$Similarity(S_i, S_j) = \frac{|W_k | W_k \in S_i \& W_k \in S_j|}{log(|S_i|) + log(|S_j|)}$$

3: BC–HurricaineGilbert, 09–11 339
4: BC–Hurricaine Gilbert, 0348
5: Hurricaine Gilbert heads toward Dominican Coast
6: By Ruddy Gonzalez
7: Associated Press Writer
8: Santo Domingo, Dominican Republic (AP)
9: Hurricaine Gilbert Swept towrd the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains, and high seas.
10: The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph.
11: "There is no need for alarm," Civil Defense Director Eugenio Cabral said in a television alert shortly after midnight Saturday.
12: Cabral said residents of the province of Barahona should closely follow Gilbert's movement.
13: An estimated 100,000 people live in the province, including 70,000 in the city of Barahona, about 125 miles west of Santo Domingo.
14. Tropical storm Gilbert formed in the eastern Carribean and strenghtened into a hurricaine Saturday night.
15: The National Hurricaine Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo.
16: The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westard at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the storm.
17. The weather service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday.
18: Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds, and up to 12 feet to Puerto Rico's south coast.
19: There were no reports on casualties.
20: San Juan, on the north coast, had heavy rains and gusts Saturday, but they subsided during the night.
21: On Saturday, Hurricane Florence was downgraded to a tropical storm, and its remnants pushed inland from the U.S. Gulf Coast.
22: Residents returned home, happy to find little damage from 90 mph winds and sheets of rain.
23: Florence, the sixth named storm of the 1988 Atlantic storm season, was the second hurricane.
24: The first, Debby, reached minimal hurricane strength briefly before hitting the Mexican coast last month.

Figure 1: Sample graph build for sentence extraction from a newspaper article.

Rada Mihalcea, ACL 2004

Outline

- What is summarization?
- Data and metrics
- Workshop time
- Modern approaches

Outline

- What is summarization?
- Data and metrics
- Workshop time
- Modern approaches

Modern approaches

Preparation: Research Concepts

Generation Way

- `gen-ext` : Extractive Summarization
- `gen-abs` : Abstractive Summarization
- `gen-2stage` Two-stage Summarization (compressive, hybrid)

Regressive Way

- `regr-auto` : Autoregressive Decoder (Pointer network)
- `regr-nonauto` : Non-autoregressive Decoder (Sequence labeling)

Supervision

- `sup-sup` : Supervised Learning
- `sup-weak` (implies `sup-sup`): Weakly Supervised Learning
- `sup-unsup` : Unsupervised Learning

Task Settings

rich of task settings!

- `task-single` : Single-document Summarization
- `task-multi` : Multi-document Summarization
- `task-senCompre:` Sentence Compression
- `task-sci` : Scientific Paper
- `task-multimodal` : Multi-modal Summarization
- `task-aspect` : Aspect-based Summarization
- `task-opinion` : Opinion Summarization
- `task-questoin` : Question-based Summarization

Architecture (Mechanism)

- `arch-rnn` : Recurrent Neural Networks (LSTM, GRU)
- `arch-cnn` : Convolutional Neural Networks (CNN)
- `arch-transformer` : Transformer
- `arch-graph` : Graph Neural Networks or Statistic Graph Models
- `arch-gnn` : Graph Neural Networks
- `arch-att` : Attention Mechanism
- `arch-pointer` : Pointer Layer
- `arch-coverage` : Coverage Mechanism

Training

- `train-multitask` : Multi-task Learning
- `train-multilingual` : Multi-lingual Learning
- `train-multimodal` : Multi-modal Learning
- `train-auxiliary` : Joint Training
- `train-transfer` : Cross-domain Learning, Transfer Learning, Domain Adaptation
- `train-active` : Active Learning, Boostrapping
- `train-adver` : Adversarial Learning
- `train-template` : Template-based Summarization
- `train-augment` : Data Augmentation
- `train-curriculum` : Curriculum Learning
- `train-lowresource` : Low-resource Summarization
- `train-retrieval` : Retrieval-based Summarization
- `train-meta` : Meta-learning

Pre-trained Models

- `pre-word2vec` : word2vec
- `pre-glove` : GLoVe
- `pre-bert` : BERT

Modern approaches

Modern approaches

Modern approaches

Modern approaches

Modern approaches

Modern approaches

2018: booming!

Modern approaches

2019

Abstractive Text Summarization using seq-to-seq RNNs

Implements many tricks (nmt, copy, coverage, hierarchical, external knowledge)

Figure 1: Feature-rich-encoder: We use one embedding vector each for POS, NER tags and discretized TF and IDF values, which are concatenated together with word-based embeddings as input to the encoder.

Figure 2: Switching generator/pointer model: When the switch shows 'G', the traditional generator consisting of the softmax layer is used to produce a word, and when it shows 'P', the pointer network is activated to copy the word from one of the source document positions. When the pointer is activated, the embedding from the source is used as input for the next time-step as shown by the arrow from the encoder to the decoder at the bottom.

Nallapati et al., CoNLL 2016

Abstractive Text Summarization using seq-to-seq RNNs

Figure 3: Hierarchical encoder with hierarchical attention: the attention weights at the word level, represented by the dashed arrows are re-scaled by the corresponding sentence-level attention weights, represented by the dotted arrows. The dashed boxes at the bottom of the top layer RNN represent sentence-level positional embeddings concatenated to the corresponding hidden states.

Nallapati et al., CoNLL 2016

Abstractive Text Summarization using seq-to-seq RNNs

Implements many tricks (nmt, copy, coverage, hierarchical, external knowledge)

#	Model name	Rouge-1	Rouge-2	Rouge-L	Src. copy rate (%)
Full length F1 on our internal test set					
1	words-lvt2k-1sent	34.97	17.17	32.70	75.85
2	words-lvt2k-2sent	35.73	17.38	33.25	79.54
3	words-lvt2k-2sent-hieratt	36.05	18.17	33.52	78.52
4	feats-lvt2k-2sent	35.90	17.57	33.38	78.92
5	feats-lvt2k-2sent-ptr	***36.40**	**17.77**	***33.71**	78.70
Full length F1 on the test set used by (Rush et al., 2015)					
6	ABS+ (Rush et al., 2015)	29.78	11.89	26.97	91.50
7	words-lvt2k-1sent	32.67	15.59	30.64	74.57
8	RAS-Elman (Chopra et al., 2016)	33.78	15.97	31.15	
9	words-lvt5k-1sent	***35.30**	**16.64**	***32.62**	

Nallapati et al., CoNLL 2016

Copy mechanism

- OOV, Extraction
- "Pointer networks" (Vinyals et al., 2015 NIPS)
- "Pointing the Unknown Words" (Gulcehre et al., ACL 2016)
- " Incorporating Copying Mechanism in Sequence-to-Sequence Learning " (Gu et al., ACL 2016)
- " Get To The Point: Summarization with Pointer-Generator Networks " (See et al., ACL 2017)

Copy mechanism

Copy words from the source text

See et al., ACL 2017

Copy mechanism

See et al., ACL 2017

$$P_{\text{vocab}} = \text{softmax}(V'(V[s_t, h_t^*] + b) + b')$$

$$p_{\text{gen}} = \sigma(w_{h^*}^T h_t^* + w_s^T s_t + w_x^T x_t + b_{\text{ptr}})$$

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

Copy mechanism

Article: andy murray (...) is into the semi-finals of the miami open , but not before getting a scare from 21 year-old austrian dominic thiem, who pushed him to 4-4 in the second set before going down 3-6 6-4, 6-1 in an hour and three quarters. (...)
Summary: andy murray **defeated** dominic thiem 3-6 6-4, 6-1 in an hour and three quarters.

Article: (...) wayne rooney smashes home during manchester united 's 3-1 win over aston villa on saturday. (...)
Summary: manchester united **beat** aston villa 3-1 at old trafford on saturday.

See et al., ACL 2017

Other lines of research

- Coverage Mechanism
  - "Modeling Coverage for Neural Machine Translation" (Tu et al., 2016 ACL)
- Graph-based attentional neural model
  - "Abstractive document summarization with a graph-based attentional neural model" (Tan et al., ACL 2017)
- Reinforcement Learning
  - "A deep reinforced model for abstractive summarization." (Paulus et al., ICLR 2018)
- Remaining Challenges
  - Long text abstractive summarization
  - Abstractive multi-document summarization

Historical Overview

- The development of deep NNs lags behind other tasks
- Summarization tasks requires some customized techniques (e.g. *copy*)
- Only technique-ready is not enough … dataset also matters!
- A good match between “techniques” and “datasets”

Let's explore

Text Summarization with Pretrained Encoders

Yang Liu and **Mirella Lapata**
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
yang.liu2@ed.ac.uk, mlap@inf.ed.ac.uk

In this paper, we showcase how BERT can be usefully applied in text summarization and propose a general framework for both extractive and abstractive models. We introduce a novel document-level encoder based on BERT which is able to express the semantics of a document and obtain representations for its sentences. Our extractive model is built on top of this encoder by stacking several inter-sentence Transformer layers. For abstractive summarization, we propose a new fine-tuning schedule which adopts different optimizers for the encoder and the decoder as a means of alleviating the mismatch between the two (the former is pretrained while the latter is not).

SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization

Yixin Liu
Carnegie Mellon University
yixinl2@cs.cmu.edu

Pengfei Liu *
Carnegie Mellon University
pliu3@cs.cmu.edu

In this paper, we present a conceptually simple while empirically powerful framework for abstractive summarization, SIMCLS, which can bridge the gap between the learning objective and evaluation metrics resulting from the currently dominated sequence-to-sequence learning framework by formulating text generation as a reference-free evaluation problem (i.e., quality estimation) assisted by contrastive learning