# Text summarization

DECEMBER 6, 2018

Elena Voita
Yandex Research,
University of Amsterdam
lena-voita@yandex-team.ru

# Plan

- Text summarization examples
- Text summarization methods
- Extractive vs Abstractive
- A piece of Extractive summarization
- Abstractive summarization
- Hack of the day

# Motivation

> “People spend 12 hours everyday consuming media in 2018.
>
> – eMarketer
>
> https://www.emarketer.com/topics/topic/time-spent-with-media

From http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

# Text Summarization

- To condense a piece of text to a shorter version while maintaining the important points


From http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

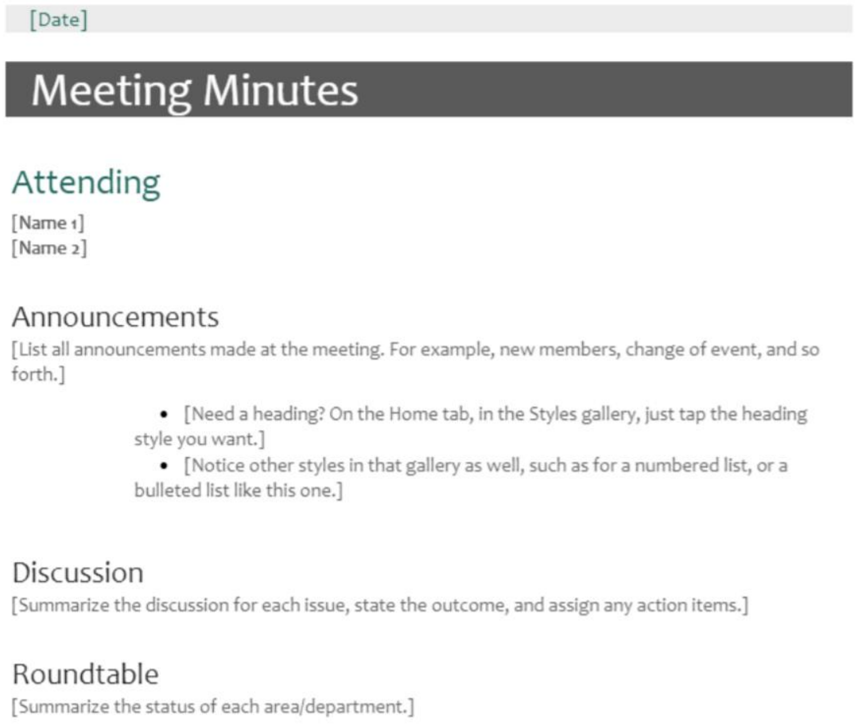
# Examples of Text Summarization

- Article headlines


From http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

# Examples of Text Summarization

- Article headlines
- Meeting minutes



[Date]

Meeting Minutes

Attending

[Name 1]
[Name 2]

Announcements

[List all announcements made at the meeting. For example, new members, change of event, and so forth.]

- [Need a heading? On the Home tab, in the Styles gallery, just tap the heading style you want.]
- [Notice other styles in that gallery as well, such as for a numbered list, or a bulleted list like this one.]

Discussion

[Summarize the discussion for each issue, state the outcome, and assign any action items.]

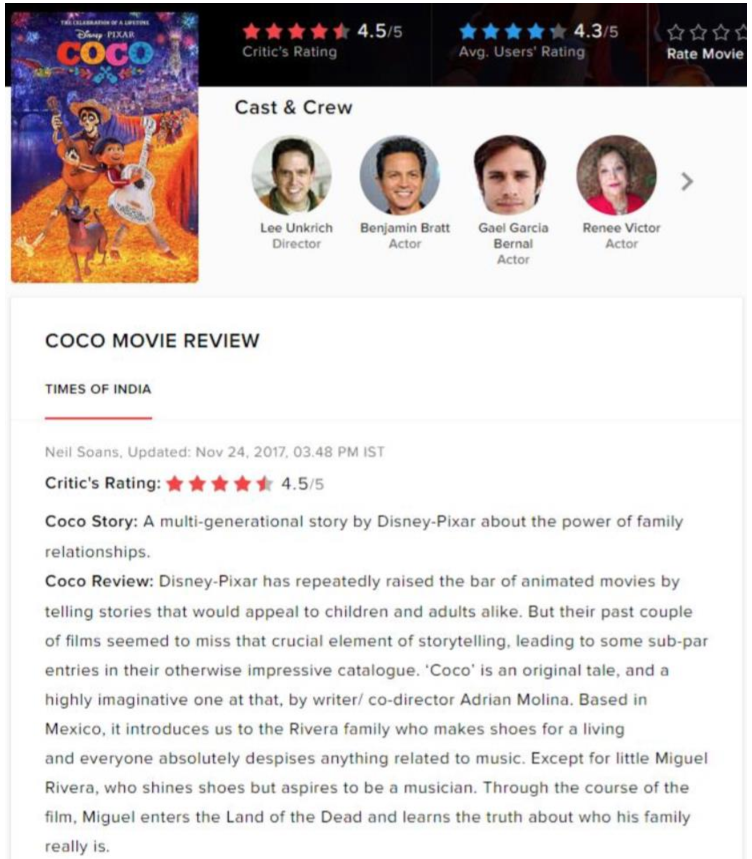
Roundtable

[Summarize the status of each area/department.]

From http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

# Examples of Text Summarization

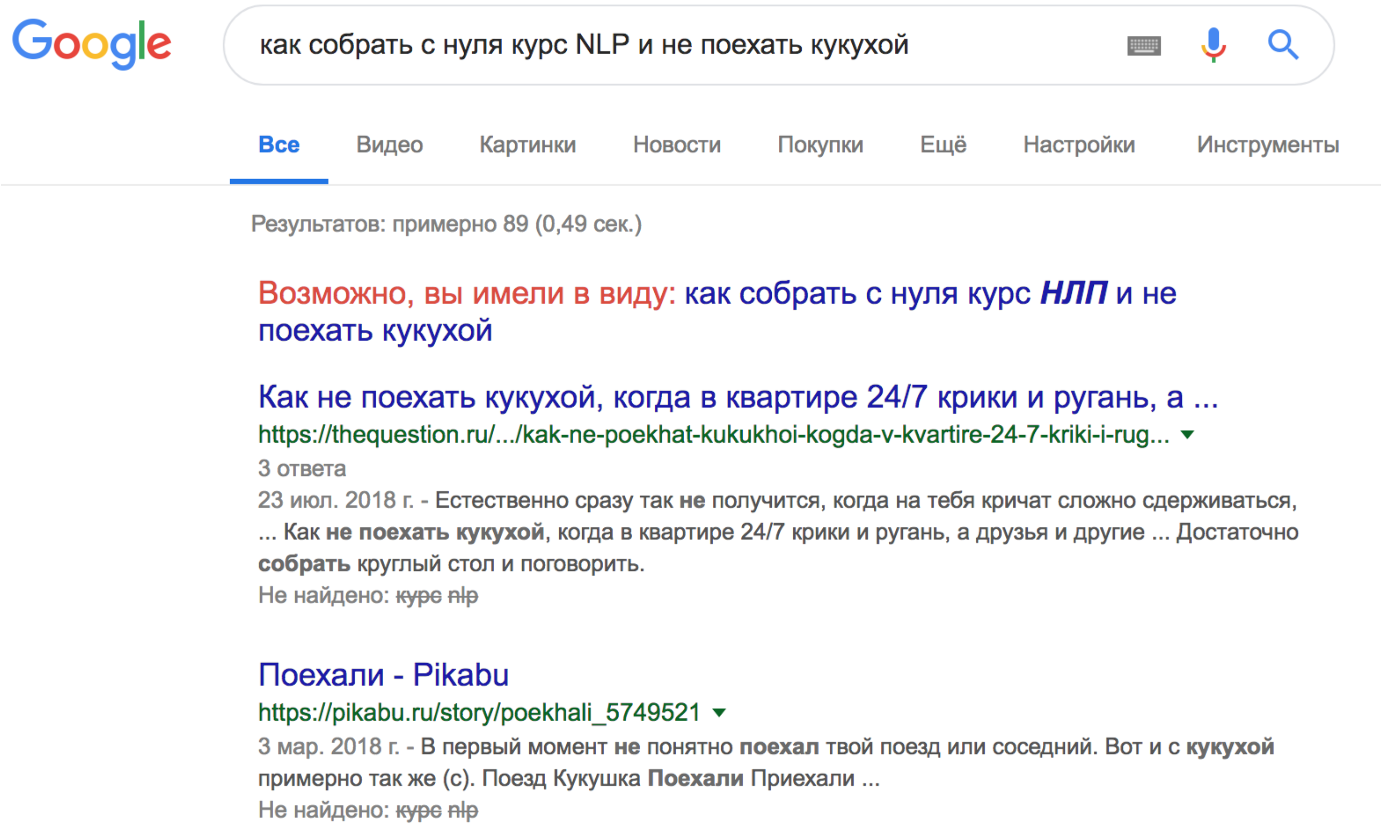
- ➢ Article headlines
- ➢ Meeting minutes
- ➢ Movie/book review




From http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

# Examples of Text Summarization

- Article headlines
- Meeting minutes
- Movie/book review
- Snippets in Web search

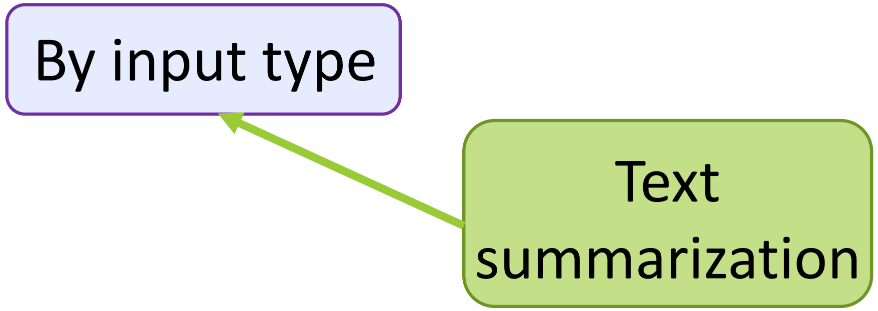


# Text Summarization Methods

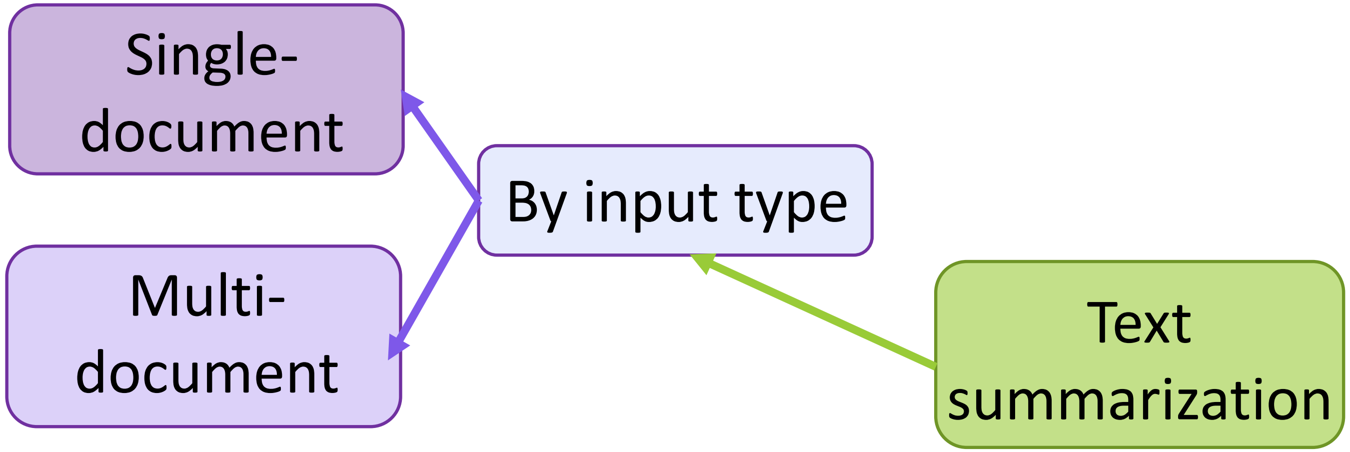
# Text summarization methods

Text summarization

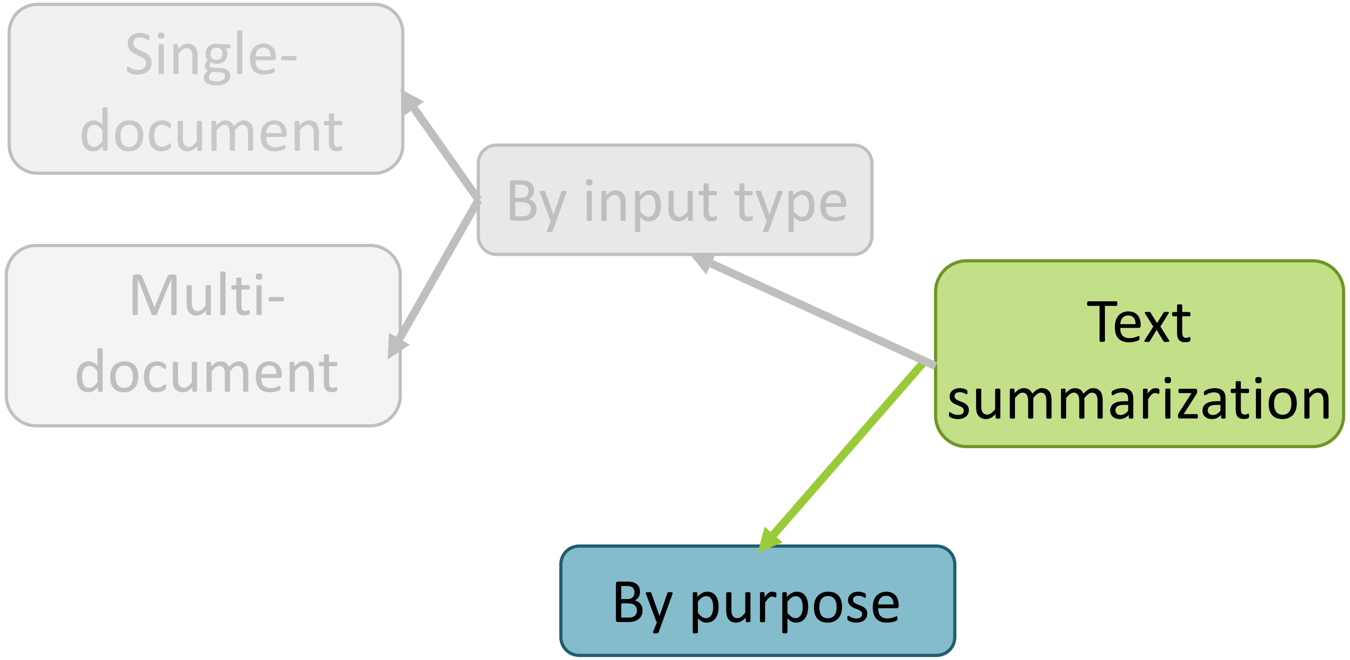
# Text summarization methods



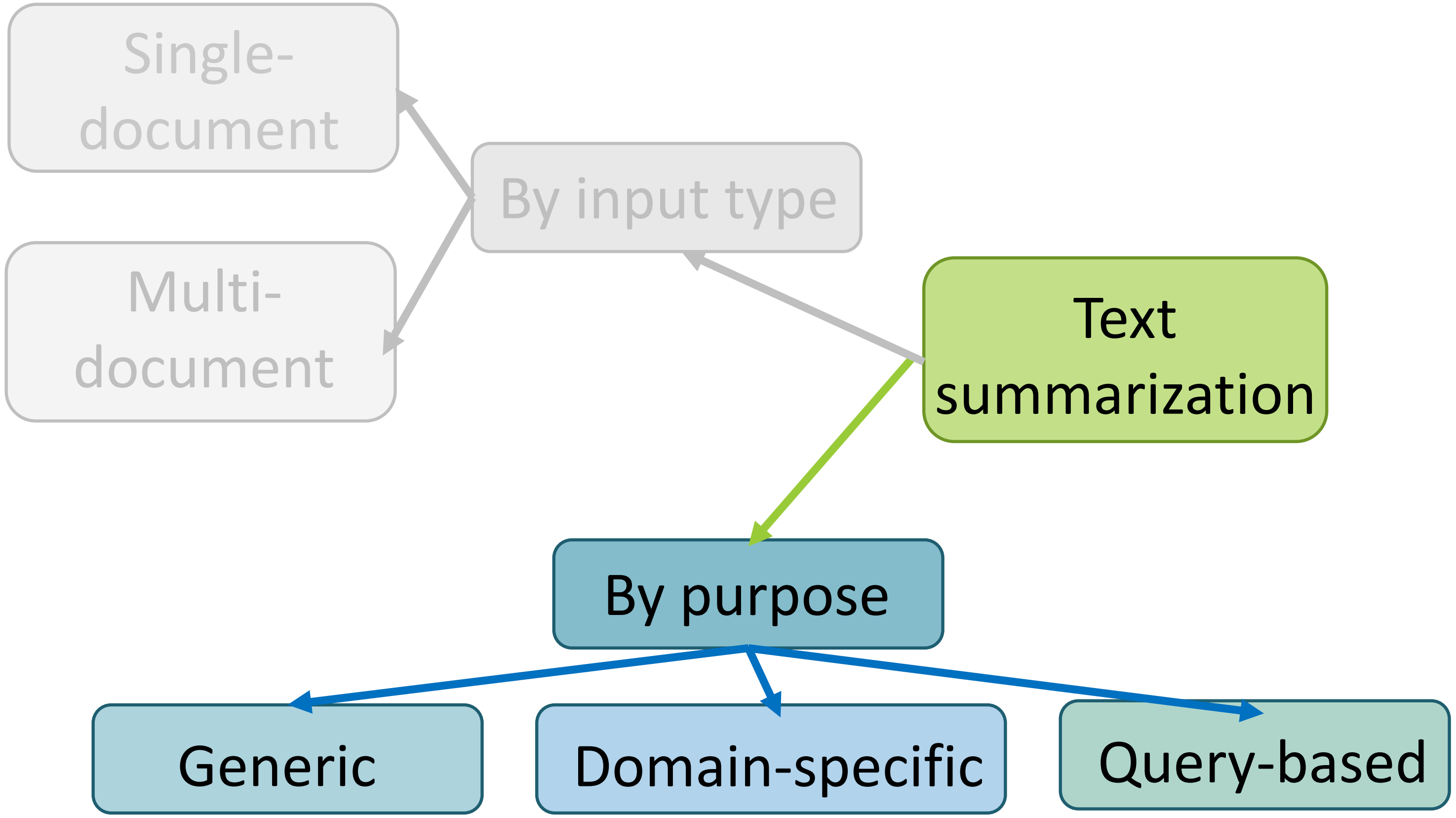
# Text summarization methods



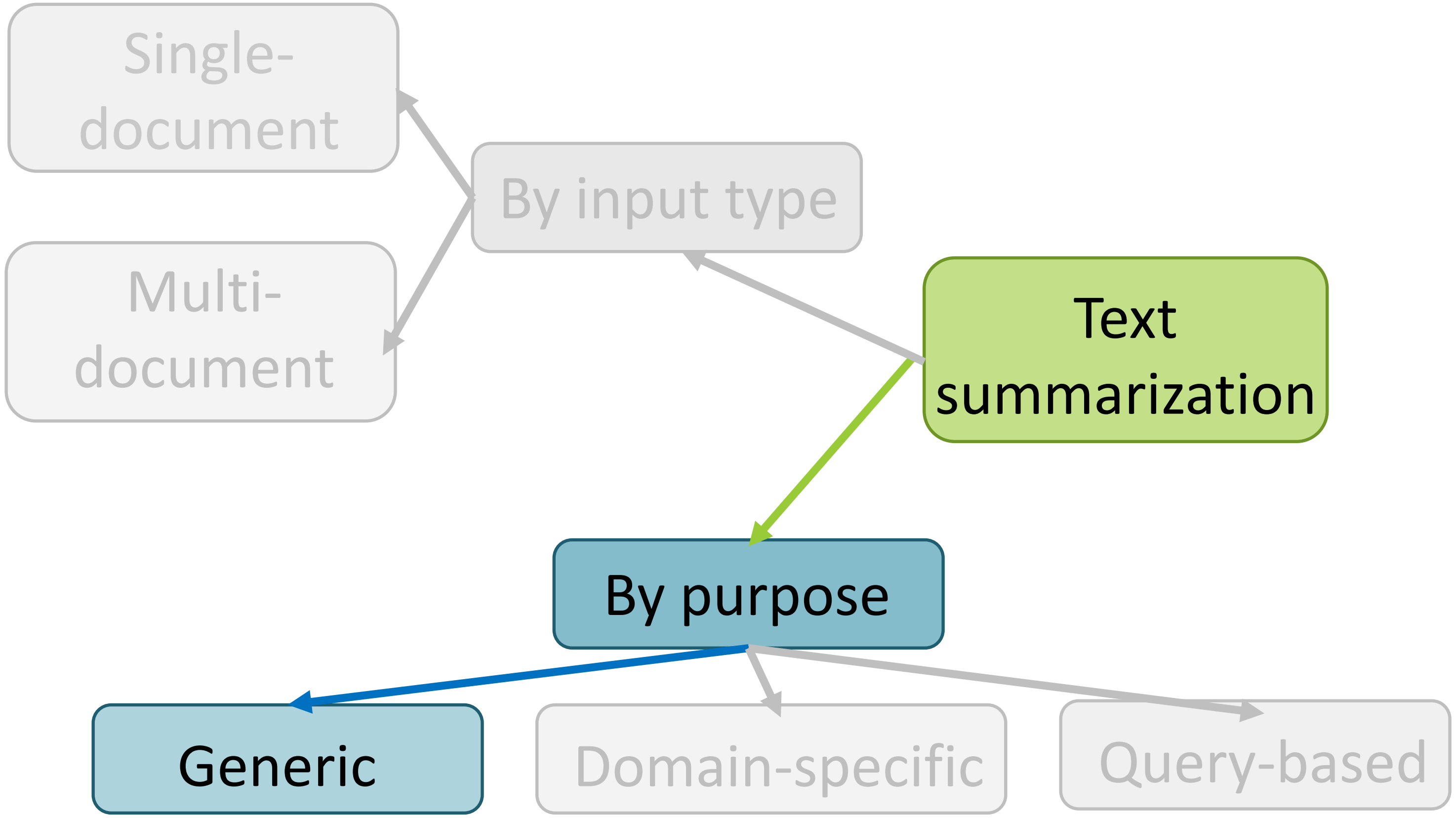
# Text summarization methods



# Text summarization methods

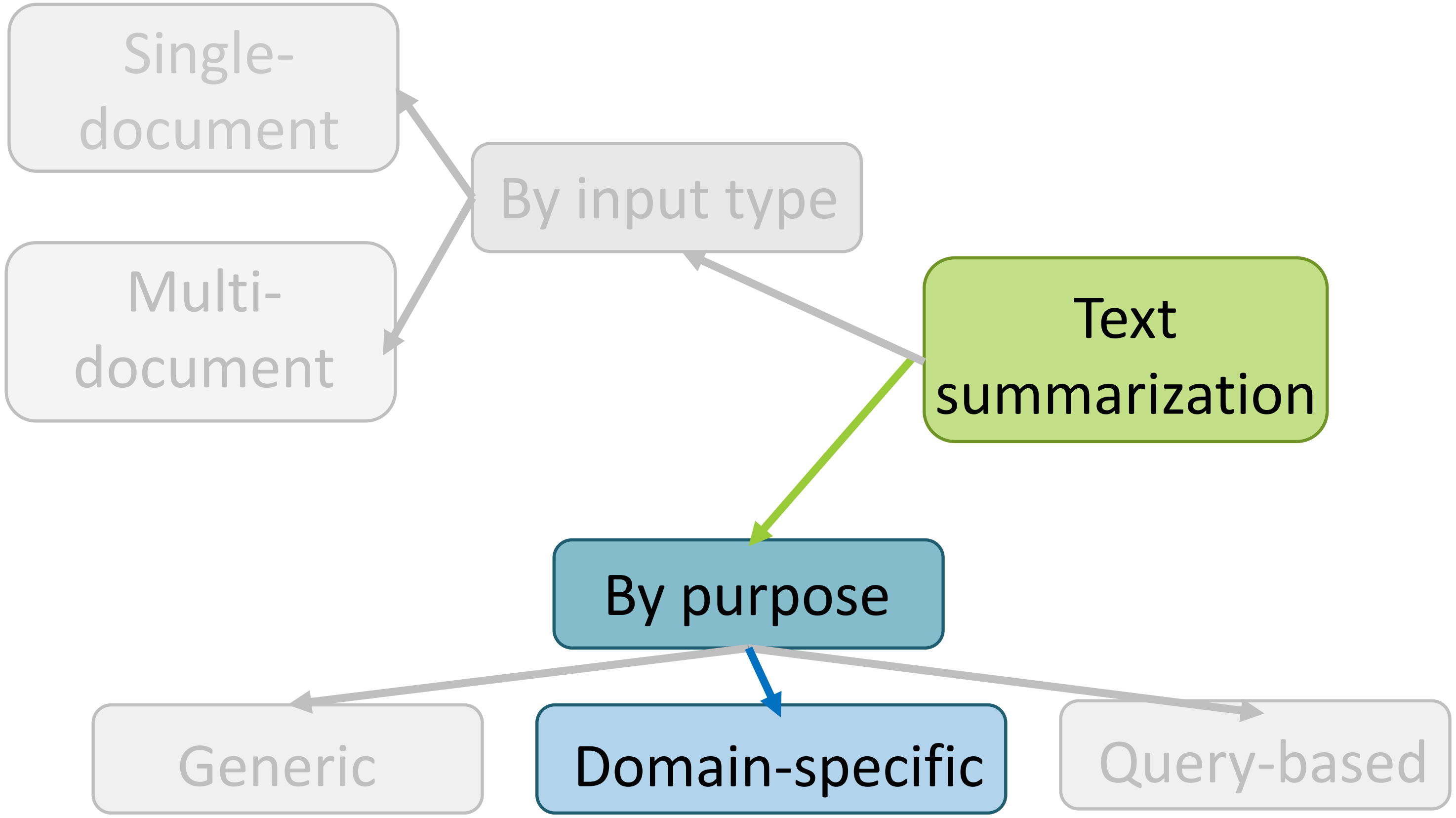


# Text summarization methods



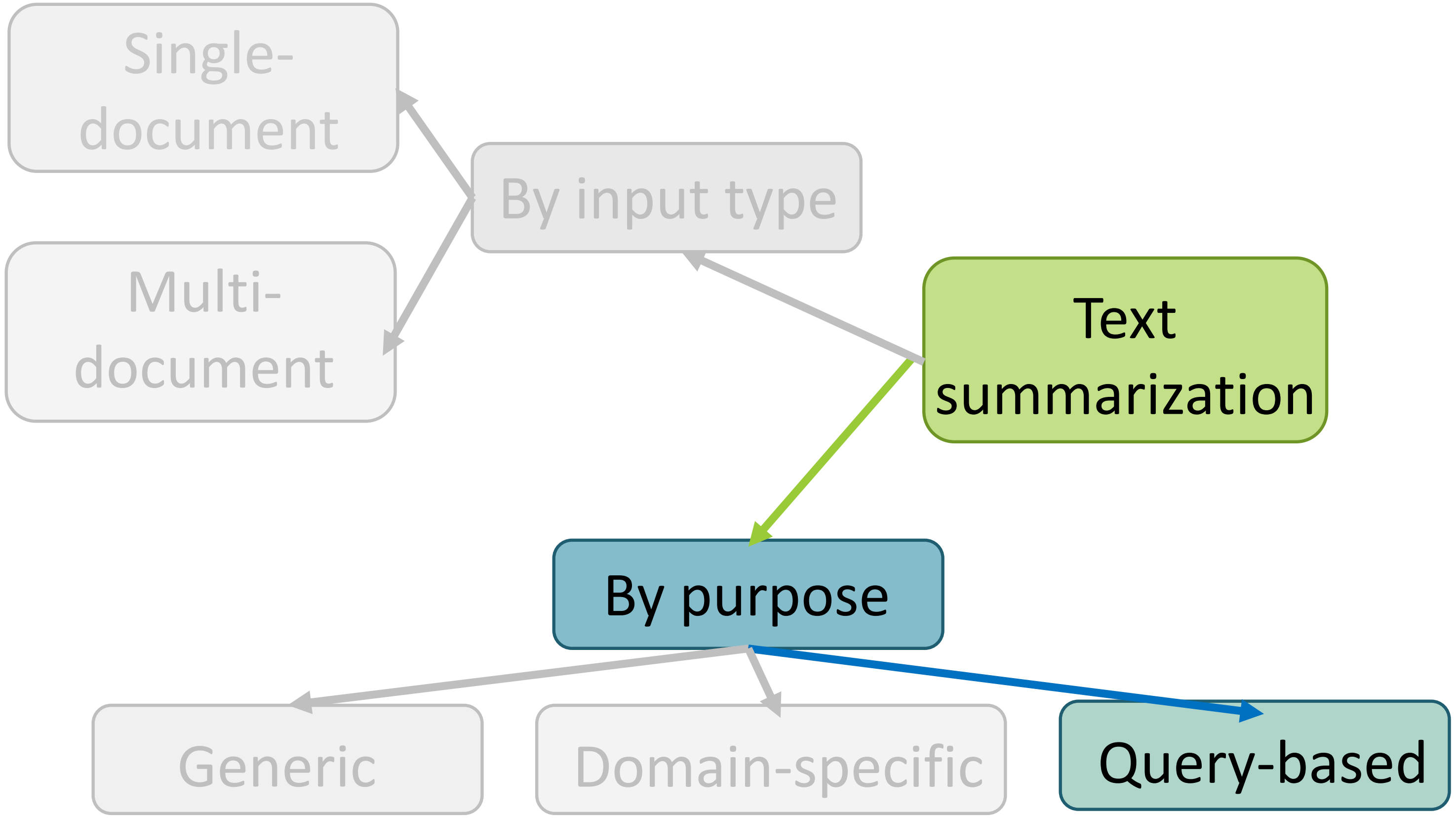

Given any document, make a summary (no assumptions on domain or content)

# Text summarization methods



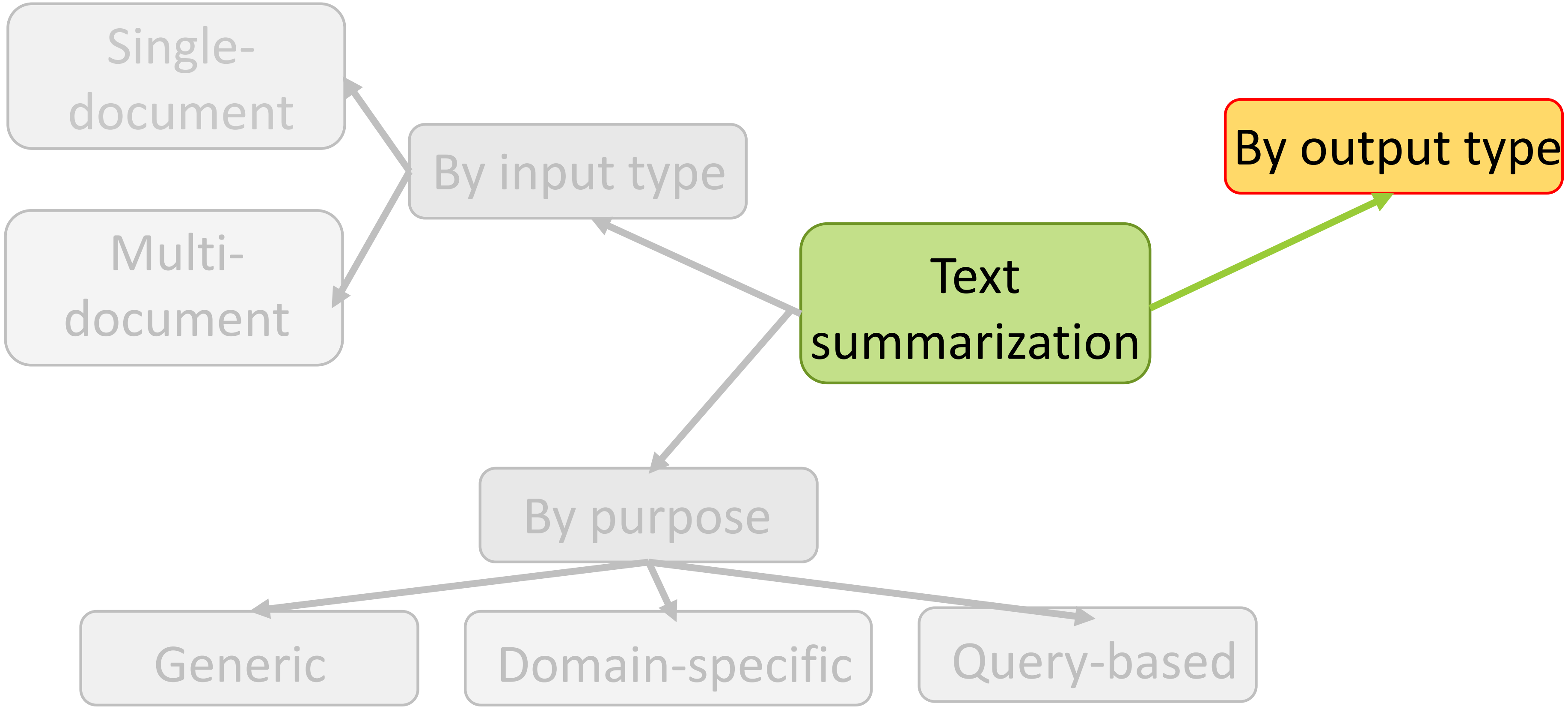

- Biomedical documents
- Research papers

# Text summarization methods

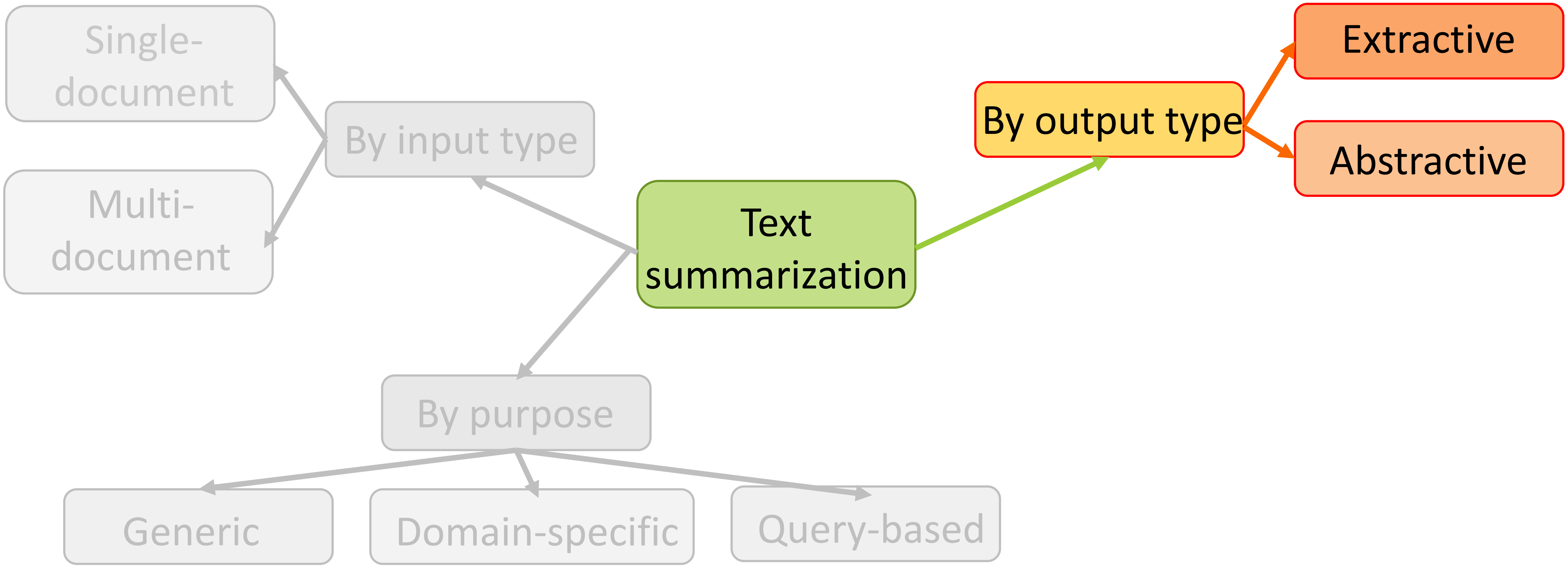



Summary contains only information answering a query

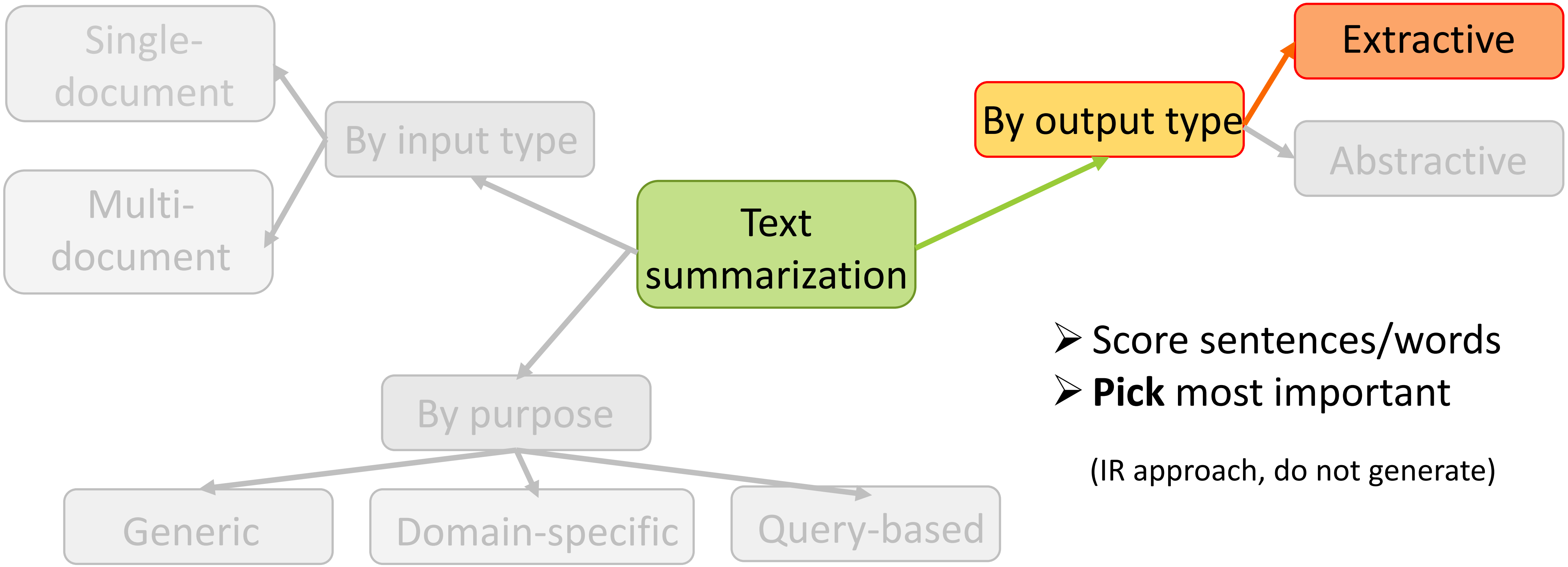
# Text summarization methods



# Text summarization methods



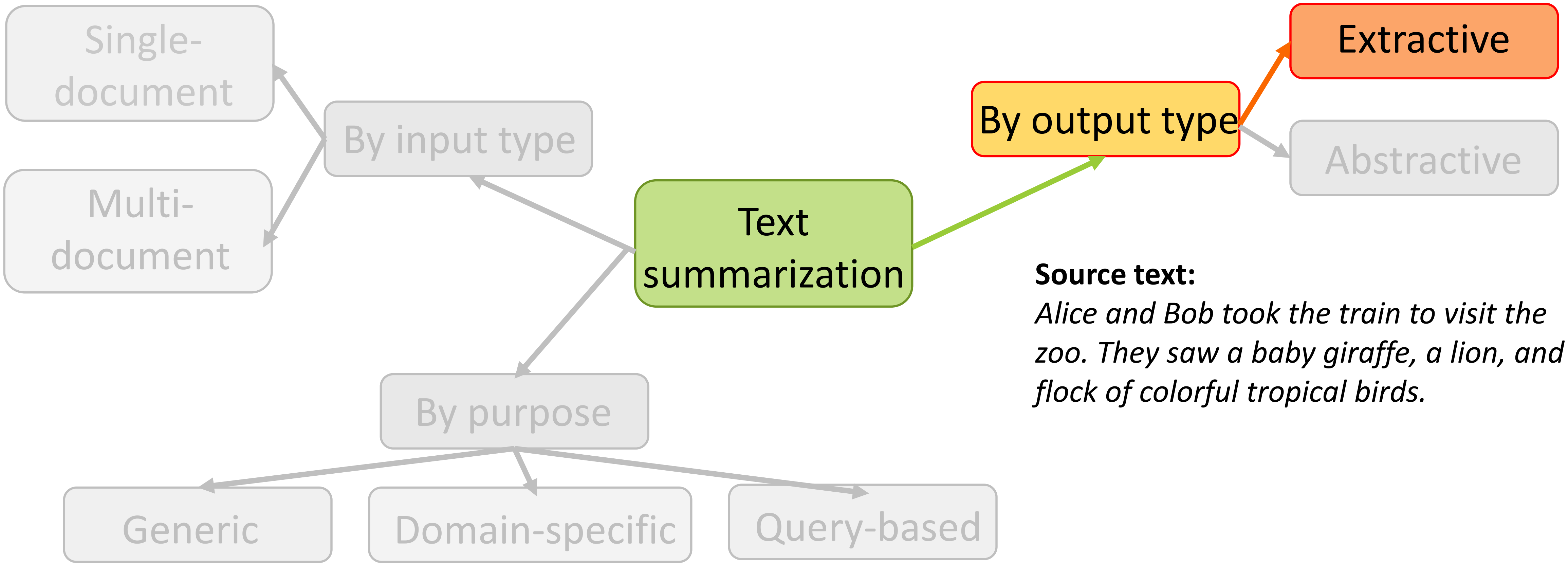
# Text summarization methods




- ➢ Score sentences/words
- ➢ **Pick** most important

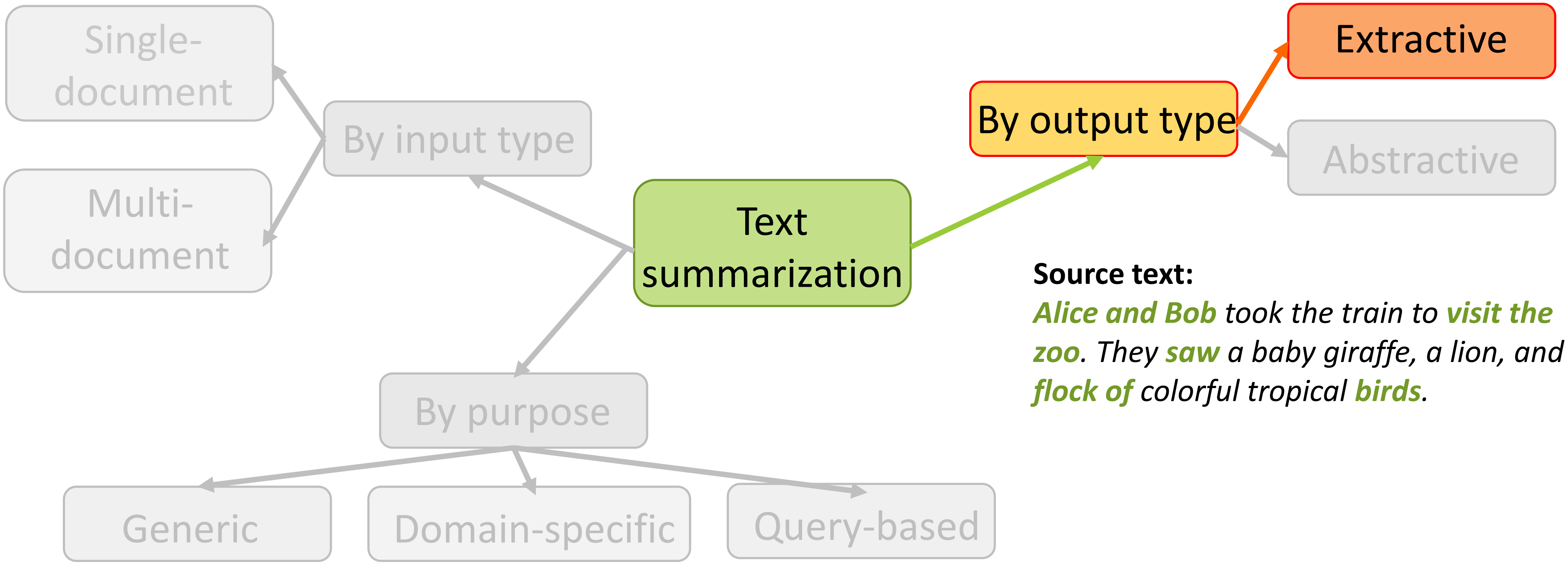
(IR approach, do not generate)

# Text summarization methods




**Source text:**
*Alice and Bob took the train to visit the zoo. They saw a baby giraffe, a lion, and a flock of colorful tropical birds.*

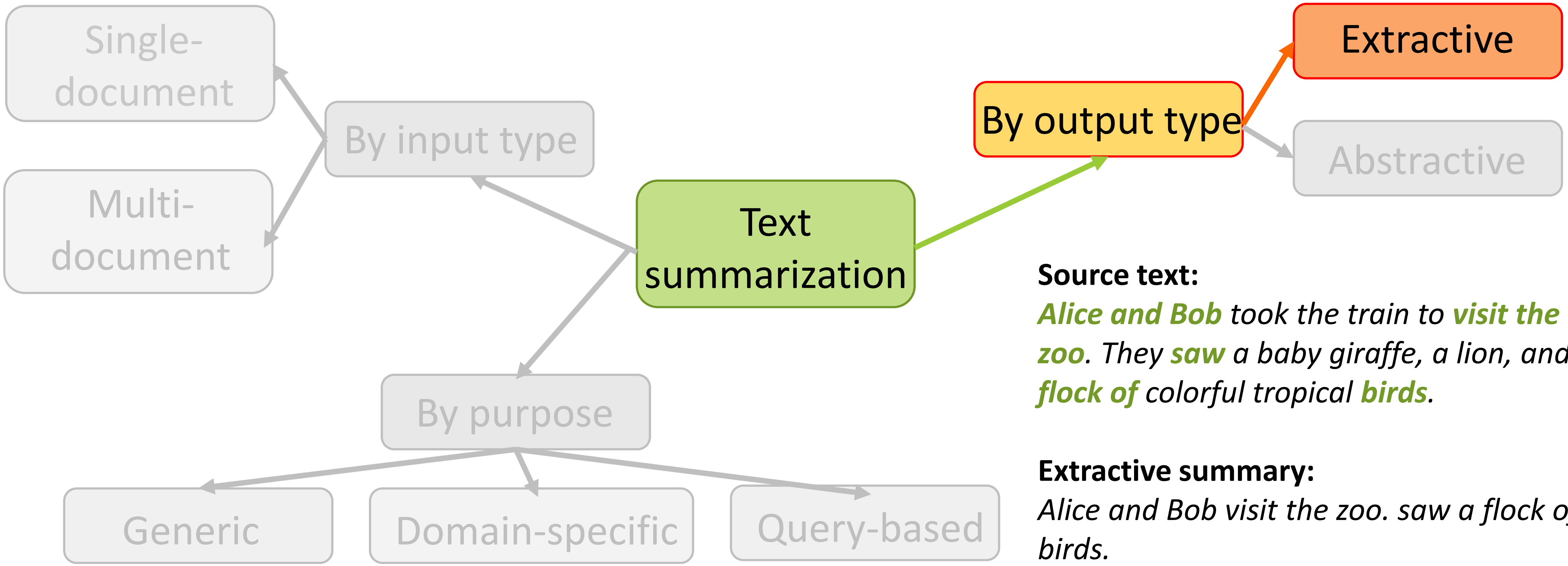
# Text summarization methods




**Source text:**

***Alice and Bob*** *took the train to* ***visit the zoo****. They* ***saw*** *a baby giraffe, a lion, and* ***a flock of*** *colorful tropical* ***birds****.*

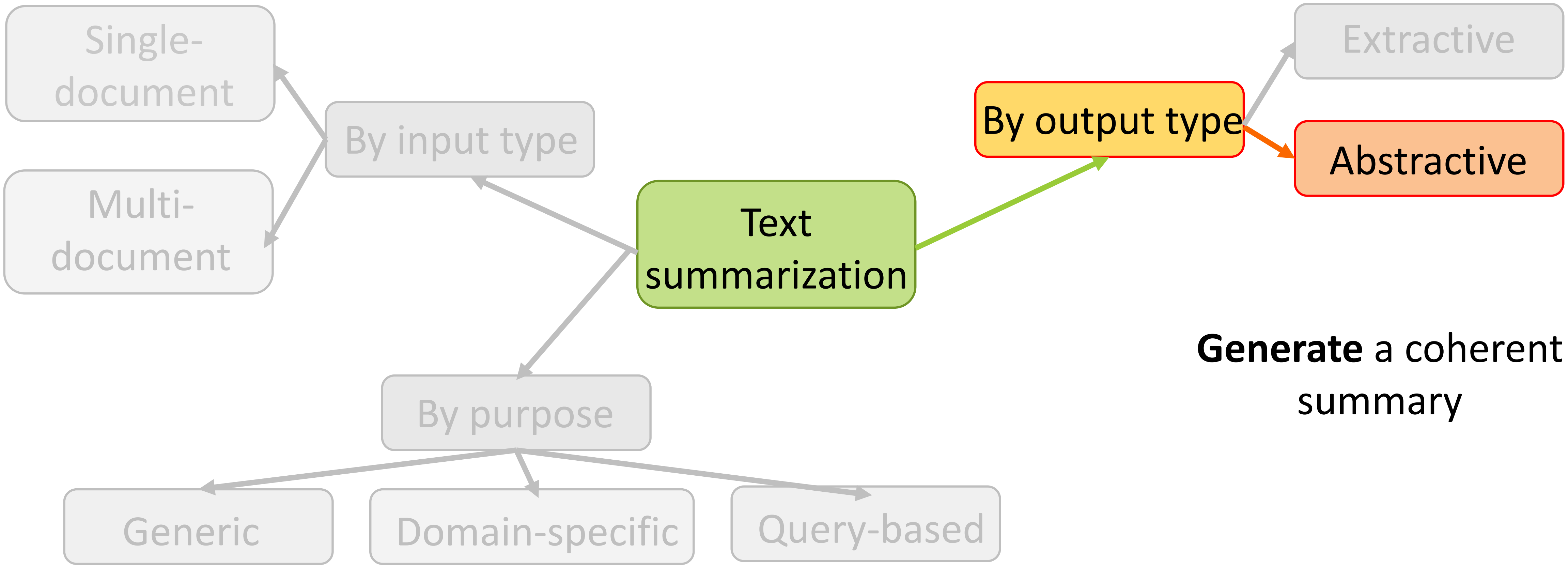
# Text summarization methods




**Source text:**
***Alice and Bob*** *took the train to* ***visit the zoo****. They* ***saw*** *a baby giraffe, a lion, and* ***a flock of*** *colorful tropical* ***birds****.*

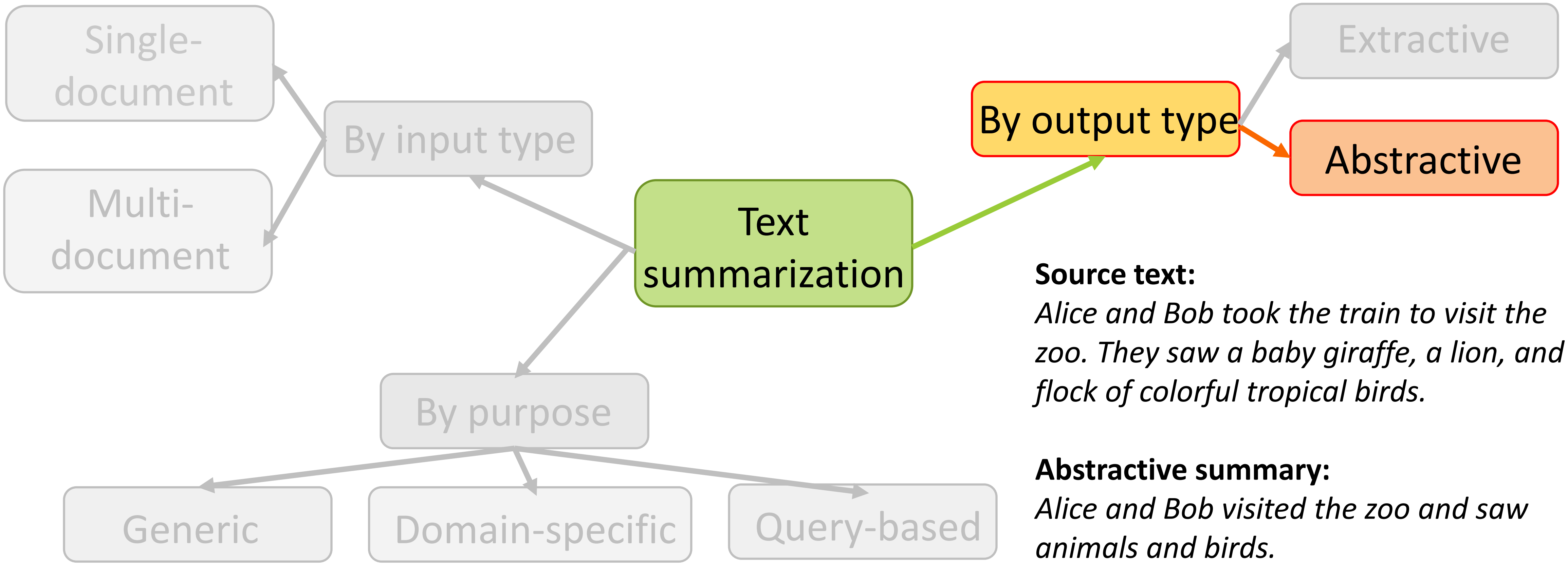
**Extractive summary:**
*Alice and Bob visit the zoo. saw a flock of birds.*

# Text summarization methods




**Generate** a coherent summary

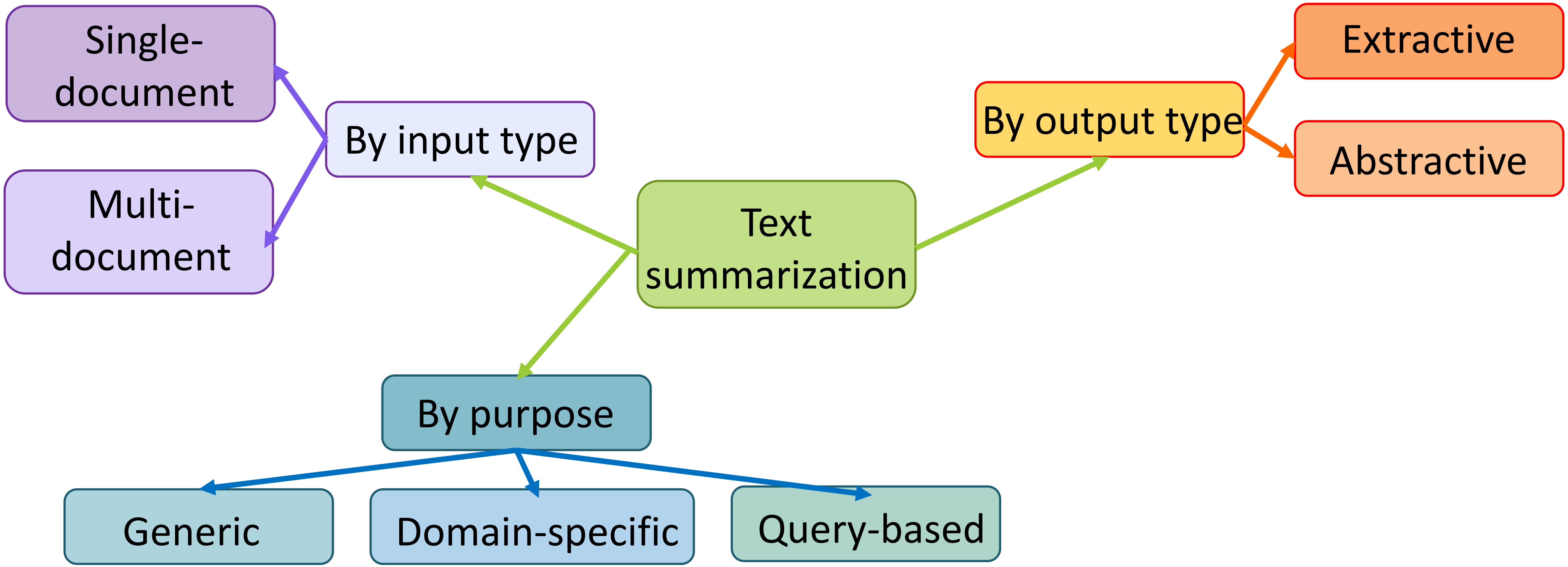
# Text summarization methods



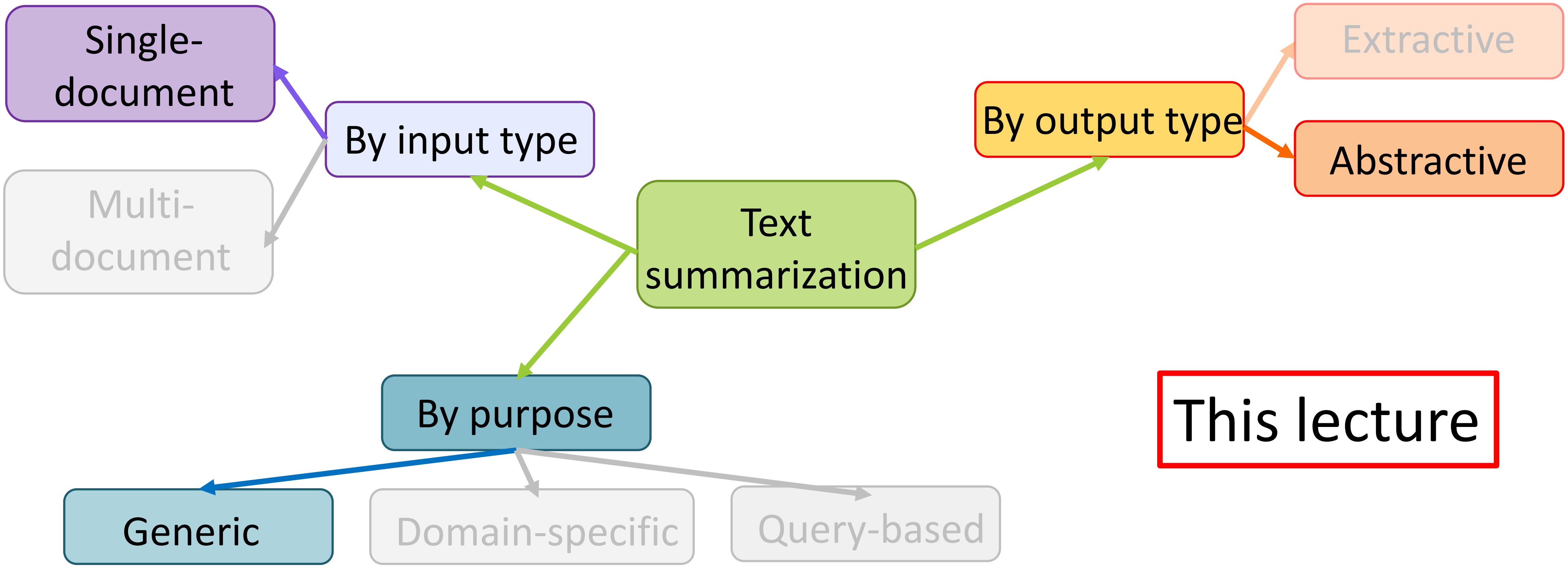

**Source text:**
*Alice and Bob took the train to visit the zoo. They saw a baby giraffe, a lion, and a flock of colorful tropical birds.*

**Abstractive summary:**
*Alice and Bob visited the zoo and saw animals and birds.*

# Text summarization methods



# Text summarization methods



# Extractive vs Abstractive

## Extractive Summarization

select sentences from the article

## Abstractive Summarization

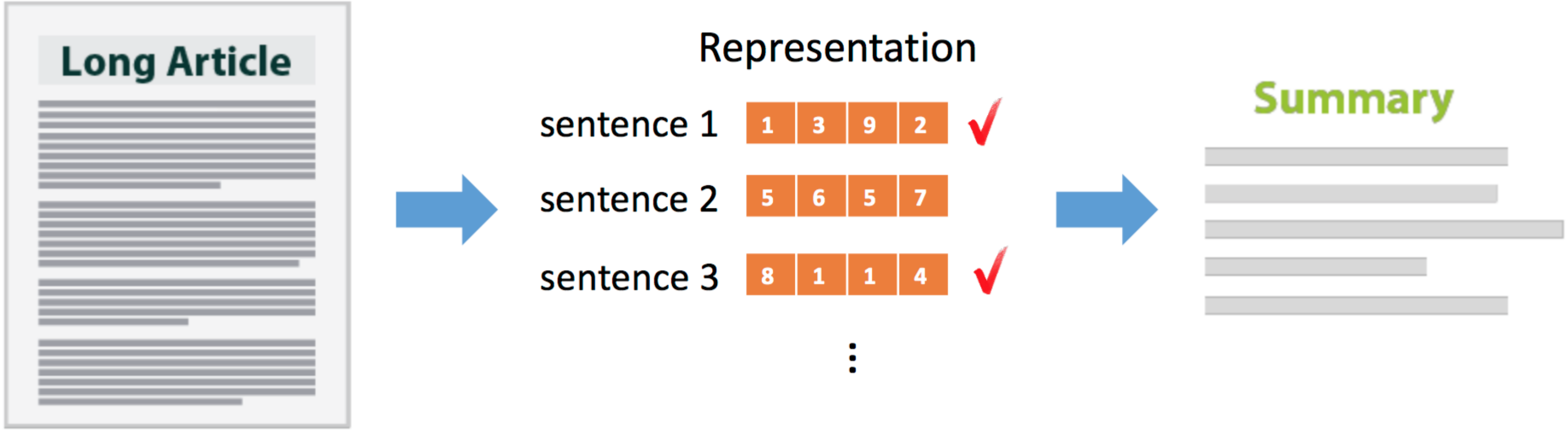
generate the summary word-by-word

From http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

# Extractive vs Abstractive

## Extractive Summarization

- Select phrases or sentences from the source document




From http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

# Extractive vs Abstractive

## Abstractive Summarization

- Select phrases or sentences from the source document


From http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

# Extractive vs Abstractive

## Motivation

- Extractive summary (select sentences):
  - important, correct
  - incoherent or not concise
- Abstractive summary (generate word-by-word):
  - readable, concise
  - may lose or mistake some facts

not concise

Italian artist Johannes Stoetter has painted two naked women to look like a chameleon.

The 37-year-old has previously transformed his models into frogs and parrots but this may be his most intricate and impressive artwork to date.

concise

Justin Bieber

~~Johannes Stoetter~~ has previously transformed his models into frogs and parrots but this chameleon may be his most impressive artwork to date.

From http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

# Evaluation

- Human evaluation
- ROUGE

ROUGE-N

$$= \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

- **ROUGE-N** – measures unigram, bigram, trigram and higher order n-gram overlap
- **ROUGE-L** – measures longest matching sequence of words using LCS. An advantage of using LCS is that it does not require consecutive matches but in-sequence matches that reflect sentence level word order. Since it automatically includes longest in-sequence common n-grams, you don't need a predefined n-gram length.

ROUGE paper: http://www.aclweb.org/anthology/W04-1013

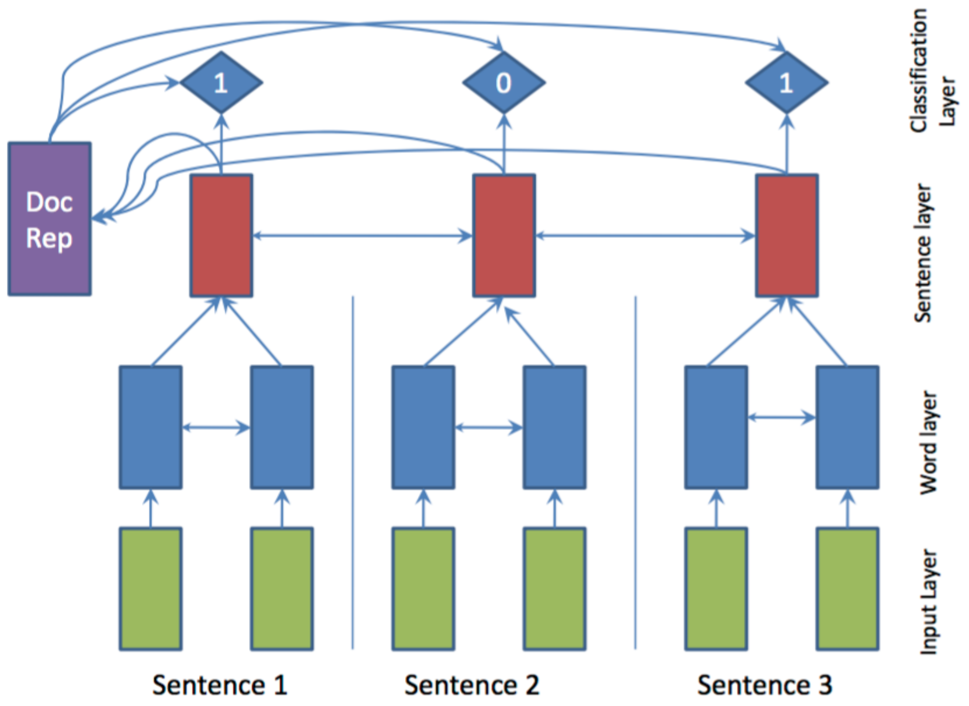
# A piece of
# Extractive Text Summarization

# SummaRuNNer: A RNN based Sequence Model for Extractive Summarization of Documents

$$
\begin{aligned}
P(y_j = 1|\mathbf{h}_j, \mathbf{s}_j, \mathbf{d}) = \sigma(&W_c\mathbf{h}_j && \#\texttt{(content)}\\
&+\mathbf{h}_j^T W_s \mathbf{d} && \#\texttt{(salience)}\\
&-\mathbf{h}_j^T W_r \tanh(\mathbf{s_j}) && \#\texttt{(novelty)}\\
&+W_{ap}\mathbf{p}_j^a && \#\texttt{(abs. pos. imp.)}\\
&+W_{rp}\mathbf{p}_j^r && \#\texttt{(rel. pos. imp.)}\\
&+b), && \#\texttt{(bias term)}
\end{aligned}
$$

$$\mathbf{d} = \tanh(W_d \frac{1}{N_d} \sum_{j=1}^{N^d} [\mathbf{h}_j^f, \mathbf{h}_j^b] + \mathbf{b})$$ - Document representation

$$\mathbf{s}_j = \sum_{i=1}^{j-1} \mathbf{h}_i P(y_i = 1|\mathbf{h}_i, \mathbf{s}_i, \mathbf{d})$$ - Summary representation

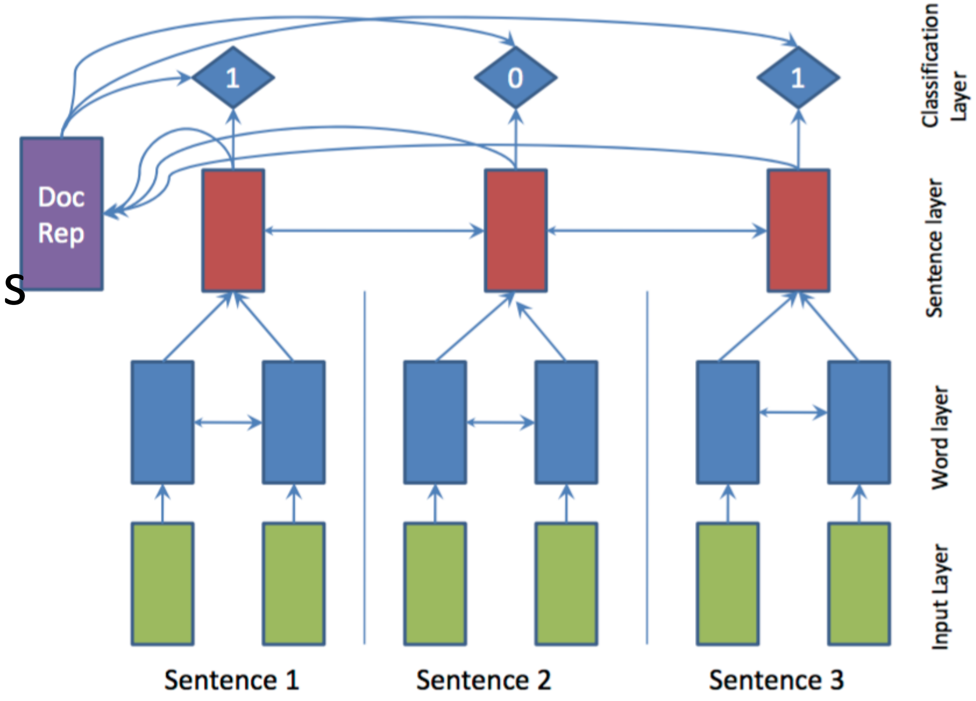



Nallapati et al, AAAI 2017, https://arxiv.org/abs/1611.04230

# SummaRuNNer: A RNN based Sequence Model for Extractive Summarization of Documents

## Extractive training

Need binary labels: how do we get them?
We have only human-generated summaries



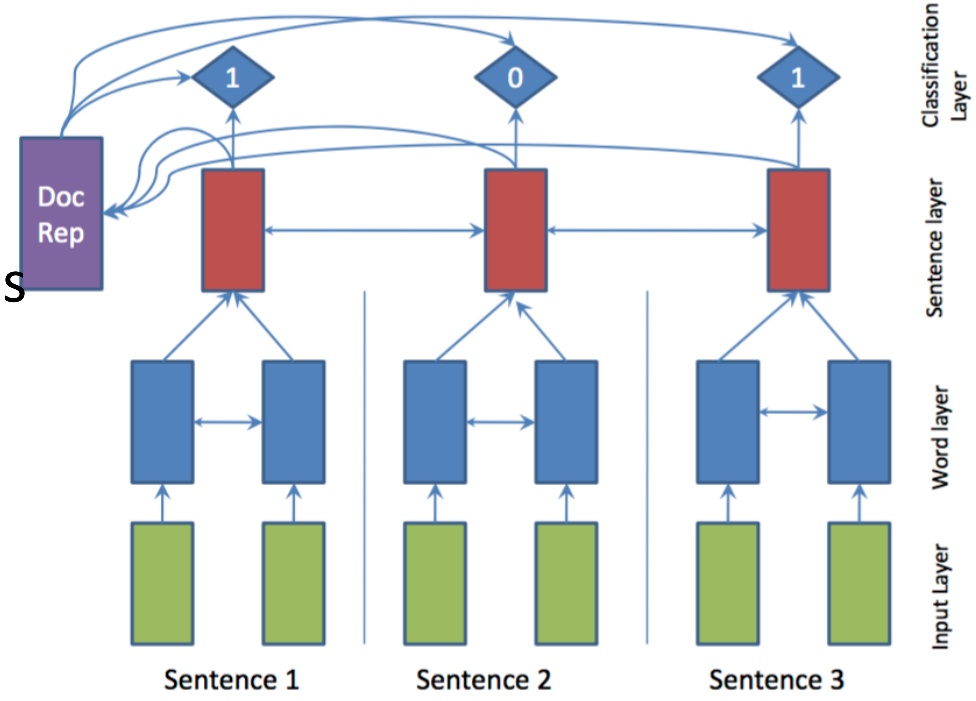

Nallapati et al, AAAI 2017, https://arxiv.org/abs/1611.04230

# SummaRuNNer: A RNN based Sequence Model for Extractive Summarization of Documents

## Extractive training

Need binary labels: how do we get them?
We have only human-generated summaries

Greedily add sentences with the highest ROUGE



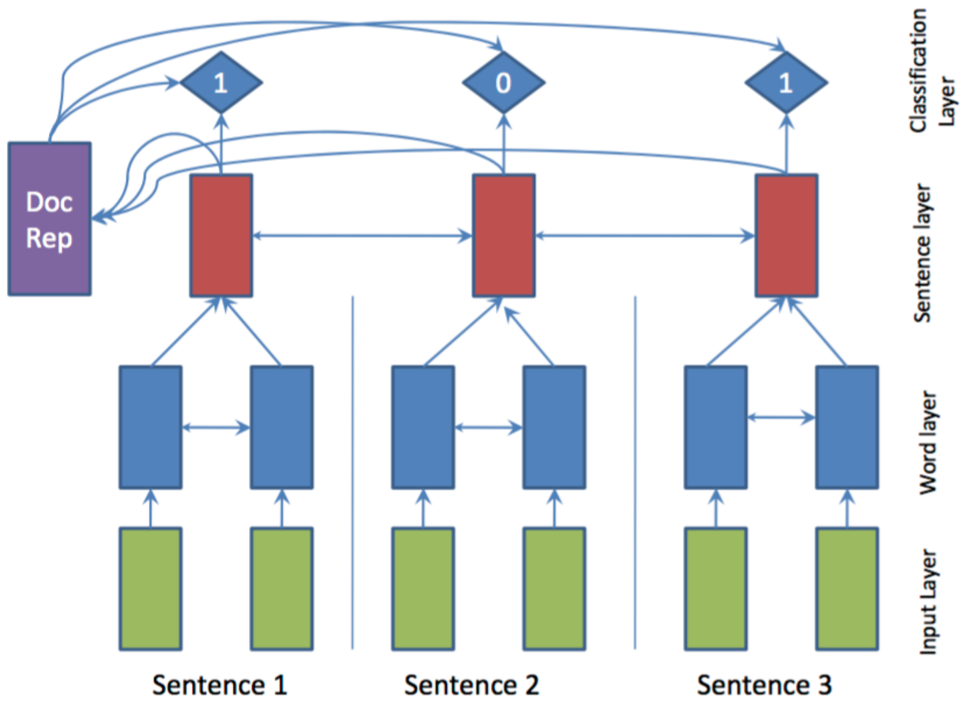

Nallapati et al, AAAI 2017, https://arxiv.org/abs/1611.04230

# SummaRuNNer: A RNN based Sequence Model for Extractive Summarization of Documents

## Abstractive training

Why would we want abstractive training?

We then use binary labels




Nallapati et al, AAAI 2017, https://arxiv.org/abs/1611.04230

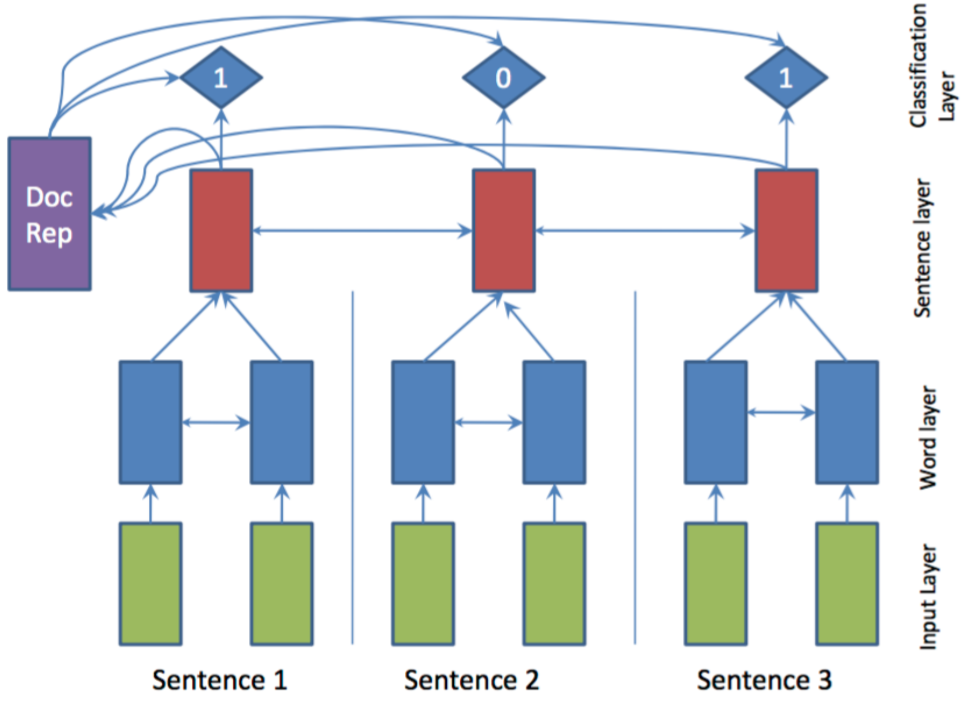
# SummaRuNNer: A RNN based Sequence Model for Extractive Summarization of Documents

## Abstractive training

Why would we want abstractive training?

We then use binary labels

No need to generate approximate abstractive labels!



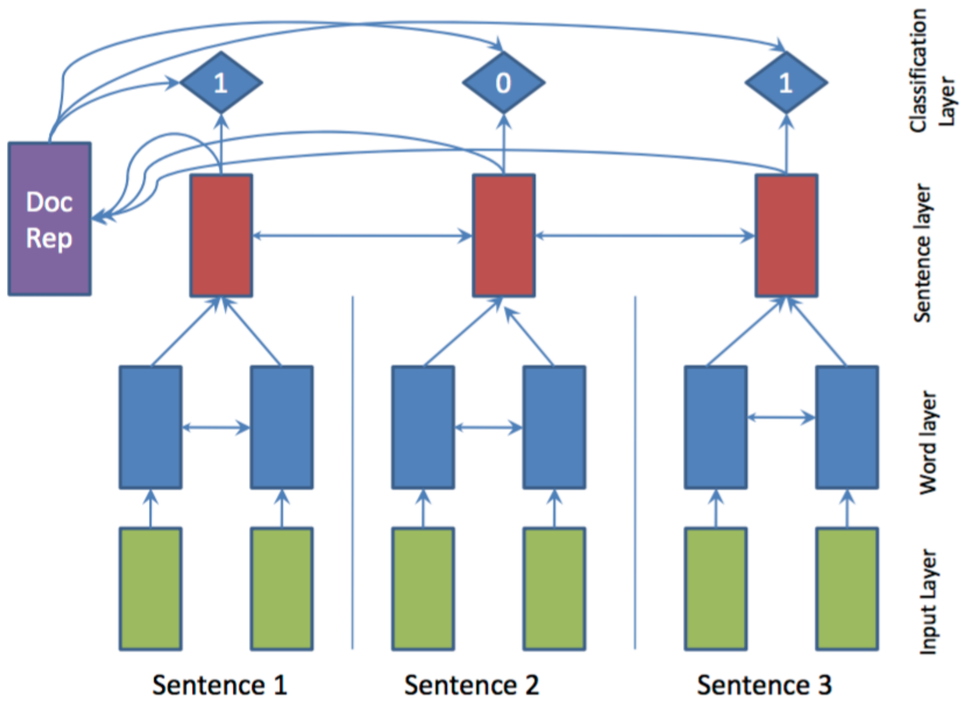

Nallapati et al, AAAI 2017, https://arxiv.org/abs/1611.04230

# SummaRuNNer: A RNN based Sequence Model for Extractive Summarization of Documents

## Abstractive training

No need to generate approximate abstractive labels!

- ➢ Add decoder (in training only!)
- ➢ Use last summary representation from encoder




Nallapati et al, AAAI 2017, https://arxiv.org/abs/1611.04230

# SummaRuNNer: A RNN based Sequence Model for Extractive Summarization of Documents

| **Gold Summary:** Redpath has ended his eight-year association with Sale Sharks. Redpath spent five years as a player and three as a coach at sale. He has thanked the owners, coaches and players for their support. | Salience | Content | Novelty | Position | Prob. |
|---|---|---|---|---|---|
| Bryan Redpath has left his coaching role at Sale Sharks with immediate effect. | 0.1 | 0.1 | 0.9 | 0.1 | 0.3 |
| The 43 - year - old Scot ends an eight-year association with the Aviva Premiership side, having spent five years with them as a player and three as a coach. | 0.9 | 0.6 | 0.9 | 0.9 | 0.7 |
| Redpath returned to Sale in June 2012 as director of rugby after starting a coaching career at Gloucester and progressing to the top job at Kingsholm . | 0.8 | 0.5 | 0.5 | 0.9 | 0.6 |
| Redpath spent five years with Sale Sharks as a player and a further three as a coach but with Sale Sharks struggling four months into Redpath's tenure, he was removed from the director of rugby role at the Salford-based side and has since been operating as head coach . | 0.8 | 0.9 | 0.7 | 0.8 | **0.9** |
| 'I would like to thank the owners, coaches, players and staff for all their help and support since I returned to the club in 2012. | 0.4 | 0.1 | 0.1 | 0.7 | 0.2 |
| Also to the supporters who have been great with me both as a player and as a coach,' Redpath said. | 0.6 | 0.0 | 0.2 | 0.3 | 0.2 |

Nallapati et al, AAAI 2017, https://arxiv.org/abs/1611.04230

# SummaRuNNer: A RNN based Sequence Model for Extractive Summarization of Documents

| | Rouge-1 | Rouge-2 | Rouge-L |
|---|---|---|---|
| Lead-3 | 39.2 | 15.7 | **35.5** |
| (Nallapati et al. 2016) | 35.4 | 13.3 | 32.6 |
| SummaRuNNer-abs | 37.5 | 14.5 | 33.4 |
| SummaRuNNer | **39.6**±0.2* | **16.2**±0.2* | 35.3±0.2 |

Table 3: Performance comparison of abstractive and extractive models on the entire CNN Daily Mail test set using **full-length F1** variants of Rouge. SummaRuNNer is able to significantly outperform the abstractive state-of-the-art as well as the Lead-3 baseline (on Rouge-1 and Rouge-2).

Nallapati et al, AAAI 2017, https://arxiv.org/abs/1611.04230

# SummaRuNNer: A RNN based Sequence Model for Extractive Summarization of Documents

| |
|---|
| *Document:* **@entity0 have an interest in @entity3 defender @entity2 but are unlikely to make a move until january . the 00 - year - old @entity6 captain has yet to open talks over a new contract at @entity3 and his current deal runs out in 0000 .** @entity3 defender @entity2 could be targeted by @entity0 in the january transfer window @entity0 like @entity2 but do n't expect @entity3 to sell yet they know he will be free to talk to foreign clubs from january . @entity12 will make a 0million offer for @entity3 goalkeeper @entity14 this summer . the 00 - year - old is poised to leave @entity16 and wants to play for a @entity18 contender . **@entity12 are set to make a 0million bid for @entity2 's @entity3 team - mate @entity14 in the summer** |
| *Gold Summary:* @entity2 's contract at @entity3 expires at the end of next season . 00 - year - old has yet to open talks over a new deal at @entity16 . @entity14 is poised to leave @entity3 at the end of the season |

Nallapati et al, AAAI 2017, https://arxiv.org/abs/1611.04230

# SummaRuNNer: A RNN based Sequence Model for Extractive Summarization of Documents

*Document:* **today , the foreign ministry said that control operations carried out by the corvette spiro against a korean-flagged as received ship fishing illegally in argentine waters were carried out " in accordance with international law and in coordination with the foreign ministry " .** the foreign ministry thus approved the intervention by the argentine corvette when it discovered the korean ship chin yuan hsing violating argentine jurisdictional waters on 00 may . ... **the korean ship , which had been fishing illegally in argentine waters , was sunk by its own crew after failing to answer to the argentine ship 's warnings .** the crew was transferred to the chin chuan hsing , which was sailing nearby and approached to rescue the crew of the sinking ship .....

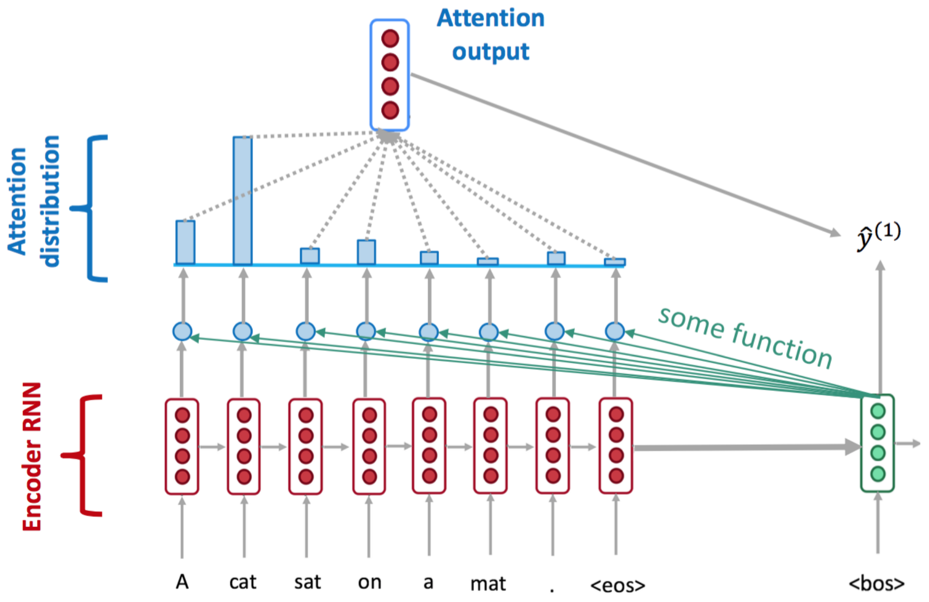
*Gold Summary:* the korean-flagged fishing vessel chin yuan hsing was scuttled in waters off argentina on 00 may 0000 . adverse weather conditions prevailed when the argentine corvette spiro spotted the korean ship fishing illegally in restricted argentine waters . the korean vessel did not respond to the corvette 's warning . instead , the korean crew sank their ship , and transferred to another korean ship sailing nearby . in accordance with a uk-argentine agreement , the argentine navy turned the surveillance of the second korean vessel over to the british when it approached within 00 nautical miles of the malvinas ( falkland ) islands .

Nallapati et al, AAAI 2017, https://arxiv.org/abs/1611.04230

# Abstractive Text Summarization

# A Neural Attention Model for Sentence Summarization

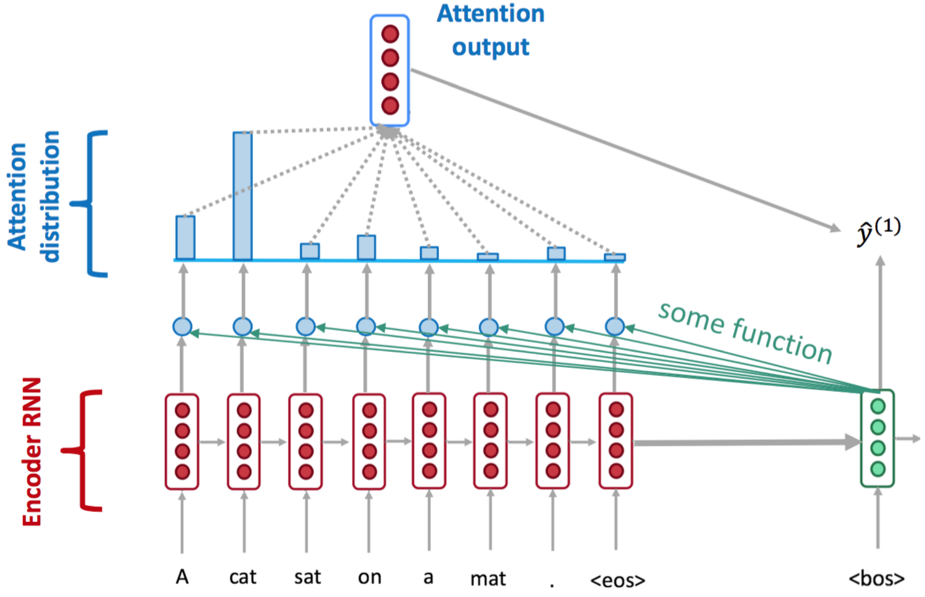
Model: RNNs with attention



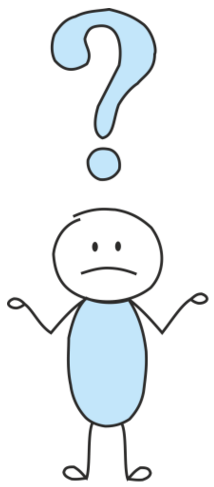

Rush et al, EMNLP 2015 https://www.aclweb.org/anthology/D/D15/D15-1044.pdf

# A Neural Attention Model for Sentence Summarization

Model: RNNs with attention



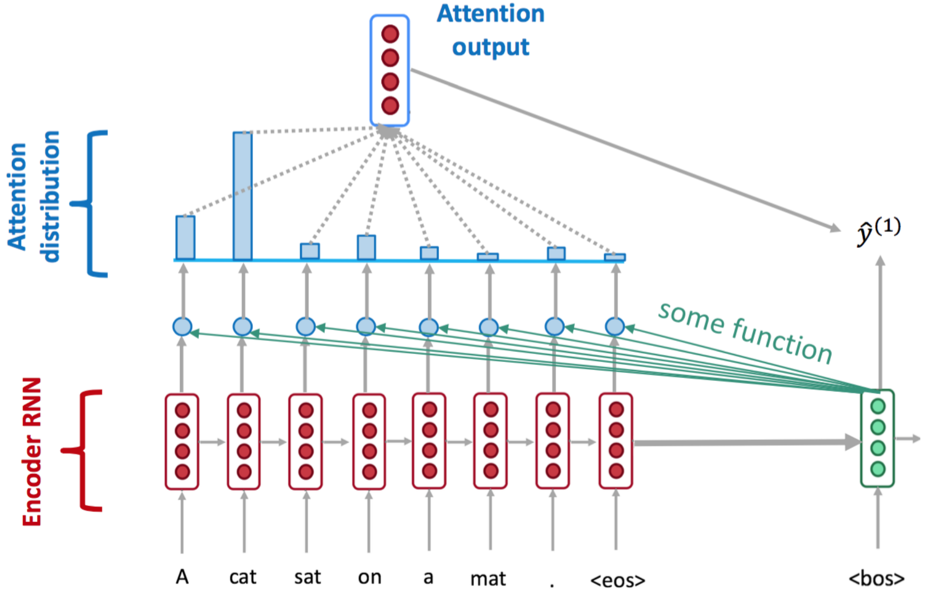

What is especially hard for this model comparing to extractive ones?



Rush et al, EMNLP 2015 https://www.aclweb.org/anthology/D/D15/D15-1044.pdf

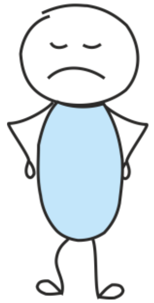
# A Neural Attention Model for Sentence Summarization

Model: RNNs with attention




What is especially hard for this model comparing to extractive ones?

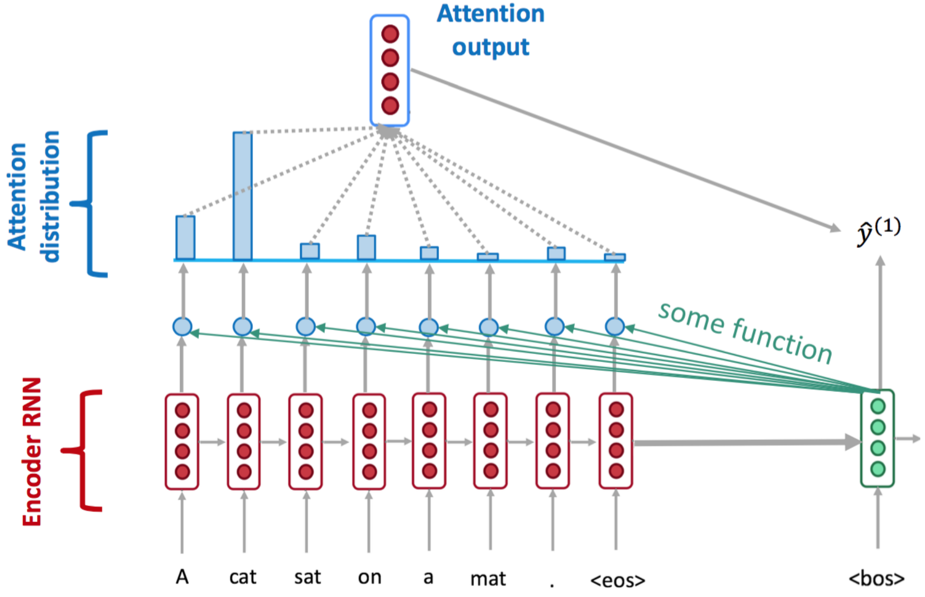
Hard to handle unseen proper noun phrase (named entities)



Rush et al, EMNLP 2015 https://www.aclweb.org/anthology/D/D15/D15-1044.pdf

# A Neural Attention Model for Sentence Summarization

Model: RNNs with attention




Solution: Extractive Tuning

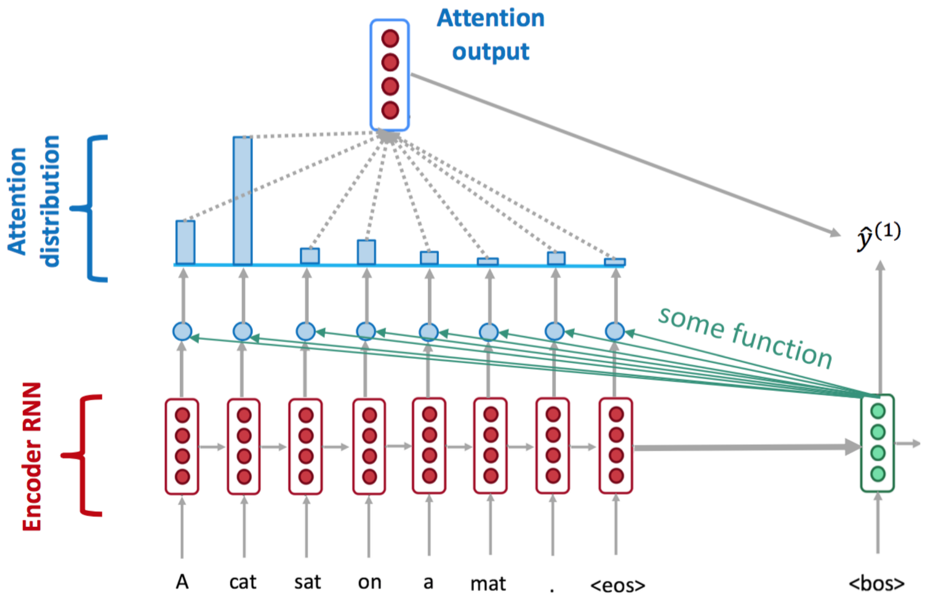
$$p(\mathbf{y}|\mathbf{x};\theta,\alpha) \quad \propto \quad \exp(\alpha^\top \sum_{i=0}^{N-1} f(\mathbf{y}_{i+1},\mathbf{x},\mathbf{y}_c))$$

$$\begin{aligned} f(\mathbf{y}_{i+1},\mathbf{x},\mathbf{y}_c) = [\, & \log p(\mathbf{y}_{i+1}|\mathbf{x},\mathbf{y}_c;\theta), \\ & \mathbb{1}\{\exists j.\ \mathbf{y}_{i+1} = \mathbf{x}_j\ \}, \\ & \mathbb{1}\{\exists j.\ \mathbf{y}_{i+1-k} = \mathbf{x}_{j-k}\ \forall k \in \{0,1\}\}, \\ & \mathbb{1}\{\exists j.\ \mathbf{y}_{i+1-k} = \mathbf{x}_{j-k}\ \forall k \in \{0,1,2\}\}, \\ & \mathbb{1}\{\exists k > j.\ \mathbf{y}_i = \mathbf{x}_k, \mathbf{y}_{i+1} = \mathbf{x}_j\}\ ]. \end{aligned}$$

Rush et al, EMNLP 2015 https://www.aclweb.org/anthology/D/D15/D15-1044.pdf

# A Neural Attention Model for Sentence Summarization

## Model: RNNs with attention




## Solution: Extractive Tuning

$$p(\mathbf{y}|\mathbf{x};\theta,\alpha) \quad \propto \quad \exp(\alpha^\top \sum_{i=0}^{N-1} f(\mathbf{y}_{i+1},\mathbf{x},\mathbf{y}_c))$$

$$\begin{aligned} f(\mathbf{y}_{i+1},\mathbf{x},\mathbf{y}_c) = [\, & \log p(\mathbf{y}_{i+1}|\mathbf{x},\mathbf{y}_c;\theta), \\ & \left.\begin{aligned} & \mathbb{1}\{\exists j.\ \mathbf{y}_{i+1} = \mathbf{x}_j\ \}, \\ & \mathbb{1}\{\exists j.\ \mathbf{y}_{i+1-k} = \mathbf{x}_{j-k}\ \forall k \in \{0,1\}\}, \\ & \mathbb{1}\{\exists j.\ \mathbf{y}_{i+1-k} = \mathbf{x}_{j-k}\ \forall k \in \{0,1,2\}\}, \\ & \mathbb{1}\{\exists k > j.\ \mathbf{y}_i = \mathbf{x}_k, \mathbf{y}_{i+1} = \mathbf{x}_j\}\ ]. \end{aligned}\right\} \text{N-gram match} \end{aligned}$$

Rush et al, EMNLP 2015 https://www.aclweb.org/anthology/D/D15/D15-1044.pdf

# A Neural Attention Model for Sentence Summarization

| Model | DUC-2004 ROUGE-1 | DUC-2004 ROUGE-2 | DUC-2004 ROUGE-L | Gigaword ROUGE-1 | Gigaword ROUGE-2 | Gigaword ROUGE-L | Gigaword Ext. % |
|---|---|---|---|---|---|---|---|
| IR | 11.06 | 1.67 | 9.67 | 16.91 | 5.55 | 15.58 | 29.2 |
| PREFIX | 22.43 | 6.49 | 19.65 | 23.14 | 8.25 | 21.73 | 100 |
| COMPRESS | 19.77 | 4.02 | 17.30 | 19.63 | 5.13 | 18.28 | 100 |
| W&L | 22 | 6 | 17 | - | - | - | - |
| TOPIARY | 25.12 | 6.46 | 20.12 | - | - | - | - |
| MOSES+ | 26.50 | 8.13 | 22.85 | 28.77 | 12.10 | 26.44 | 70.5 |
| ABS | 26.55 | 7.06 | 22.05 | 30.88 | 12.22 | 27.77 | 85.4 |
| ABS+ | 28.18 | 8.49 | 23.81 | 31.00 | 12.65 | 28.34 | 91.5 |
| REFERENCE | 29.21 | 8.38 | 24.46 | - | - | - | 45.6 |

Table 1: Experimental results on the main summary tasks on various ROUGE metrics . Baseline models are described in detail in Section 7.2. We report the percentage of tokens in the summary that also appear in the input for Gigaword as `Ext %`.

Rush et al, EMNLP 2015 https://www.aclweb.org/anthology/D/D15/D15-1044.pdf

# A Neural Attention Model for Sentence Summarization

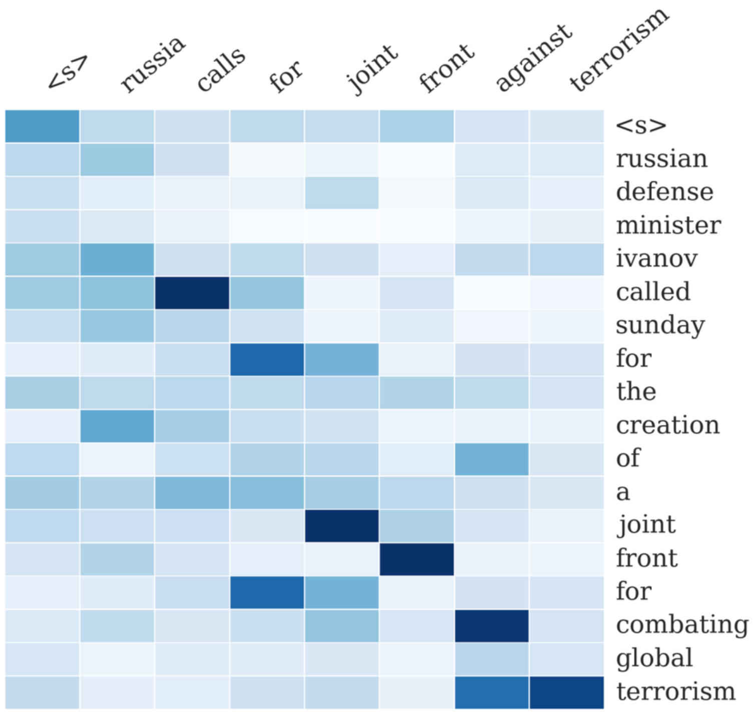



Figure 1: Example output of the attention-based summarization (ABS) system. The heatmap represents a soft alignment between the input (right) and the generated summary (top). The columns represent the distribution over the input after generating each word.

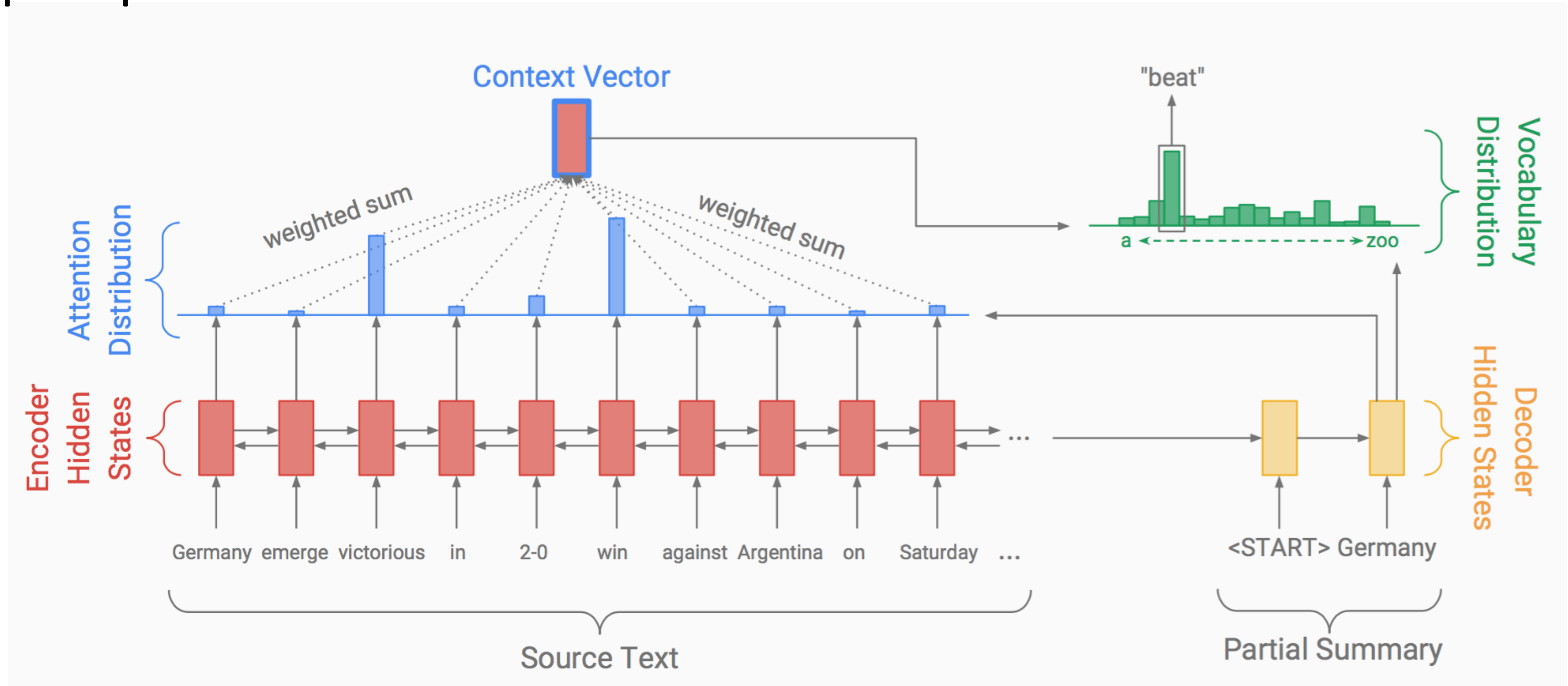
**I(8):** thousands of kashmiris chanting pro-pakistan slogans on sunday attended a rally to welcome back a hardline separatist leader who underwent cancer treatment in mumbai .
**G:** thousands attend rally for kashmir hardliner
**A:** thousands rally in support of hardline kashmiri separatist leader
**A+:** thousands of kashmiris rally to welcome back cancer treatment

**I(7):** the white house on thursday warned iran of possible new sanctions after the un nuclear watchdog reported that tehran had begun sensitive nuclear work at a key site in defiance of un resolutions .
**G:** us warns iran of step backward on nuclear issue
**A:** iran warns of possible new sanctions on nuclear work
**A+:** un nuclear watchdog warns iran of possible new sanctions

Rush et al, EMNLP 2015 https://www.aclweb.org/anthology/D/D15/D15-1044.pdf

# Get To The Point: Summarization with Pointer-Generator Networks

Seq2seq with attention:




From the authors' slides: http://anthology.aclweb.org/attachments/P/P17/P17-1099.Presentation.pdf

See et al, ACL 2017, http://aclweb.org/anthology/P17-1099

# Get To The Point: Summarization with Pointer-Generator Networks

**Problem 1:** The summaries sometimes reproduce factual details inaccurately.

e.g. *Germany beat Argentina 3-2*

Incorrect **rare** or **out-of-vocabulary** word

**Problem 2:** The summaries sometimes repeat themselves.

e.g. *Germany beat Germany beat Germany beat…*

From the authors' slides: http://anthology.aclweb.org/attachments/P/P17/P17-1099.Presentation.pdf

See et al, ACL 2017, http://aclweb.org/anthology/P17-1099

# Get To The Point: Summarization with Pointer-Generator Networks

**Problem 1:** The summaries sometimes reproduce factual details inaccurately.

e.g. *Germany beat Argentina 3-2*

Incorrect **rare** or **out-of-vocabulary** word

**Solution:** Use a pointer to copy words.

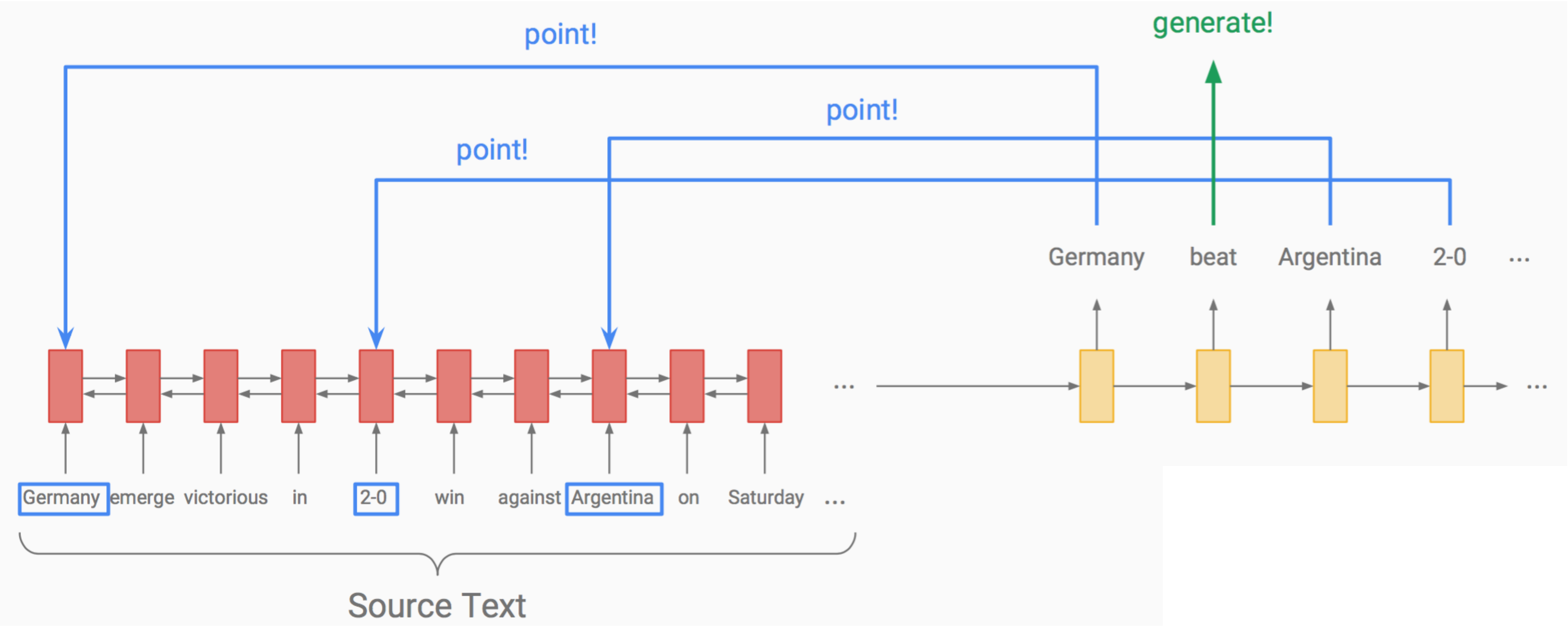
**Problem 2:** The summaries sometimes repeat themselves.

e.g. *Germany beat Germany beat Germany beat…*

From the authors' slides: http://anthology.aclweb.org/attachments/P/P17/P17-1099.Presentation.pdf

See et al, ACL 2017, http://aclweb.org/anthology/P17-1099

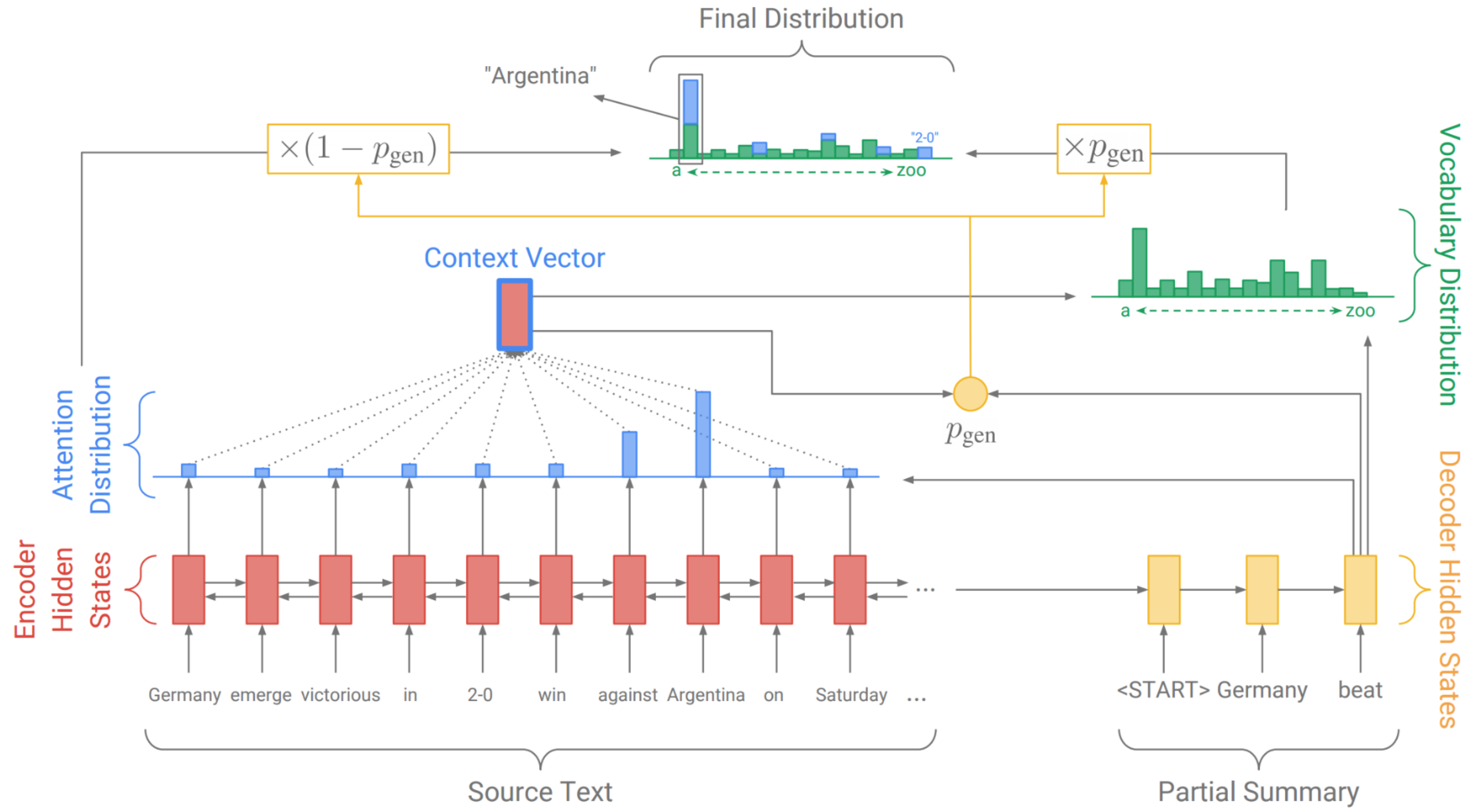
# Get To The Point: Summarization with Pointer-Generator Networks




From the authors' slides: http://anthology.aclweb.org/attachments/P/P17/P17-1099.Presentation.pdf

See et al, ACL 2017, http://aclweb.org/anthology/P17-1099

# Get To The Point: Summarization with Pointer-Generator Networks




From the authors' slides: http://anthology.aclweb.org/attachments/P/P17/P17-1099.Presentation.pdf

See et al, ACL 2017, http://aclweb.org/anthology/P17-1099

# Get To The Point: Summarization with Pointer-Generator Networks

| Before | After |
| --- | --- |
| *UNK UNK was expelled from the dubai open chess tournament* | *gaioz nigalidze was expelled from the dubai open chess tournament* |
| *the 2015 rio olympic games* | *the 2016 rio olympic games* |

From the authors' slides: http://anthology.aclweb.org/attachments/P/P17/P17-1099.Presentation.pdf

See et al, ACL 2017, http://aclweb.org/anthology/P17-1099

# Get To The Point: Summarization with Pointer-Generator Networks

**Problem 1:** The summaries sometimes reproduce factual details inaccurately.

e.g. *Germany beat Argentina 3-2*

**Solution:** Use a pointer to copy words.

**Problem 2:** The summaries sometimes repeat themselves.

e.g. *Germany beat Germany beat Germany beat…*

From the authors' slides: http://anthology.aclweb.org/attachments/P/P17/P17-1099.Presentation.pdf

See et al, ACL 2017, http://aclweb.org/anthology/P17-1099

# Get To The Point: Summarization with Pointer-Generator Networks

**Problem 1:** The summaries sometimes reproduce factual details inaccurately.

e.g. *Germany beat Argentina 3-2*

**Solution:** Use a pointer to copy words.

**Problem 2:** The summaries sometimes repeat themselves.

e.g. *Germany beat Germany beat Germany beat…*

**Solution:** Penalize repeatedly attending to same parts of the source text.

From the authors' slides: http://anthology.aclweb.org/attachments/P/P17/P17-1099.Presentation.pdf

See et al, ACL 2017, http://aclweb.org/anthology/P17-1099

# Get To The Point: Summarization with Pointer-Generator Networks

*Coverage* = cumulative attention = what has been covered so far

**Source Text:** Germany emerge victorious in 2-0 win against Argentina on Saturday

**Summary:** Germany beat ______

From the authors' slides: http://anthology.aclweb.org/attachments/P/P17/P17-1099.Presentation.pdf

See et al, ACL 2017, http://aclweb.org/anthology/P17-1099

# Get To The Point: Summarization with Pointer-Generator Networks

*Coverage* = cumulative attention = what has been covered so far

**Source Text:** Germany emerge victorious in 2-0 win against Argentina on Saturday

**Summary:** Germany beat ______

1. Use coverage as extra input to attention mechanism.

From the authors’ slides: http://anthology.aclweb.org/attachments/P/P17/P17-1099.Presentation.pdf

See et al, ACL 2017, http://aclweb.org/anthology/P17-1099

# Get To The Point: Summarization with Pointer-Generator Networks


1. Use coverage as extra input to attention mechanism.
2. Penalize attending to things that have already been covered.

From the authors' slides: http://anthology.aclweb.org/attachments/P/P17/P17-1099.Presentation.pdf

See et al, ACL 2017, http://aclweb.org/anthology/P17-1099

# Get To The Point: Summarization with Pointer-Generator Networks

*Coverage* = cumulative attention = what has been covered so far

Don't attend here

**Source Text:** Germany emerge victorious in 2-0 win against Argentina on Saturday

**Summary:** Germany beat _____

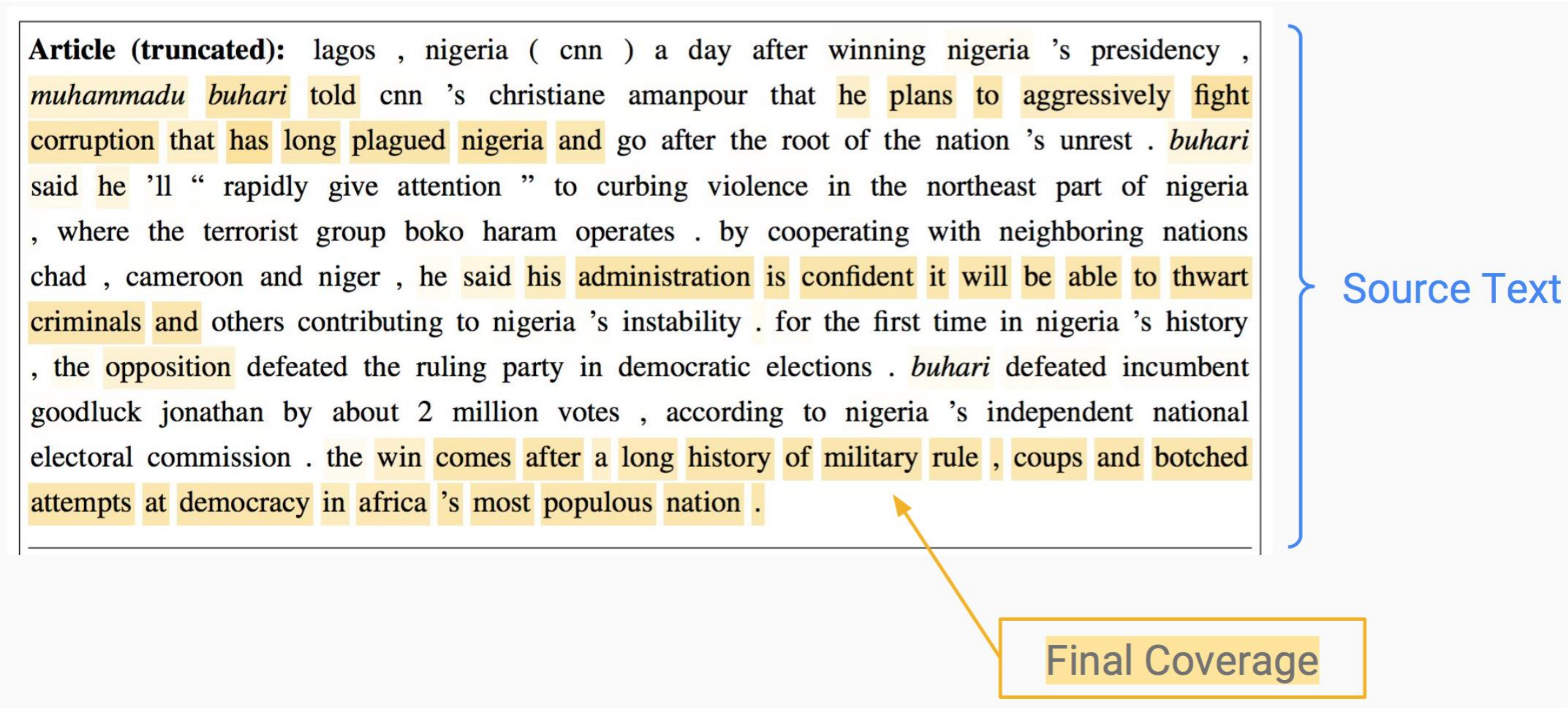
1. Use coverage as extra input to attention mechanism.
2. Penalize attending to things that have already been covered.

**Result:** repetition rate reduced to level similar to human summaries

From the authors' slides: http://anthology.aclweb.org/attachments/P/P17/P1

See et al, ACL 2017, http://aclweb.org/anthology/P17-1099

# Get To The Point: Summarization with Pointer-Generator Networks




From the authors' slides: http://anthology.aclweb.org/attachments/P/P17/P17-1099.Presentation.pdf

See et al, ACL 2017, http://aclweb.org/anthology/P17-1099

# Get To The Point: Summarization with Pointer-Generator Networks

ROUGE compares the machine-generated summary to the human-written reference summary and counts co-occurrence of 1-grams, 2-grams, and longest common sequence.

| | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| Nallapati et al. 2016 | 35.5 | 13.3 | 32.7 |
| Ours (seq2seq baseline) | 31.3 | 11.8 | 28.8 |
| Ours (pointer-generator) | 36.4 | 15.7 | 33.4 |
| Ours (pointer-generator + coverage) | **39.5** | **17.3** | **36.4** |

Previous best abstractive result

Our improvements

From the authors' slides: http://anthology.aclweb.org/attachments/P/P17/P17-1099.Presentation.pdf

See et al, ACL 2017, http://aclweb.org/anthology/P17-1099

# A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss

## Motivation

- Extractive summary (select sentences):
  - important, correct
  - incoherent or not concise
- Abstractive summary (generate word-by-word):
  - readable, concise
  - may lose or mistake some facts
- Unified summary:
  - important, correct
  - readable, concise

not concise

Italian artist Johannes Stoetter has painted two naked women to look like a chameleon.

The 37-year-old has previously transformed his models into frogs and parrots but this may be his most intricate and impressive artwork to date.

concise

Justin Bieber

~~Johannes Stoetter~~ has previously transformed his models into frogs and parrots but this chameleon may be his most impressive artwork to date.

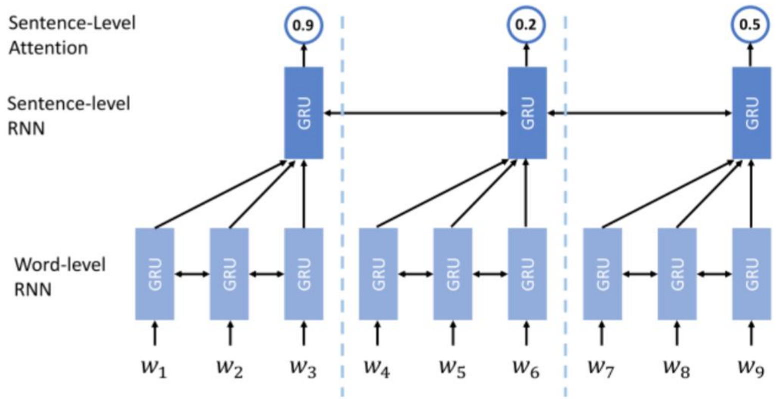
From the authors' slides: http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

Hsu et al, ACL 2018, http://aclweb.org/anthology/P18-1013

# A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss

## Models

### Extractor



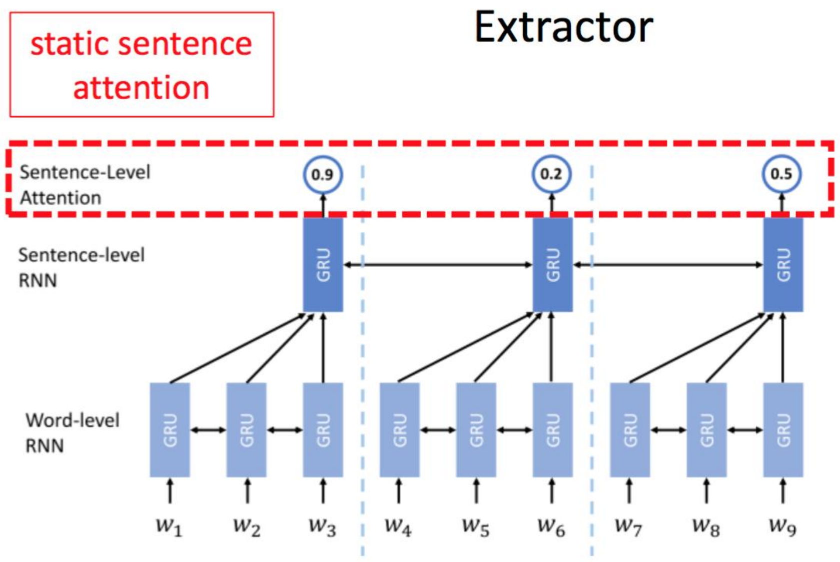

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. AAAI 2017

From the authors' slides: http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

Hsu et al, ACL 2018, http://aclweb.org/anthology/P18-1013

# A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss

## Models



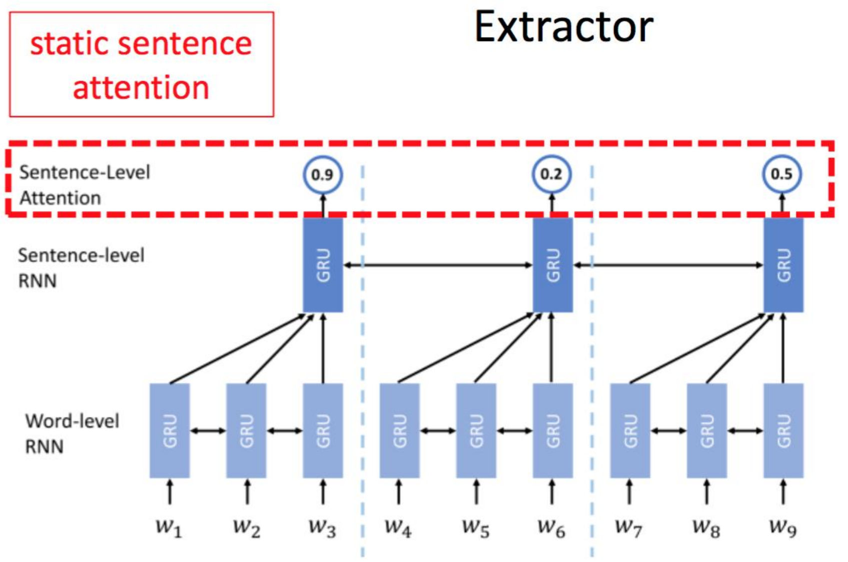

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. AAAI 2017

From the authors' slides: http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

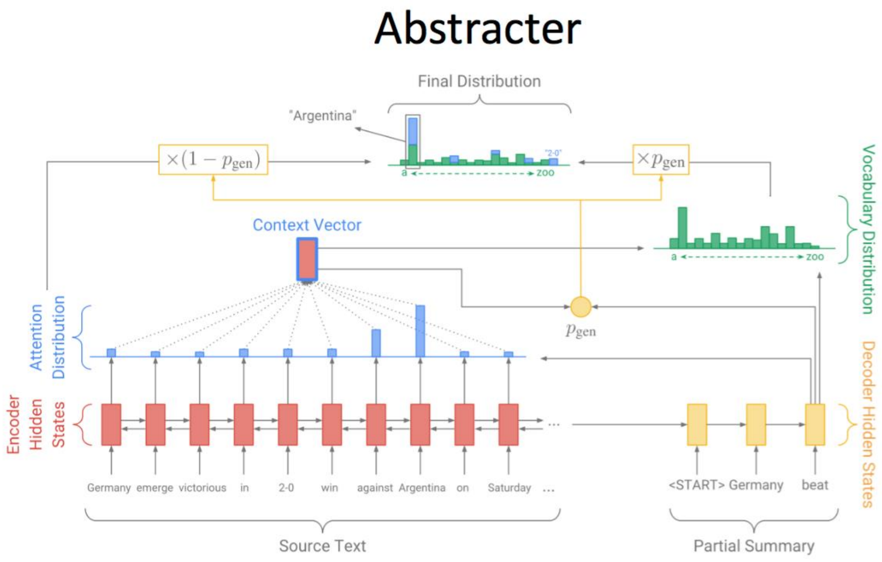
Hsu et al, ACL 2018, http://aclweb.org/anthology/P18-1013

# A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss

## Models




Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. AAAI 2017



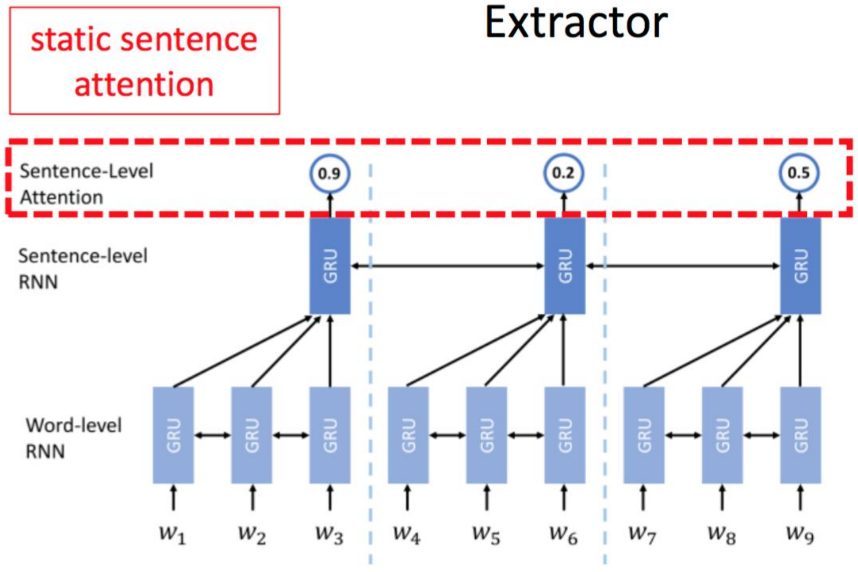

Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. ACL 2017

From the authors' slides: http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

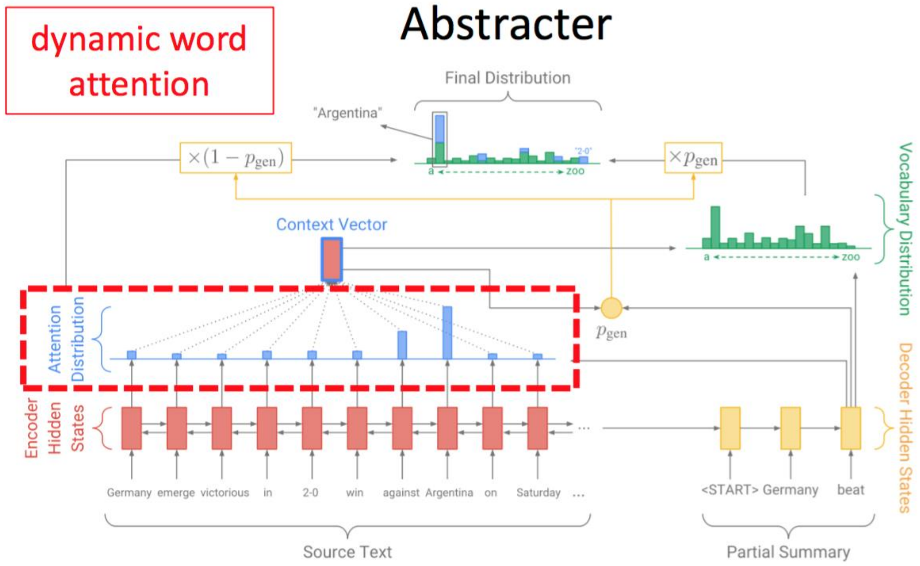
Hsu et al, ACL 2018, http://aclweb.org/anthology/P18-1013

# A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss

## Models




Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. AAAI 2017



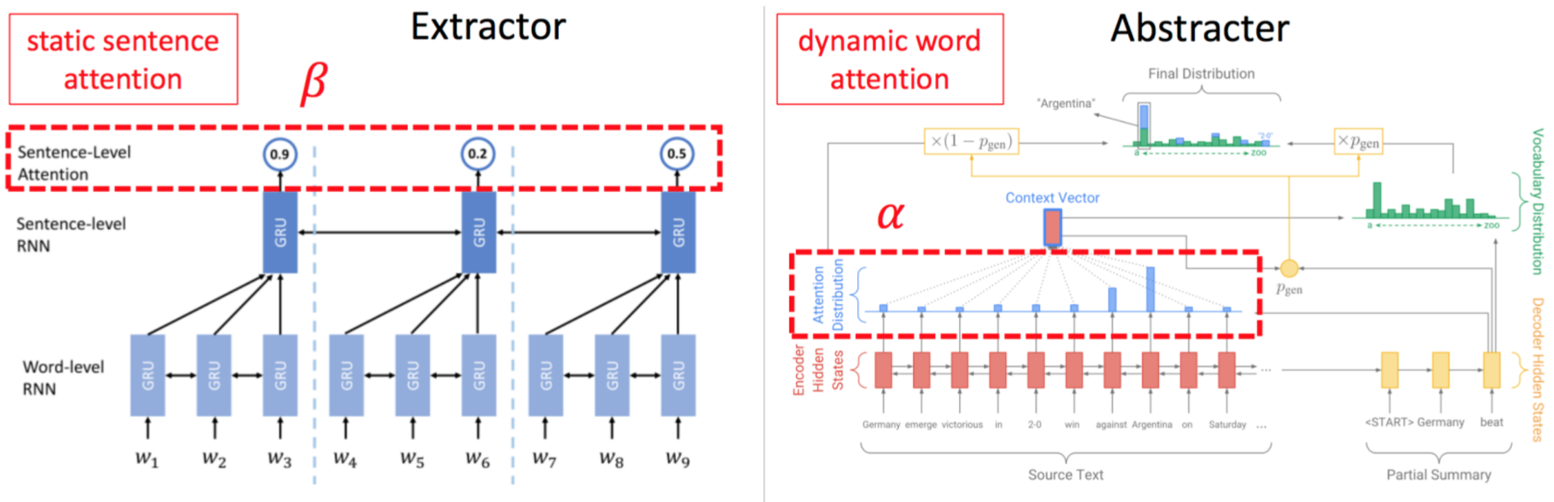

Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. ACL 2017

From the authors' slides: http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

Hsu et al, ACL 2018, http://aclweb.org/anthology/P18-1013

# A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss

## Combined Attention




$$\boxed{\hat{\alpha}_m^t = \frac{\alpha_m^t \times \beta_{n(m)}}{\sum_m \alpha_m^t \times \beta_{n(m)}}.}$$

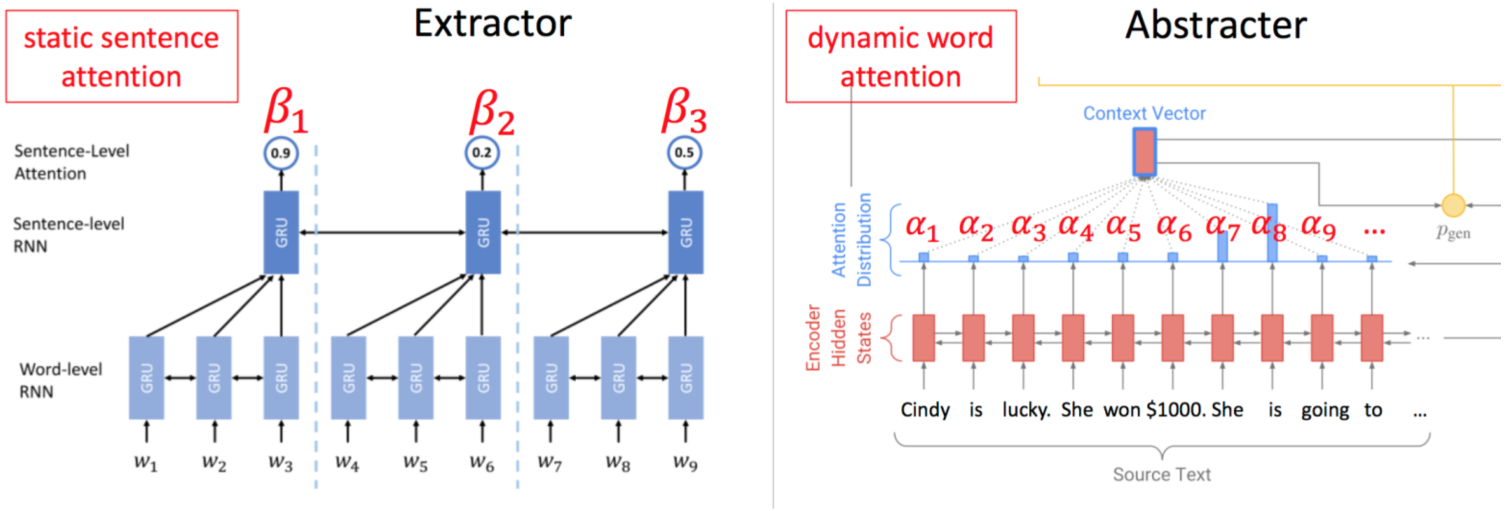
$m$: word index
$n$: sentence index
$t$: generated word index

From the authors' slides: http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

Hsu et al, ACL 2018, http://aclweb.org/anthology/P18-1013

# A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss

## Combined Attention

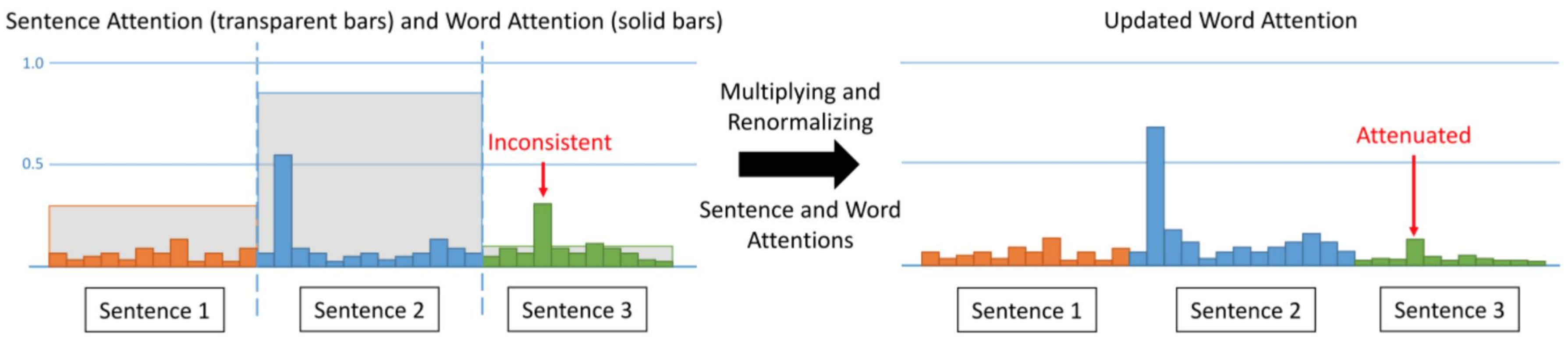


# A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss

## Combined Attention

$$\hat{\alpha}_m^t = \frac{\alpha_m^t \times \beta_{n(m)}}{\sum_m \alpha_m^t \times \beta_{n(m)}}.$$

- Our unified model combines sentence-level and word-level attentions to take advantage of both extractive and abstractive summarization approaches.




From the authors' slides: http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

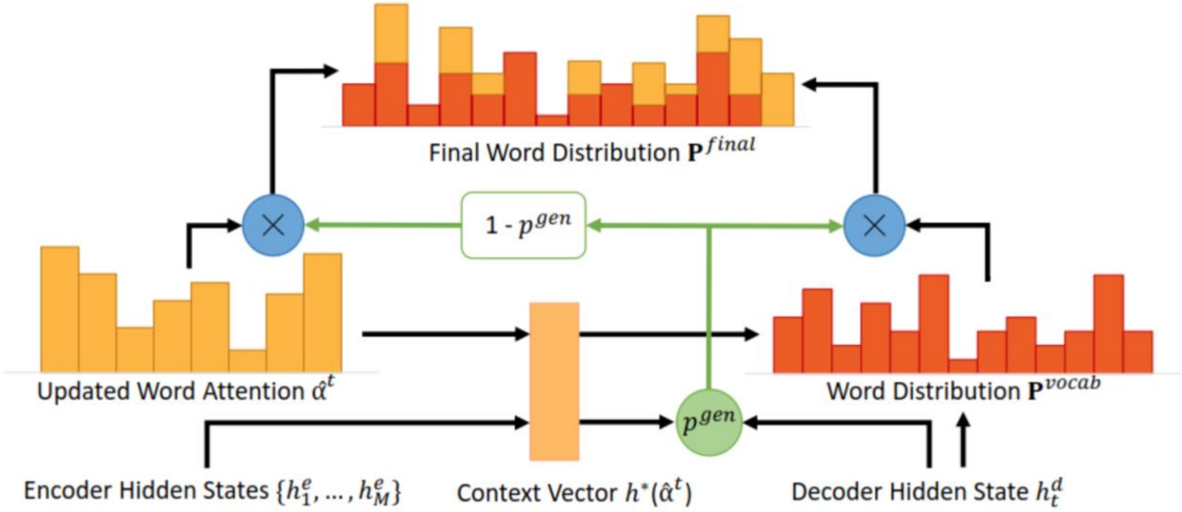
Hsu et al, ACL 2018, http://aclweb.org/anthology/P18-1013

# A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss

## Combined Attention

$$\hat{\alpha}_m^t = \frac{\alpha_m^t \times \beta_{n(m)}}{\sum_m \alpha_m^t \times \beta_{n(m)}}.$$

- Updated word attention is used for calculating the context vector and final word distribution




From the authors' slides: http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

Hsu et al, ACL 2018, http://aclweb.org/anthology/P18-1013

# A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss

## Encourage Consistency

- We propose a novel inconsistency loss function to ensure our unified model to be mutually beneficial to both extractive and abstractive summarization.

multiplied attention of top K attended words

$$L_{inc} = -\frac{1}{T}\sum_{t=1}^{T}\log\left(\frac{1}{|\mathcal{K}|}\boxed{\sum_{m\in\mathcal{K}}\alpha_m^t \times \beta_{n(m)}}\right)$$

maximize ↑

where $\mathcal{K}$ is the set of top K attended words

From the authors' slides: http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

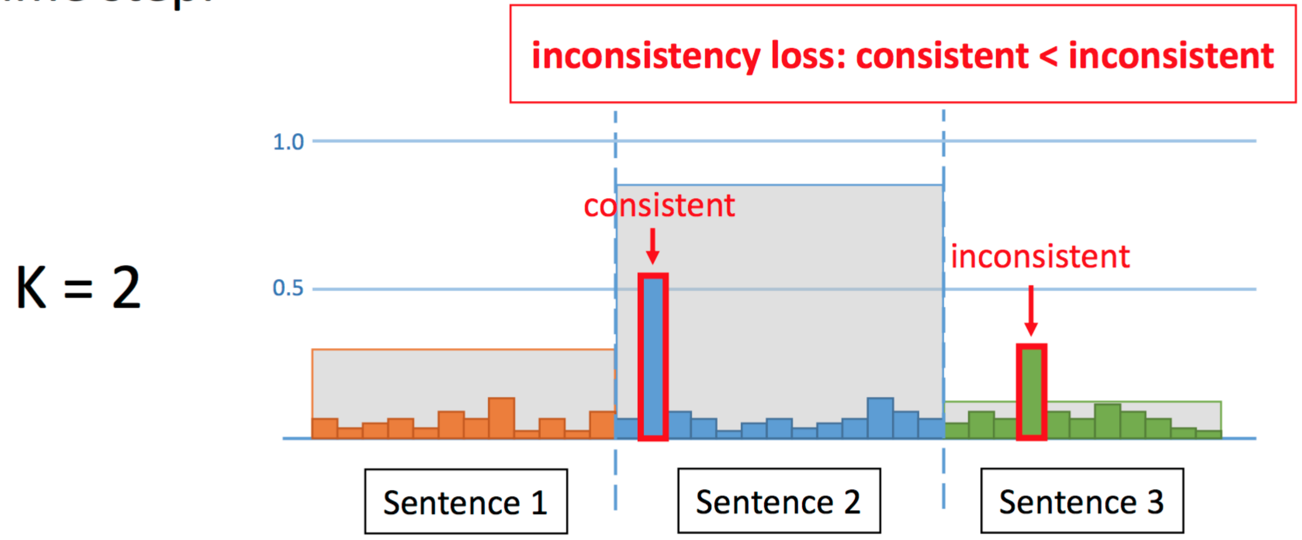
Hsu et al, ACL 2018, http://aclweb.org/anthology/P18-1013

# A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss

## Encourage Consistency

$$L_{inc} = -\frac{1}{T}\sum_{t=1}^{T}\log\left(\frac{1}{|\mathcal{K}|}\sum_{m\in\mathcal{K}}\alpha_m^t \times \beta_{n(m)}\right)$$

- encourage consistency of the top K attended words at each decoder time step.




From the authors' slides: http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

Hsu et al, ACL 2018, http://aclweb.org/anthology/P18-1013

# A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss

- 3 types of loss functions:

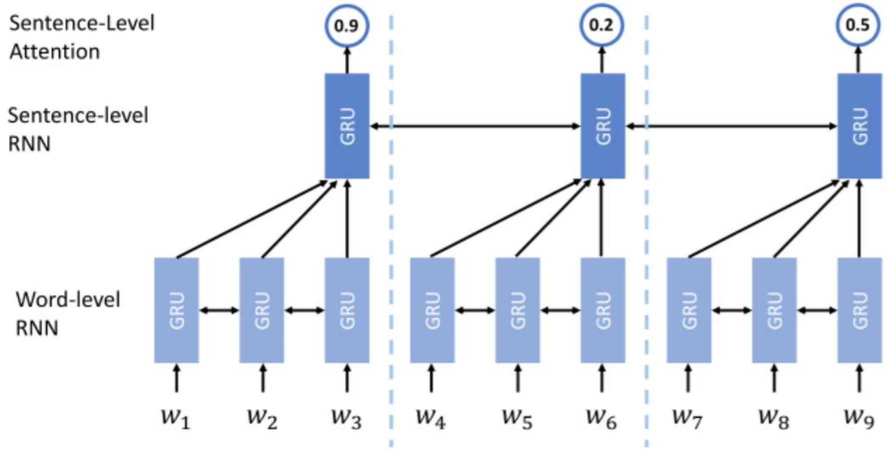
1. extractor loss
2. abstracter loss
   + coverage loss
3. inconsistency loss

From the authors' slides: http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

Hsu et al, ACL 2018, http://aclweb.org/anthology/P18-1013

# A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss

- 3 types of loss functions:
  1. extractor loss
  2. abstracter loss
     + coverage loss
  3. inconsistency loss




$$L_{ext} = -\frac{1}{N}\sum_{n=1}^{N}(g_n \log \beta_n + (1-g_n)\log(1-\beta_n))$$

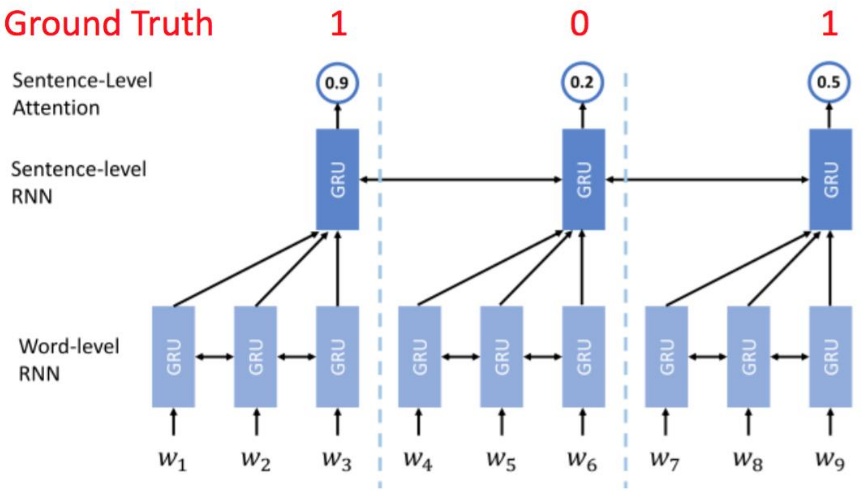
where $g_n \in \{0,1\}$ is the ground-truth label for the $n^{th}$ sentence and $N$ is the number of sentences.

From the authors' slides: http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

Hsu et al, ACL 2018, http://aclweb.org/anthology/P18-1013

# A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss

- 3 types of loss functions:
    1. extractor loss
    2. abstracter loss
       + coverage loss
    3. inconsistency loss




$$L_{ext} = -\frac{1}{N}\sum_{n=1}^{N}(g_n \log \beta_n + (1 - g_n)\log(1-\beta_n))$$

where $g_n \in \{0, 1\}$ is the ground-truth label for the $n^{th}$ sentence and $N$ is the number of sentences.

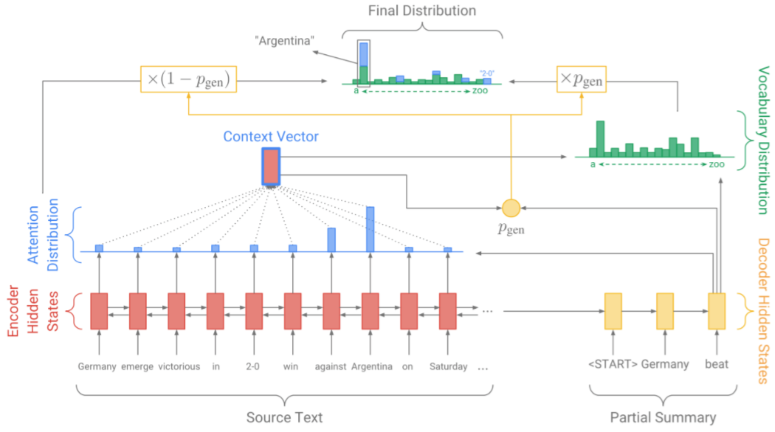
From the authors' slides: http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

Hsu et al, ACL 2018, http://aclweb.org/anthology/P18-1013

# A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss

- 3 types of loss functions:

  1. extractor loss
  2. abstracter loss
     + coverage loss
  3. inconsistency loss




$$L_{abs} = -\frac{1}{T}\sum_{t=1}^{T}\log P^{final}_{\hat{y}^t}$$

$$L_{cov} = \frac{1}{T}\sum_{t=1}^{T}\sum_{m=1}^{M}\min(\hat{\alpha}^t_m, c^t_m)$$

$$\mathbf{c}^t = \sum_{t'=1}^{t-1}\hat{\boldsymbol{\alpha}}^{t'}$$

From the authors' slides: http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

Hsu et al, ACL 2018, http://aclweb.org/anthology/P18-1013

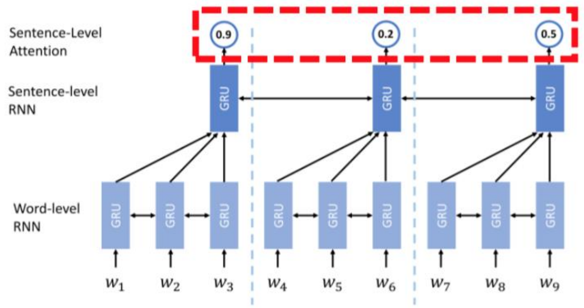
# A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss

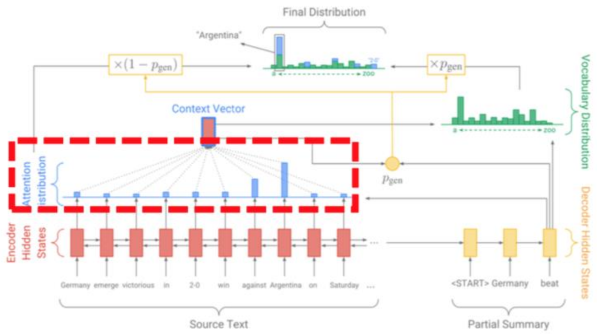
- 3 types of loss functions:

  1. extractor loss
  2. abstracter loss
     + coverage loss
  3. inconsistency loss

$$L_{inc} = -\frac{1}{T}\sum_{t=1}^{T}\log\left(\frac{1}{|\mathcal{K}|}\sum_{m\in\mathcal{K}}\alpha_m^t \times \beta_{n(m)}\right)$$

where $\mathcal{K}$ is the set of top K attended words




From the authors' slides: http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

Hsu et al, ACL 2018, http://aclweb.org/anthology/P18-1013

# A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss

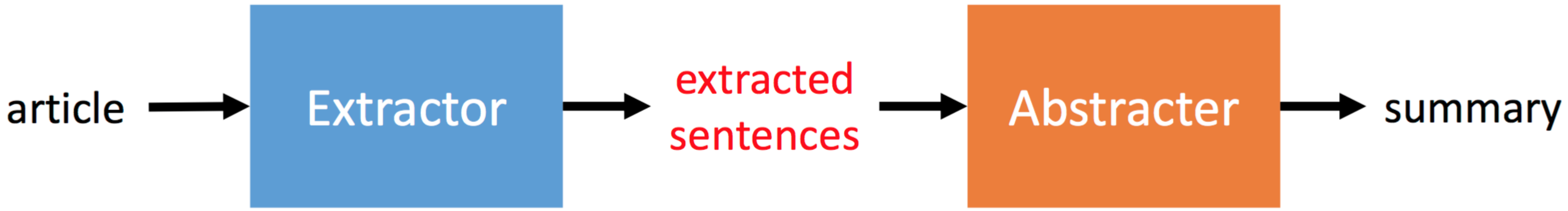
1. Two-stages training
2. End-to-end training without inconsistency loss
3. End-to-end training with inconsistency loss

From the authors' slides: http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

Hsu et al, ACL 2018, http://aclweb.org/anthology/P18-1013

# A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss

1. Two-stages training
   - The extractor is used as a classifier to select sentences with high informativity and output only those sentences. = Hard attention on the original article.
   - simply combine the extractor and abstracter by feeding the extracted sentences to the abstracter.




From the authors' slides: http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

Hsu et al, ACL 2018, http://aclweb.org/anthology/P18-1013

# A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss

2. End-to-end training without inconsistency loss
   - the sentence-level attention is soft attention and will be combined with the word-level attention
   - minimize extractor loss and abstracter loss

$$L_{e2e} = \lambda L_{ext} + L_{abs} + L_{cov}$$

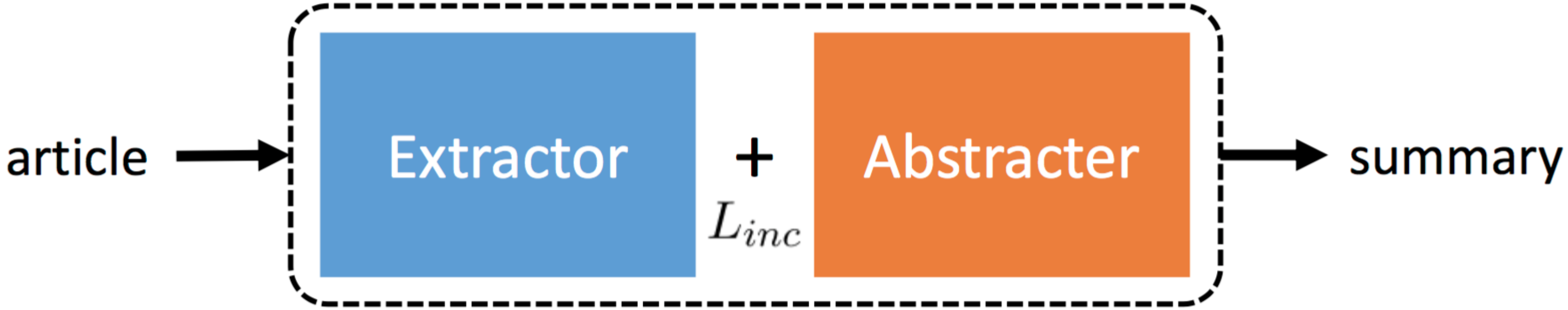

From the authors' slides: http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

Hsu et al, ACL 2018, http://aclweb.org/anthology/P18-1013

# A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss

3. End-to-end training with inconsistency loss
   - the sentence-level attention is soft attention and will be combined with the word-level attention
   - minimize extractor loss, abstracter loss and inconsistency loss:

$$L_{e2e} = \lambda L_{ext} + L_{abs} + L_{cov} \boxed{+ L_{inc}}$$




From the authors' slides: http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

Hsu et al, ACL 2018, http://aclweb.org/anthology/P18-1013

# A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss

| Method | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| HierAttn (Nallapati et al., 2016b)* | 32.75 | 12.21 | 29.01 |
| DeepRL (Paulus et al., 2017)* | 39.87 | 15.82 | 36.90 |
| pointer-generator (See et al., 2017) | 39.53 | 17.28 | 36.38 |
| GAN (Liu et al., 2017) | 39.92 | 17.65 | 36.71 |
| two-stage (ours) | 39.97 | 17.43 | 36.34 |
| end2end w/o inconsistency loss (ours) | 40.19 | 17.67 | 36.68 |
| end2end w/ inconsistency loss (ours) | **40.68** | **17.97** | **37.13** |
| lead-3 (See et al., 2017) | 40.34 | 17.70 | 36.57 |

Table 2: ROUGE F-1 scores of the generated abstractive summaries on the CNN/Daily Mail test set. Our two-stages model outperforms pointer-generator model on ROUGE-1 and ROUGE-2. In addition, our model trained end-to-end with inconsistency loss exceeds the lead-3 baseline. All our ROUGE scores have a 95% confidence interval with at most ±0.24. '*' indicates the model is trained and evaluated on the anonymized dataset and thus is not strictly comparable with ours.

From the authors' slides: http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

Hsu et al, ACL 2018, http://aclweb.org/anthology/P18-1013

# A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss

## Results – Human Evaluation

- Informativity: how well does the summary capture the important parts of the article?
- Conciseness: is the summary clear enough to explain everything without being redundant?
- Readability: how well-written (fluent and grammatical) the summary is?

| Method | informativity | conciseness | readability |
|---|---|---|---|
| DeepRL (Paulus et al., 2017) | 3.23 | 2.97 | 2.85 |
| pointer-generator (See et al., 2017) | 3.18 | 3.36 | 3.47 |
| GAN (Liu et al., 2017) | 3.22 | 3.52 | 3.51 |
| Ours | **3.58** | 3.40 | **3.70** |
| reference | 3.43 | **3.61** | 3.62 |

From the authors' slides: http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

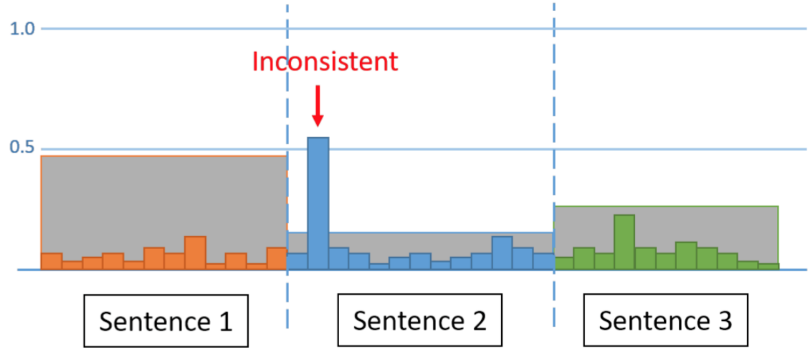
Hsu et al, ACL 2018, http://aclweb.org/anthology/P18-1013

# A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss

## Results – Inconsistency Rate $R_{inc}$

**inconsistency step $t_{inc}$:**

$$\beta_{n(\text{argmax}(\boldsymbol{\alpha}^t))} < \text{mean}(\boldsymbol{\beta})$$




sentence attention and word attention in time step $t$

**inconsistency rate:**

$$R_{inc} = \frac{\text{Count}(t_{inc})}{T}$$

where $T$ is the length of the summary.

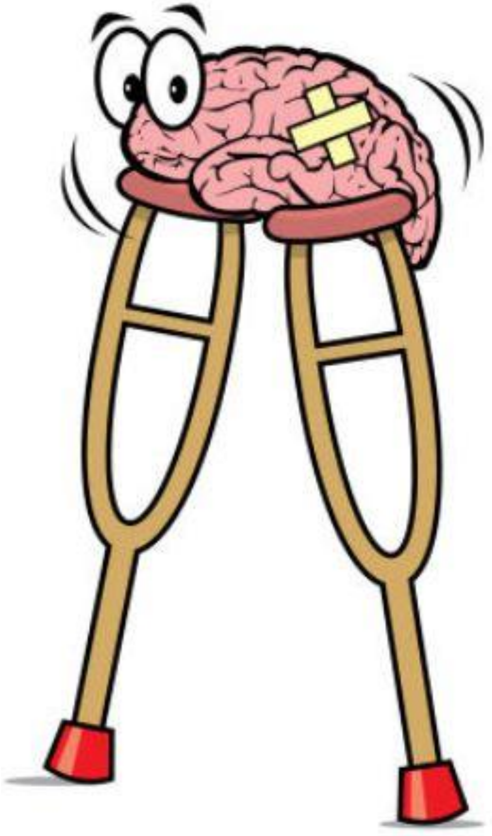
| Method | avg. $R_{inc}$ |
|---|---|
| w/o incon. loss | 0.198 |
| w/ incon. loss | 0.042 |

Table 3: Inconsistency rate of our end-to-end trained model with and without inconsistency loss.

From the authors' slides: http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

Hsu et al, ACL 2018, http://aclweb.org/anthology/P18-1013

# Hack of the day



# Learning Rate Scheduling

➢ Consider any optimizer, e.g. Adam

$$\mu_{t+1} = \beta_1 \cdot \mu_t + (1 - \beta_1) \cdot \nabla_\theta L$$

$$v_{t+1} = \beta_2 \cdot v_t + (1 - \beta_2) \cdot ||\, \nabla_\theta L||\,^2$$

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{v}_t + \epsilon} \mu_t$$

➢ The choice of α is crucial!

# Learning Rate Scheduling

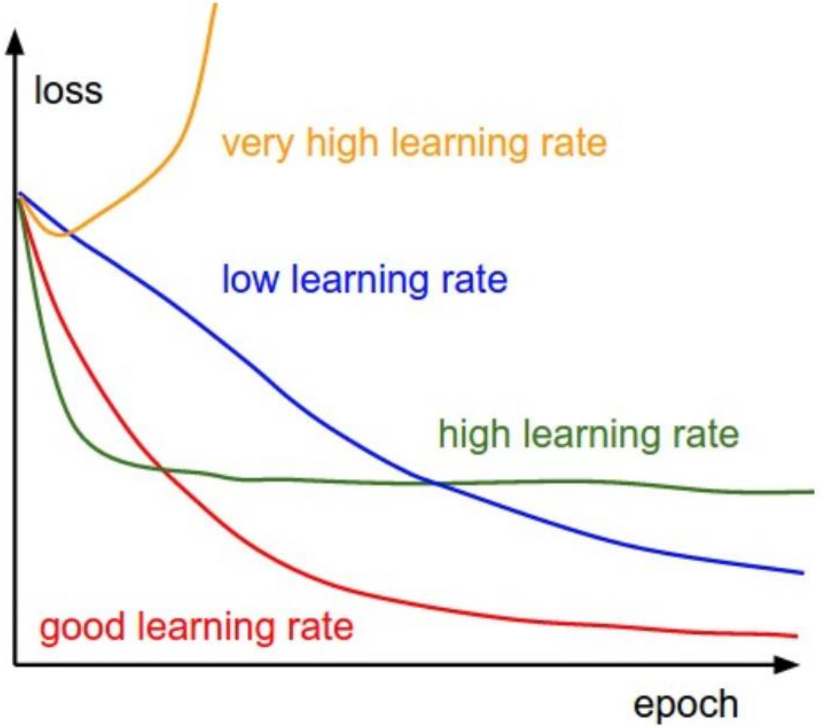
➢ Consider Adam optimizer

$$\mu_{t+1} = \beta_1 \cdot \mu_t + (1 - \beta_1) \cdot \nabla_\theta L$$

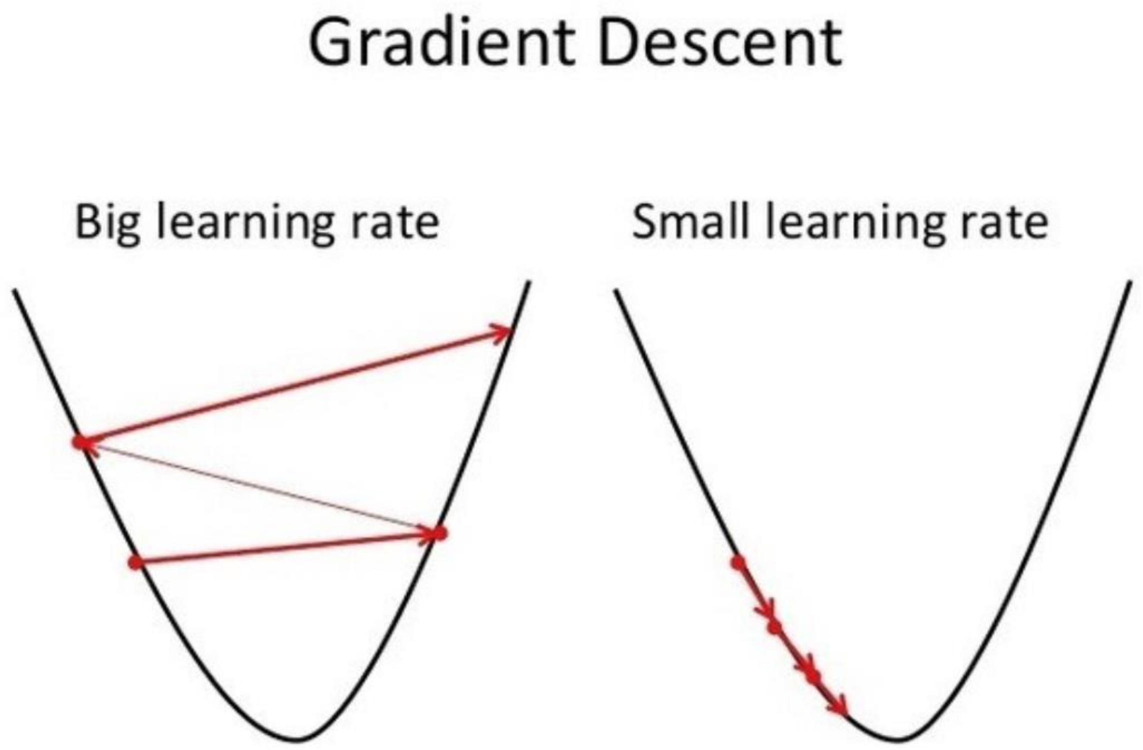
$$v_{t+1} = \beta_2 \cdot v_t + (1 - \beta_2) \cdot ||\nabla_\theta L||^2$$

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{v_t} + \epsilon} \mu_t$$

➢ **The choice of $\alpha$ is crucial!**



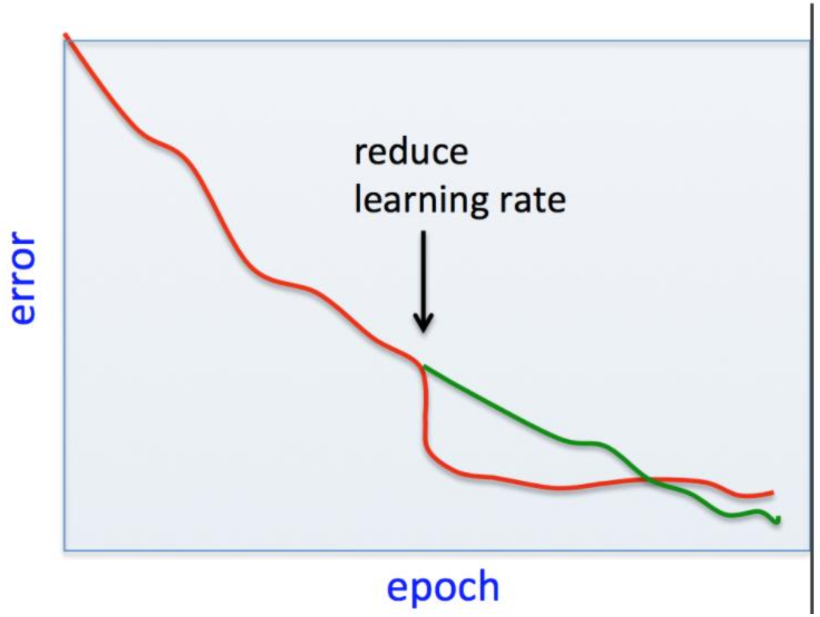
# Learning Rate Scheduling




Source: Andrew Ng’s Machine Learning course

# Learning Rate Scheduling

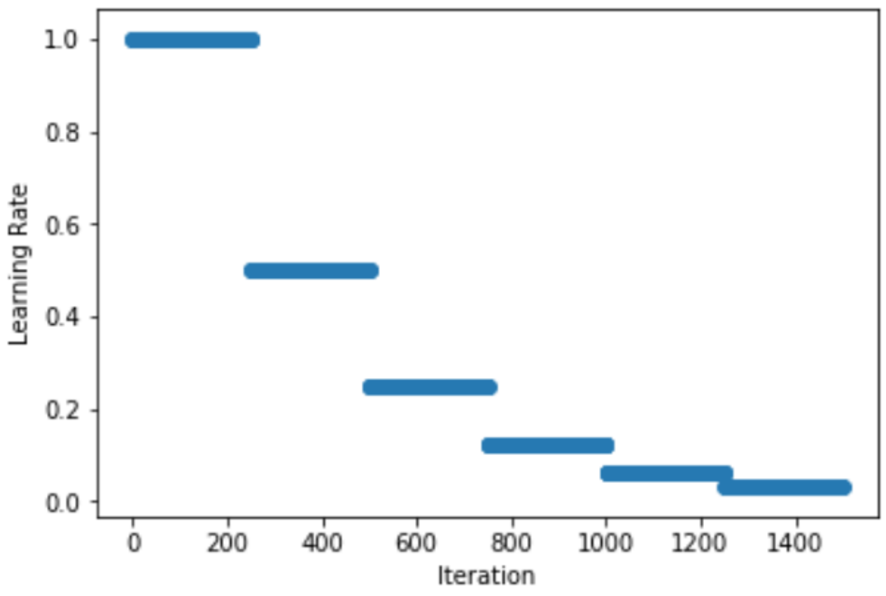
- Traditional approach: decrease learning rate in stages
  - every **k** steps or whenever progress slows down

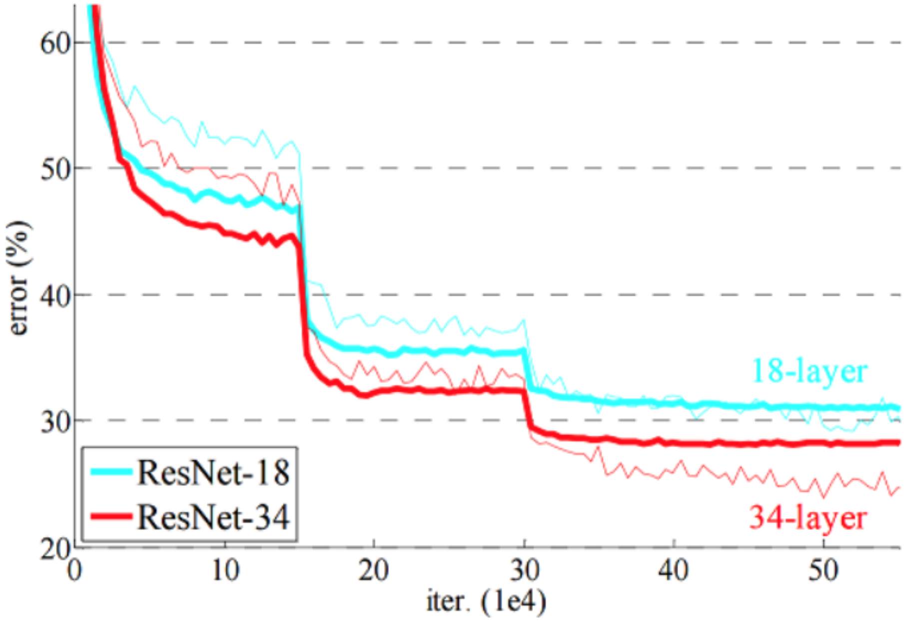



Source: Andrew Ng's Machine Learning course

# Learning Rate Scheduling

- ➢ Traditional approach: decrease learning rate in stages
  - ○ every **k** steps or whenever progress slows down



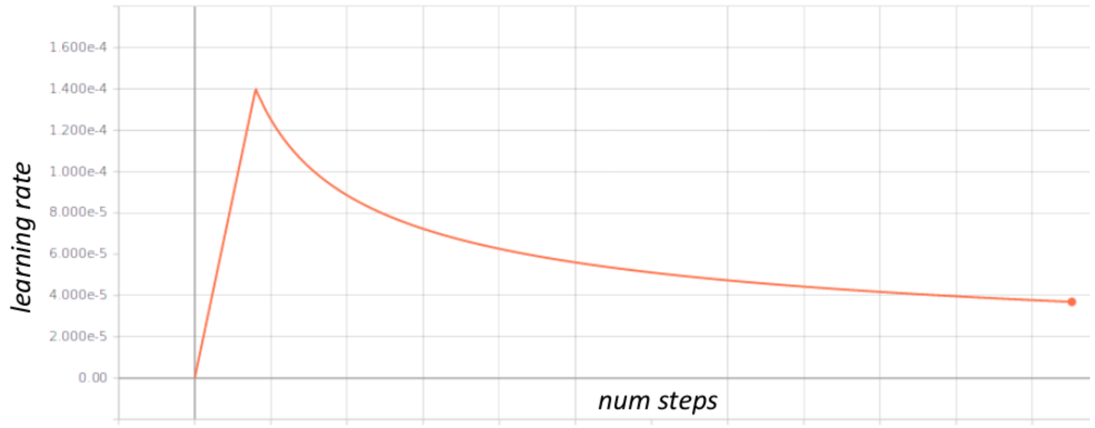

Source: Andrew Ng's Machine Learning course

# Learning Rate Scheduling

- Problem: first k steps of Adam are unstable
  - it needs time to accumulate statistics
- Use warmup time!
  keep lr small over first epochs: $\alpha = \alpha_{base} \cdot min(growth(t), decay(t))$

$$growth(t) = \frac{t}{T_{warmup}}$$

$$decay(t) = \sqrt{\frac{T_{warmup}}{t}}$$

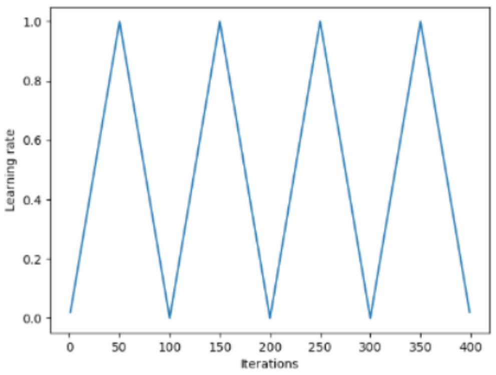



Source: Vaswani et al., Attention is All You Need

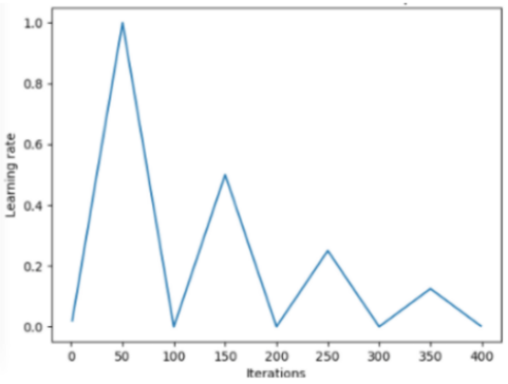
# Learning Rate Scheduling

➢ Why stop at one cycle?

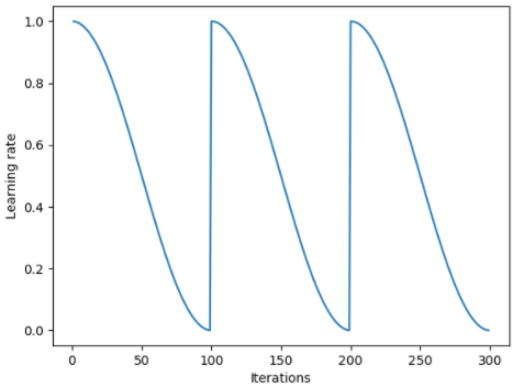
linear cyclic learning rate



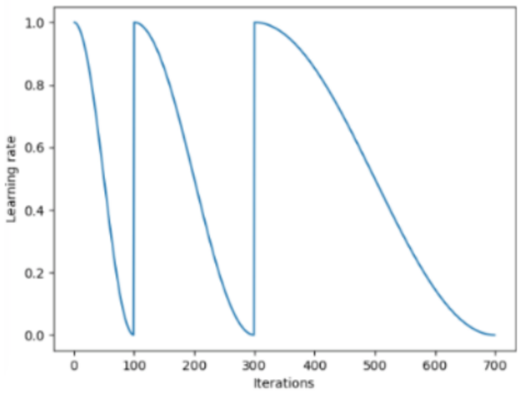

decreasing cyclic learning rate




cosine cyclic learning rate



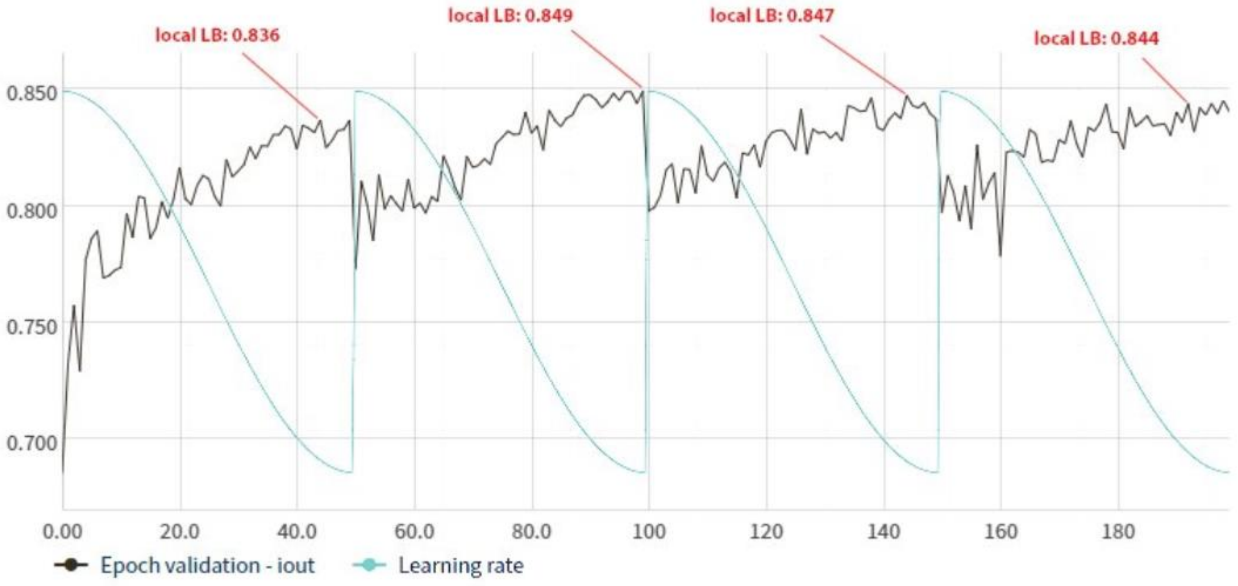

slowing cosine learning rate




Source: Smith et al., Cyclical Learning Rates for Training Neural Networks

# Learning Rate Scheduling

➢ Trains faster. Costs 1 mana per cycle. Use at your own risk




Source: Pestipeti @ kaggle

# That's all for today!

# That's all for today!

## In the next episode:

- Memory Networks?

# That's all for today!

## In the next episode:

- Memory Networks?

+ Really cool surprise seminar!