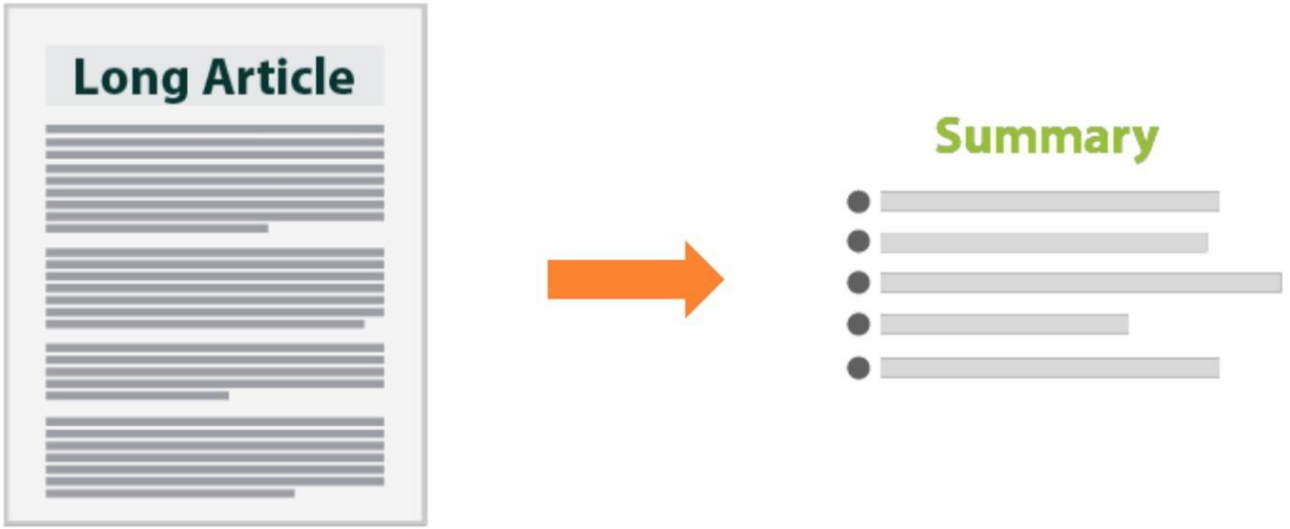
These slides are almost the exact copy of YSDA NLP course. Special thanks to YSDA team.

Summarization

Text Summarization

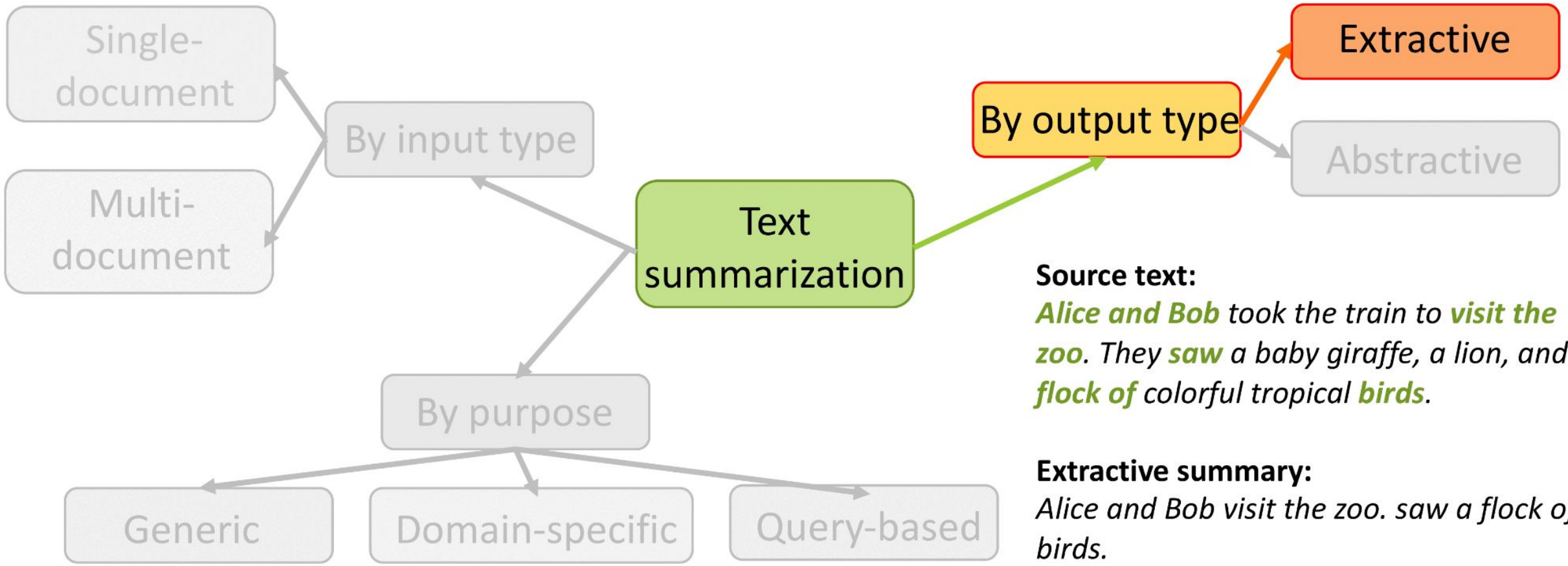
- To condense a piece of text to a shorter version while maintaining the important points



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

Text Summarization Methods

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.

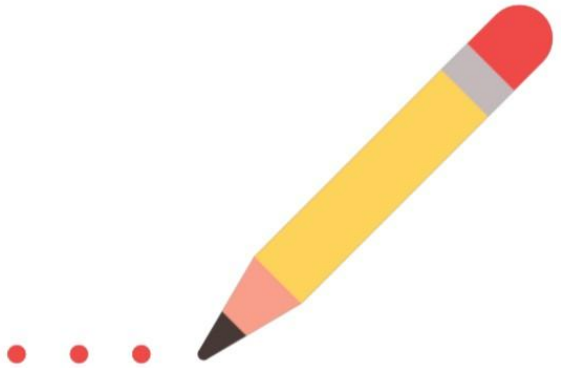
Extractive summary:
Alice and Bob visit the zoo. saw a flock of birds.

Extractive vs Abstractive

Extractive Summarization

select sentences from the article

Abstractive Summarization



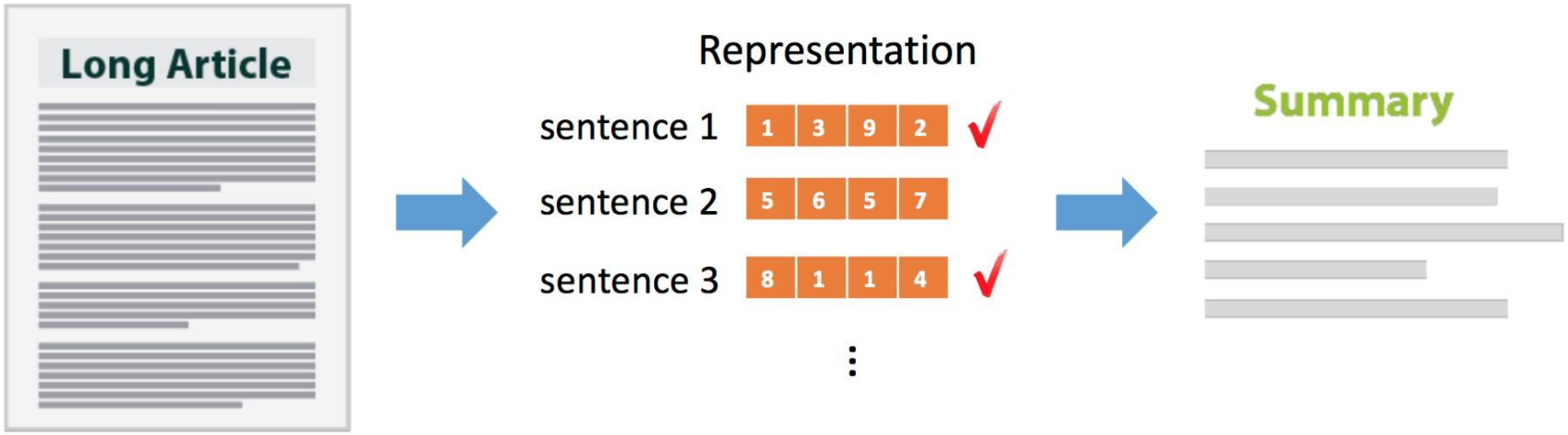
generate the summary word-by-word

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

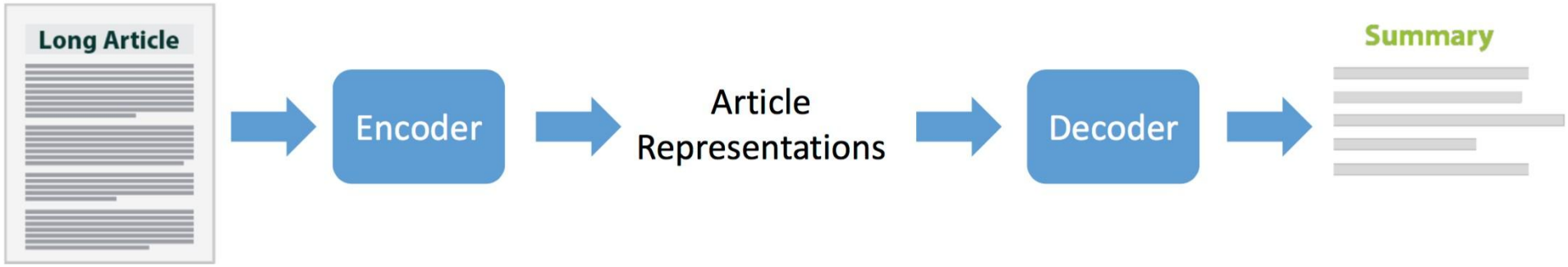


From http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

Extractive vs Abstractive

Abstractive Summarization

- Select phrases or sentences from the source document

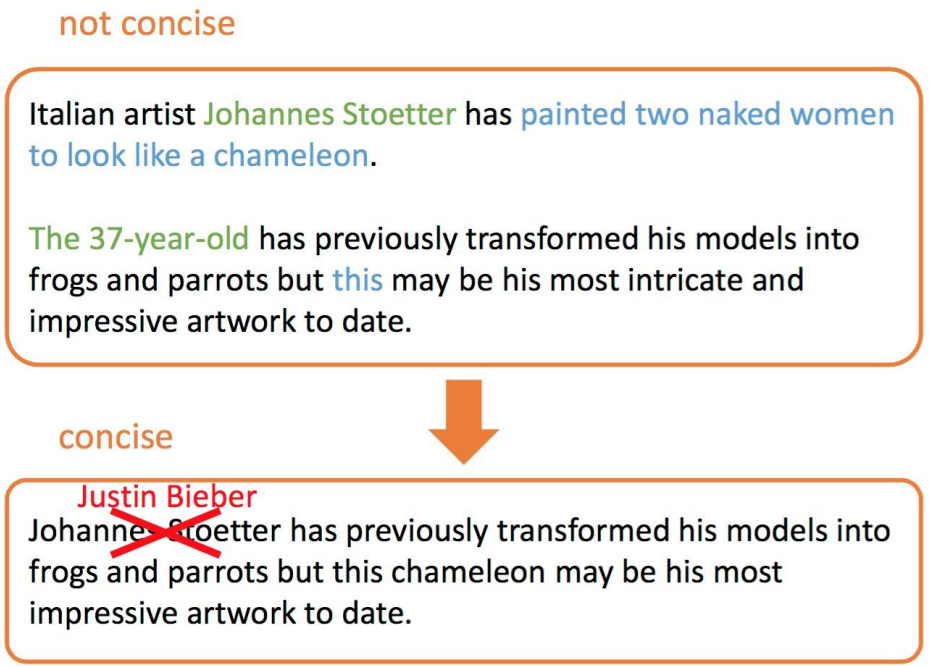


From http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

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



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

Evaluation

- Human evaluation
- ROUGE

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

- **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

Extractive Text Summarization

SummaRuNNer

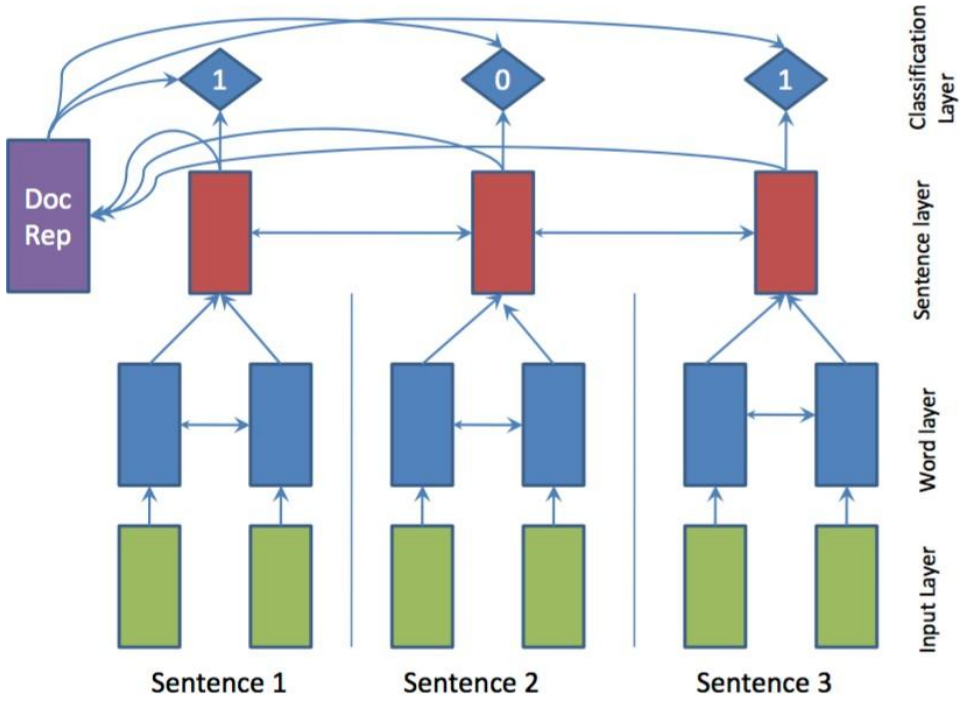
$$
\begin{aligned}
P(y_j = 1|\mathbf{h}_j, \mathbf{s}_j, \mathbf{d}) = \sigma(&W_c \mathbf{h}_j && \#(\text{content}) \\
&+ \mathbf{h}_j^T W_s \mathbf{d} && \#(\text{salience}) \\
&- \mathbf{h}_j^T W_r \tanh(\mathbf{s_j}) && \#(\text{novelty}) \\
&+ W_{ap} \mathbf{p}_j^a && \#(\text{abs. pos. imp.})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

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

-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

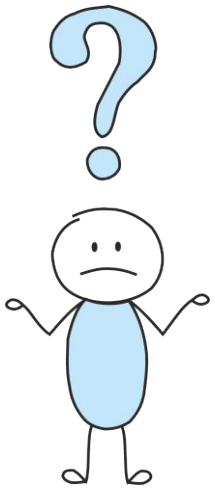


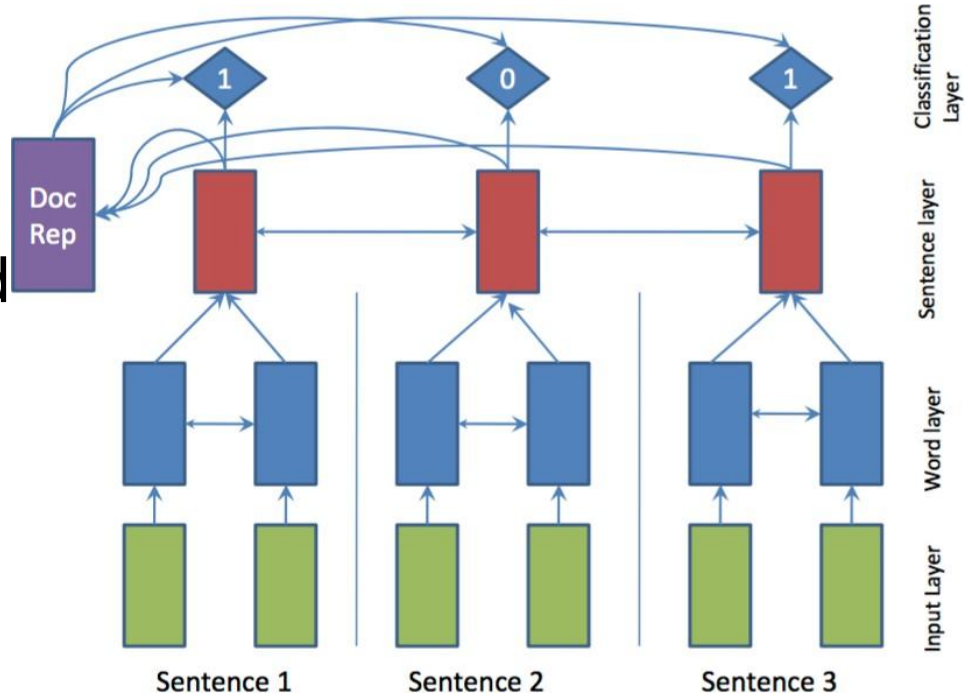
Nallapati et al, AAAI 2017, https://arxiv.org/abs/1611.04230

SummaRuNNer

Extractive training

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





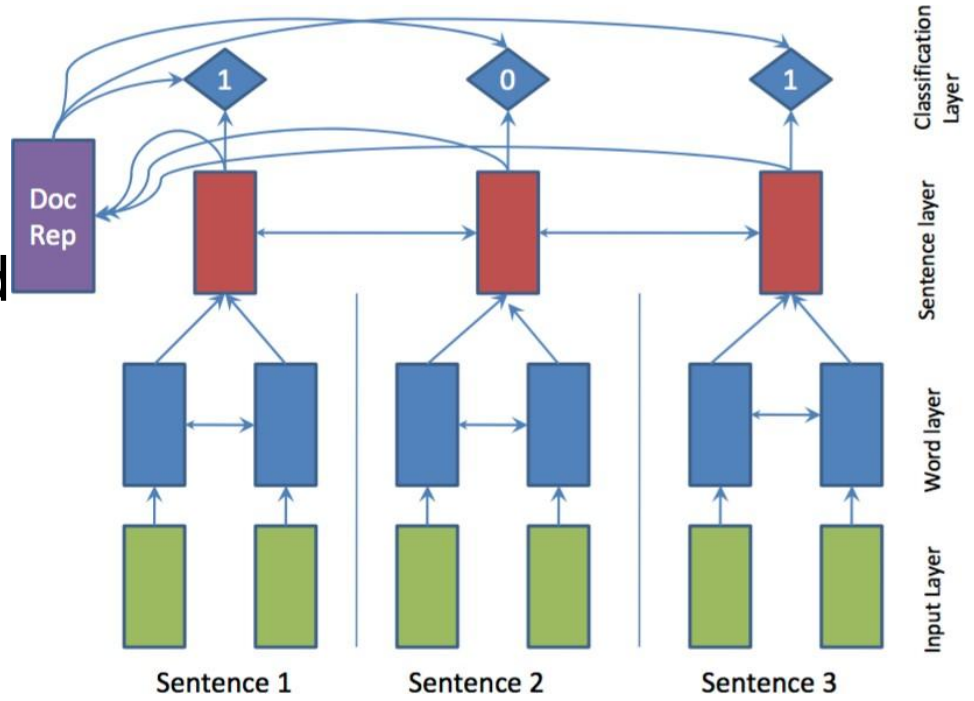
Nallapati et al, AAAI 2017, https://arxiv.org/abs/1611.04230

SummaRuNNer

Extractive training

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

Greedily add sentences with the highest ROUGE



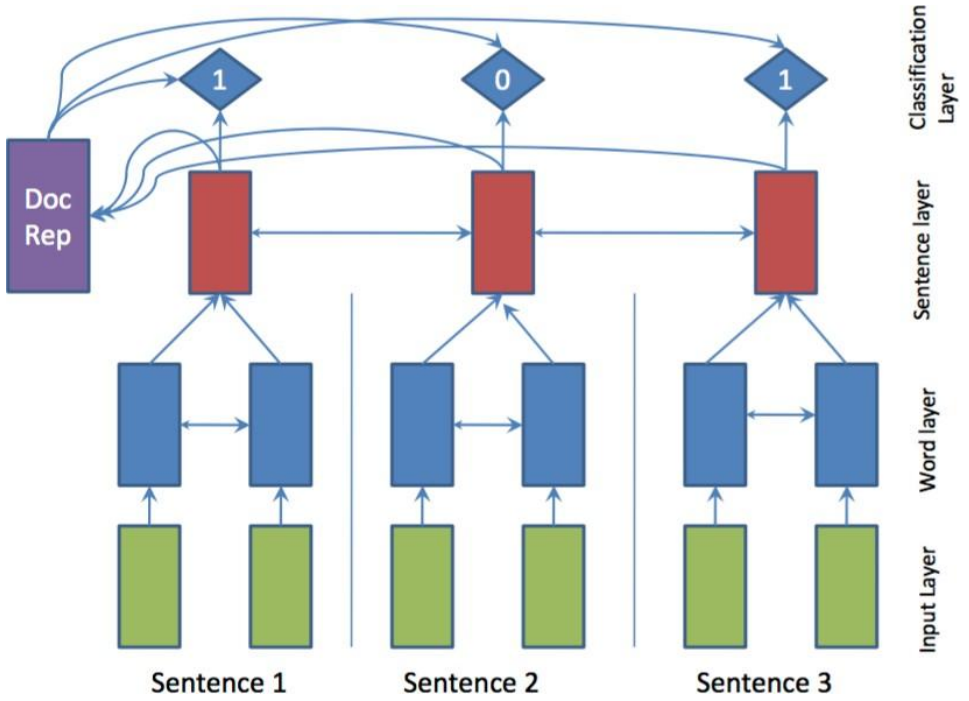
Nallapati et al, AAAI 2017, https://arxiv.org/abs/1611.04230

SummaRuNNer

Extractive training

Loss function:

$$
\begin{aligned}
l(\mathbf{W}, \mathbf{b}) &= -\sum_{d=1}^{N}\sum_{j=1}^{N_d}(y_j^d \log P(y_j^d = 1|\mathbf{h}_j^d, \mathbf{s}_j^d, \mathbf{d}_d) \\
&+ (1 - y_j^d) \log(1 - P(y_j^d = 1|\mathbf{h}_j^d, \mathbf{s}_j^d, \mathbf{d}_d))
\end{aligned}
$$

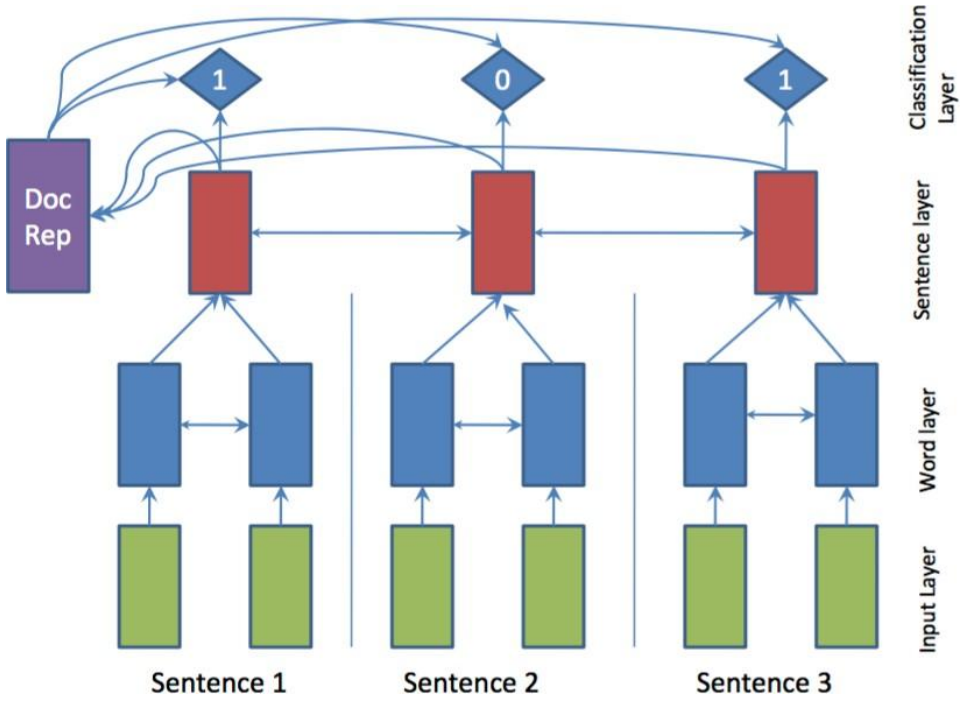


Nallapati et al, AAAI 2017, https://arxiv.org/abs/1611.04230

SummaRuNNer

Abstractive training

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



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

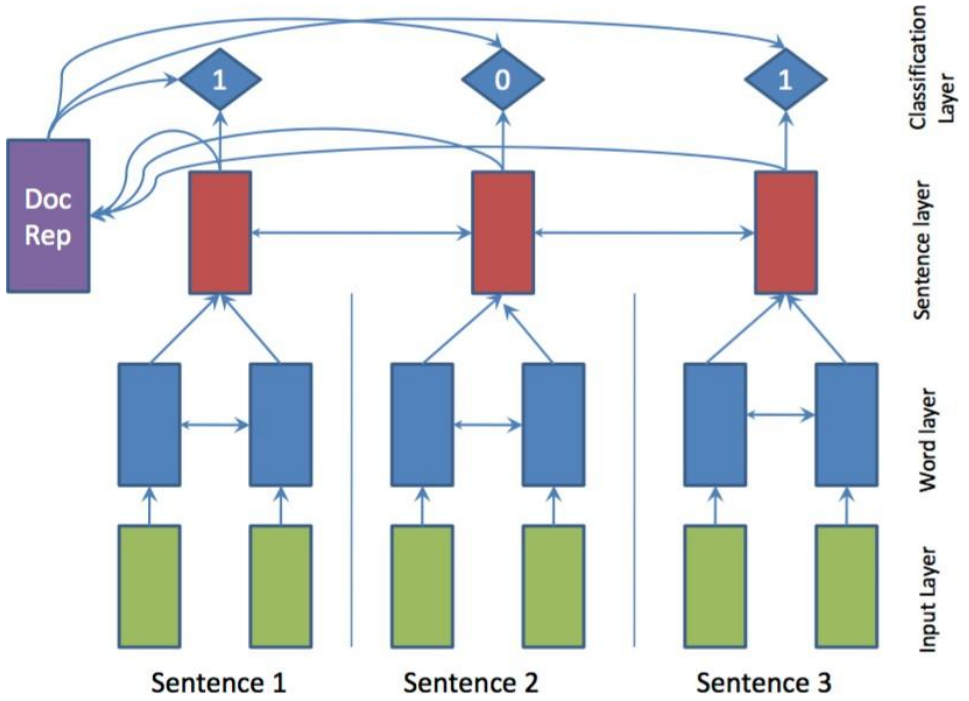
SummaRuNNer

Abstractive training

Loss function:

$$l(\mathbf{W}, \mathbf{b}, \mathbf{W}', \mathbf{b}') = -\sum_{k=1}^{N_s} \log(\mathbf{P_v}(w_k))$$

$$\begin{aligned} \mathbf{f}_k &= \tanh(\mathbf{W'}_{fh}\mathbf{h}_k + \mathbf{W'}_{fx}\mathbf{x}_k + \mathbf{W'}_{fc}\mathbf{s}_{-1} + \mathbf{b'}_f) \\ \mathbf{P_v}(\mathbf{w})_k &= \text{softmax}(\mathbf{W'}_v\mathbf{f}_k + \mathbf{b'}_v) \end{aligned}$$



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

SummaRuNNer

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

	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

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

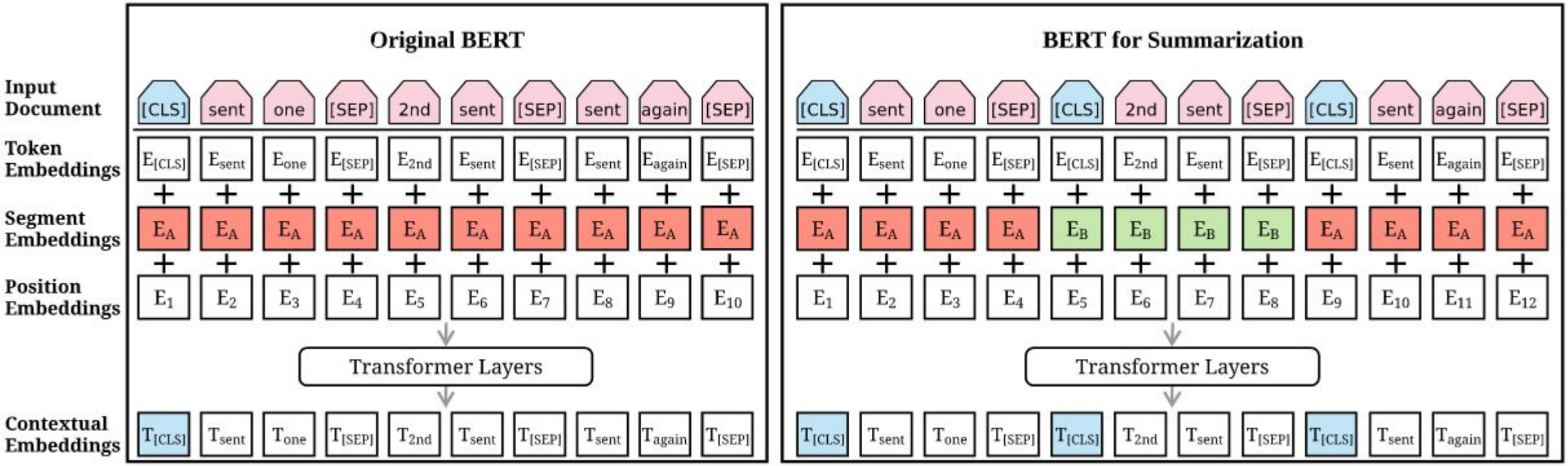
SummaRuNNer

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

BertSum



$$[sent_1, sent_2, sent_3, sent_4, sent_5] \longrightarrow [E_A, E_B, E_A, E_B, E_A]$$

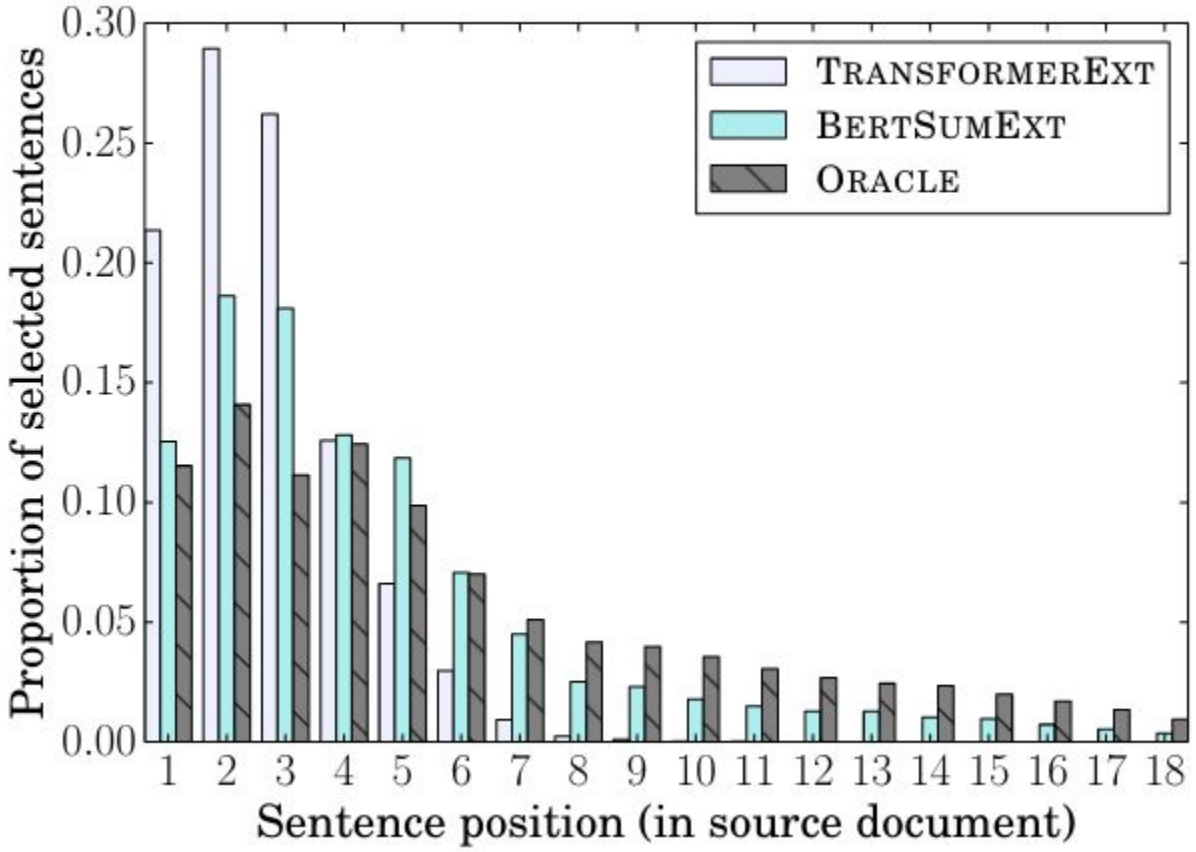
https://arxiv.org/pdf/1908.08345.pdf

BertSum

- sinusoid positional embeddings
- overcome the limitation of a maximum length of 512 by adding more position embeddings
- L= 1, 2, 3 and L= 2 performed best
- $$(\text{warmup} = 10,000)$$
  $$lr = 2e^{-3} \cdot \min(\text{step}^{-0.5}, \text{step} \cdot \text{warmup}^{-1.5})$$

https://arxiv.org/pdf/1908.08345.pdf

BertSum



https://arxiv.org/pdf/1908.08345.pdf

BertSum

Model	R1	R2	RL
ORACLE	52.59	31.24	48.87
LEAD-3	40.42	17.62	36.67
Extractive			
SUMMARUNNER (Nallapati et al., 2017)	39.60	16.20	35.30
REFRESH (Narayan et al., 2018b)	40.00	18.20	36.60
BERT-based			
BERTSUMEXT	43.25	20.24	39.63
BERTSUMEXT w/o interval embeddings	43.20	20.22	39.59
BERTSUMEXT (large)	43.85	20.34	39.90

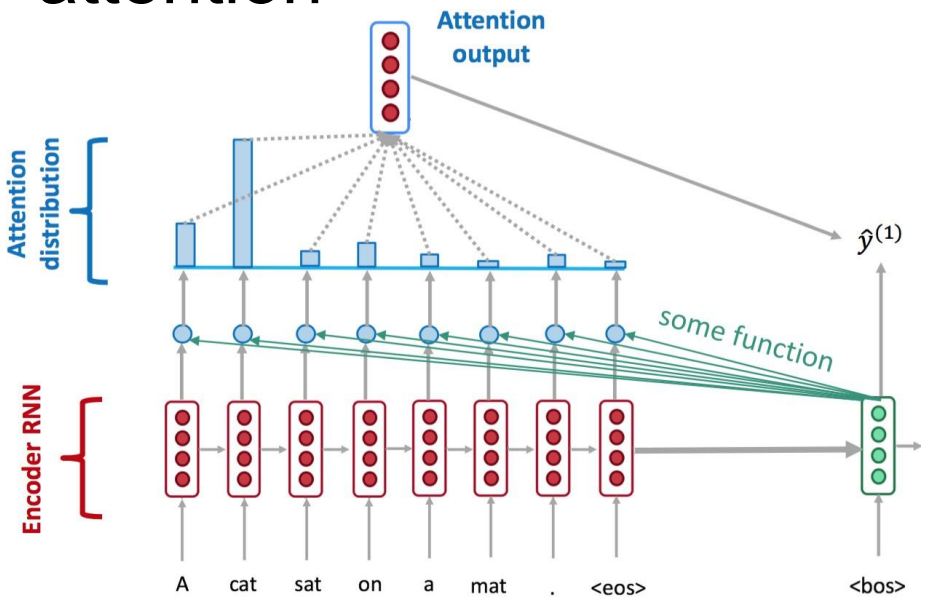
ROUGE F1 results on **CNN/DailyMail**

https://arxiv.org/pdf/1908.08345.pdf

Abstractive Text Summarization

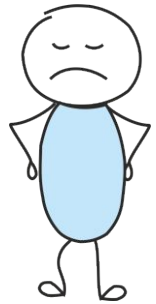
A Neural Attention Model for Sentence Summarization

Model: RNNs with attention



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

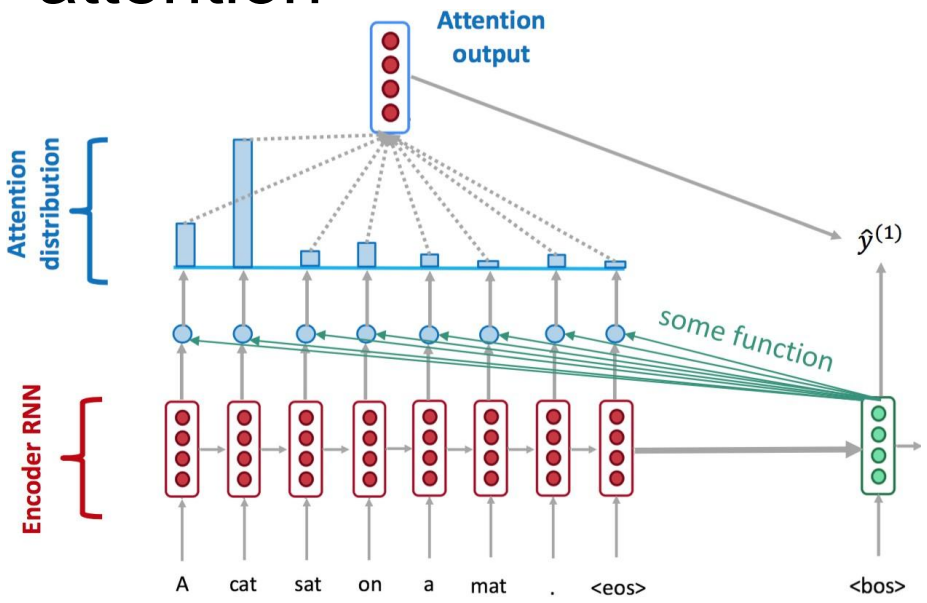
Hard to handle unseen proper noun phrase (named entities)



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

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

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

N-gram match

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

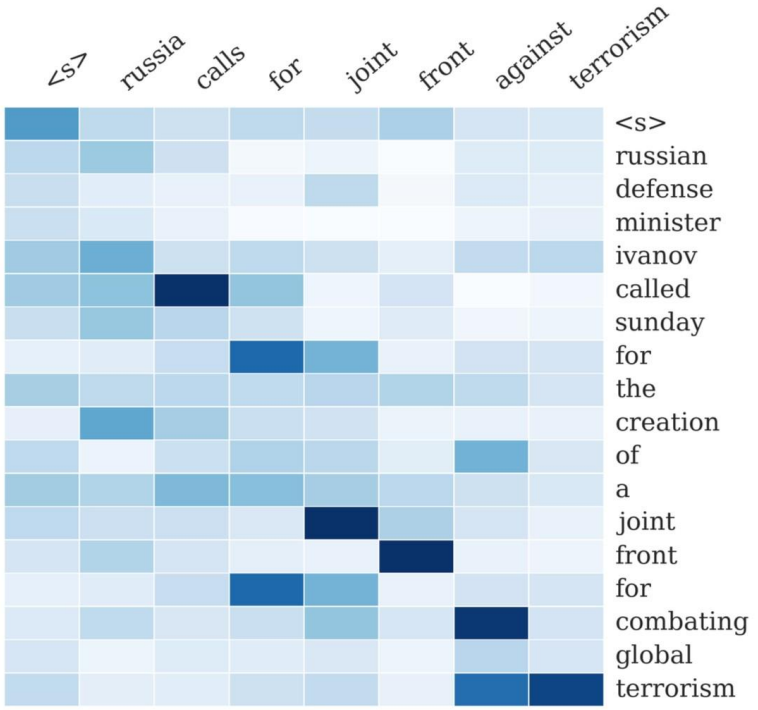


Figure 1: Example output of the attention-based summarization (ABS) system. The heatmap represents a soft alignment between the input (right) and the generated summary (top). The columns represent the distribution over the input after generating each word.

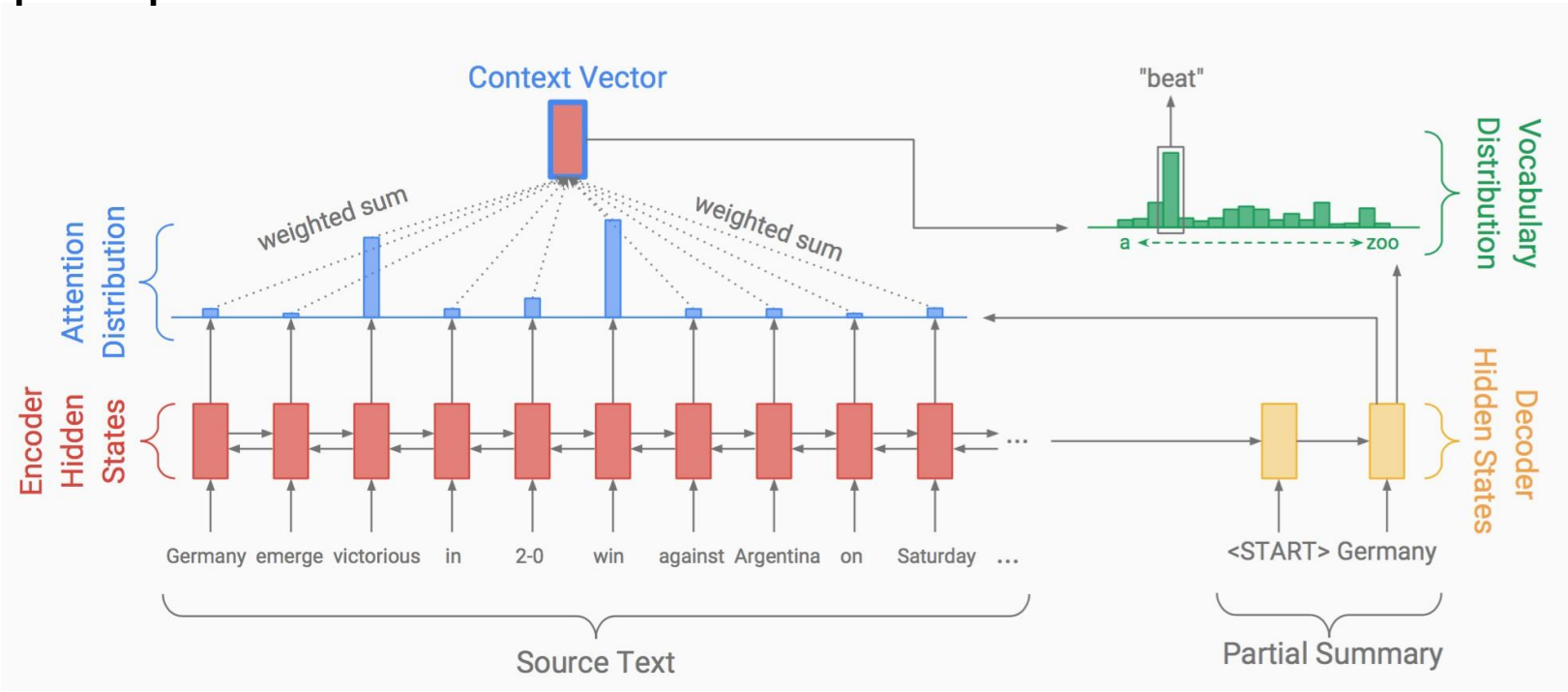
I(8): thousands of kashmiris chanting pro-pakistan slogans on sunday attended a rally to welcome back a hardline separatist leader who underwent cancer treatment in mumbai .
G: thousands attend rally for kashmir hardliner
A: thousands rally in support of hardline kashmiri separatist leader
A+: thousands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

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

Problem 1: The summaries sometimes reproduce factual details inaccurately.

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

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

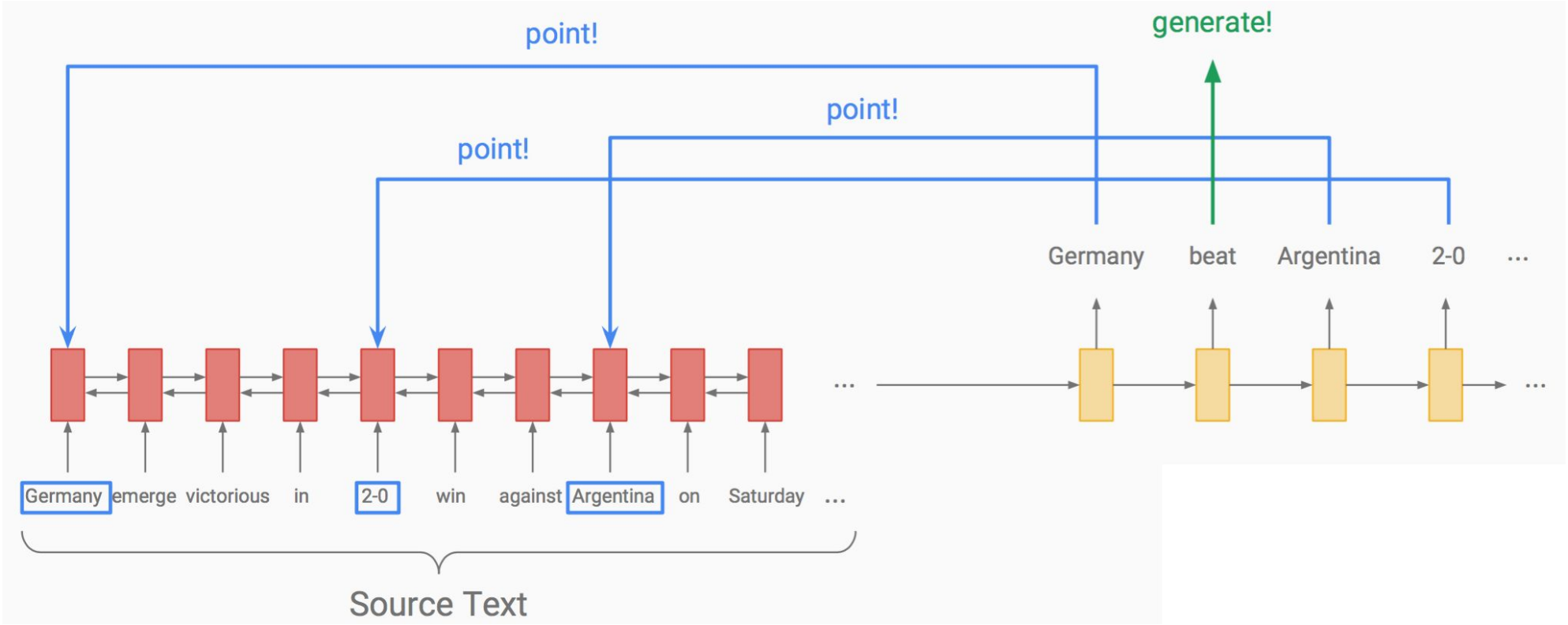
Solution: Use a pointer to copy words.

Problem 2: The summaries sometimes repeat themselves.

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

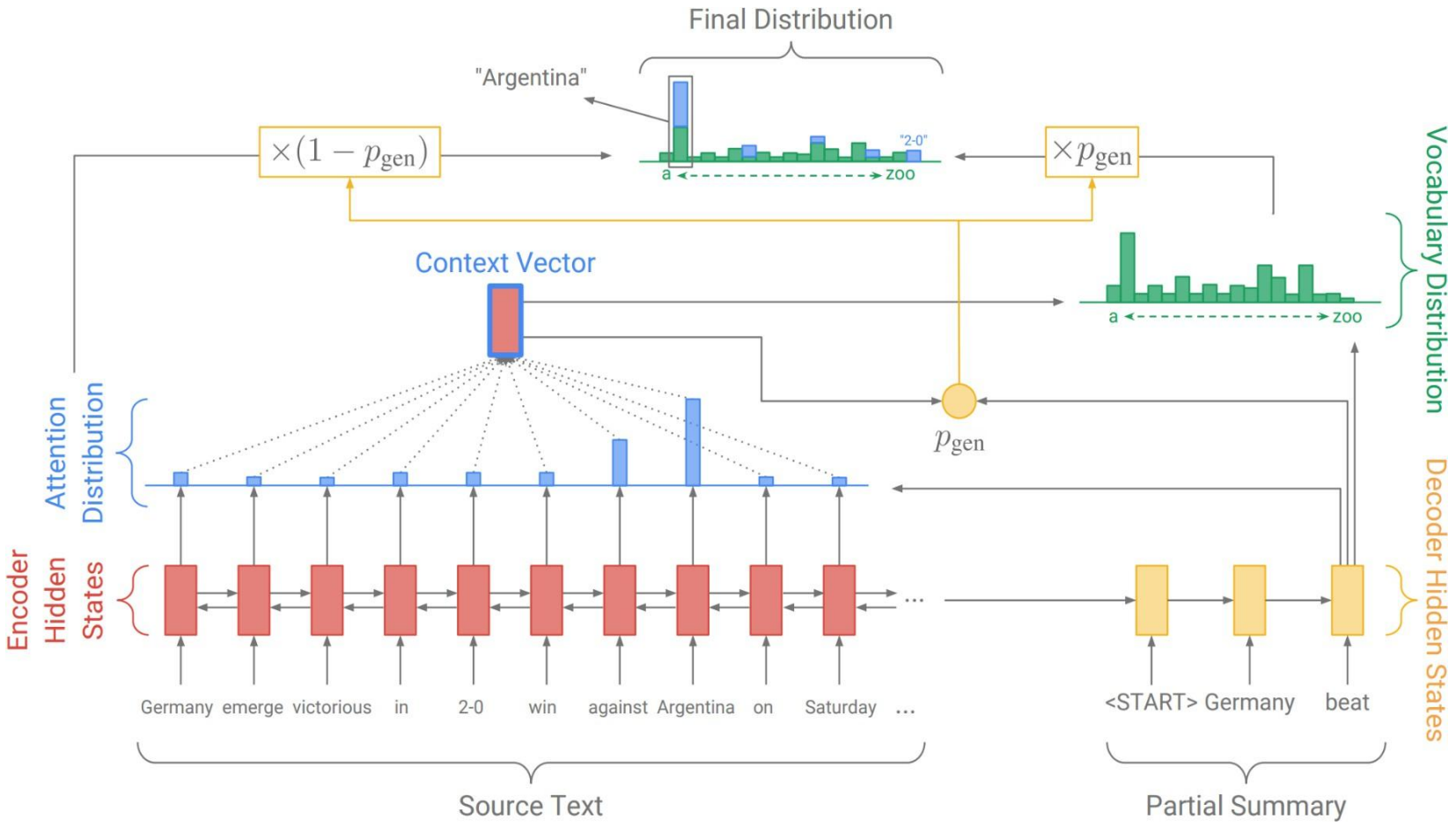
From the authors' slides:
http://anthology.aclweb.org/attachments/P/P17/P17-1099.Presentation.pdf
See et al, ACL 2017, http://aclweb.org/anthology/P17-1099

Get To The Point



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

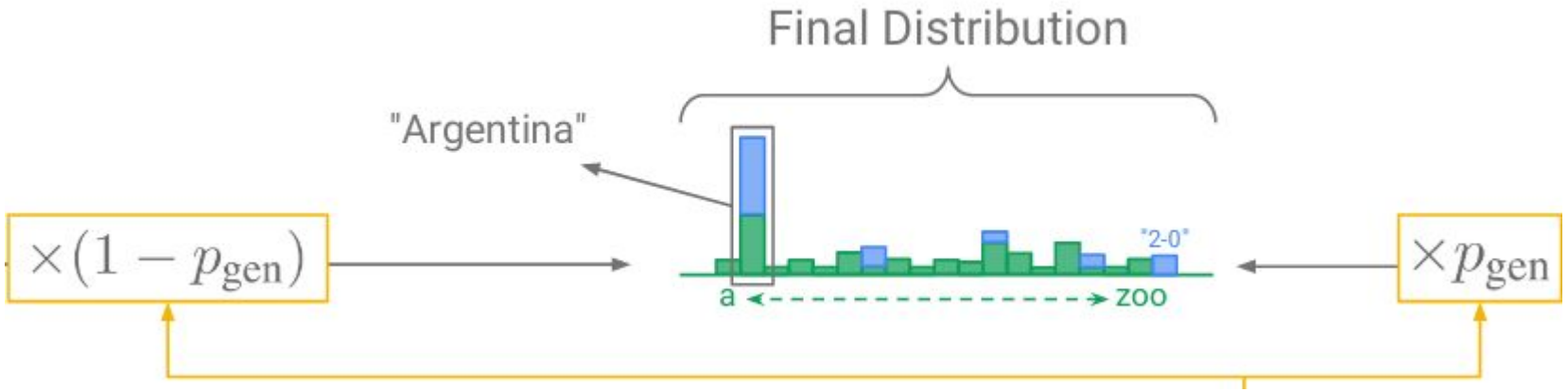


FSee et al, ACL 2017, http://aclweb.org/anthology/P17-1099

Get To The Point

$$h_t^* = \sum_i a_i^t h_i$$

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



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

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

Get To The Point

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

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

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

Coverage = cumulative attention = what has been covered so far

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

Get To The Point

- Change the definition of attention

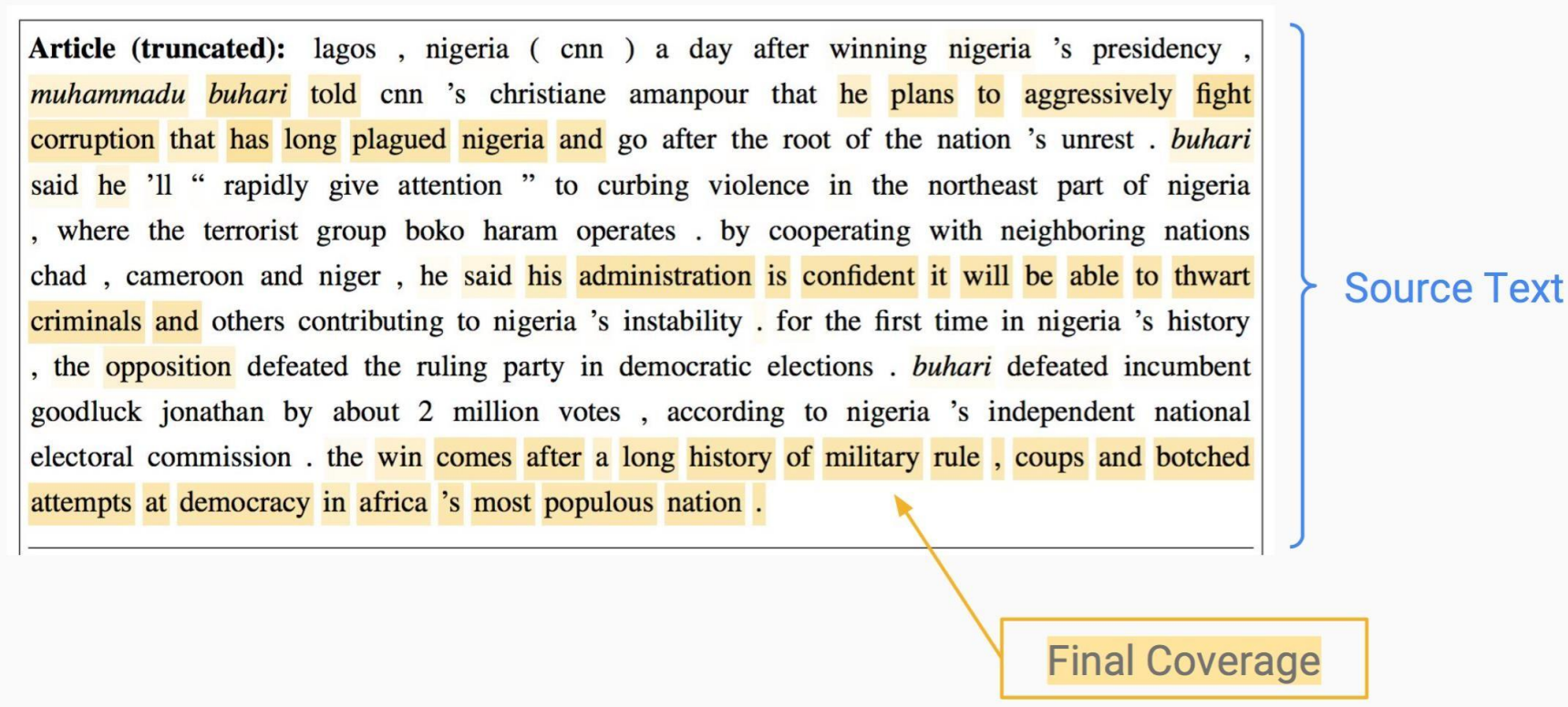
$$e_i^t = v^T \tanh(W_h h_i + W_s s_t + b_{\text{attn}}) \longrightarrow e_i^t = v^T \tanh(W_h h_i + W_s s_t + w_c c_i^t + b_{\text{attn}})$$
$$a^t = \text{softmax}(e^t)$$

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

- Change loss by adding coverage loss

$$\text{covloss}_t = \sum_i \min(a_i^t, c_i^t)$$
$$\text{loss}_t = -\log P(w_t^*) + \lambda \sum_i \min(a_i^t, c_i^t)$$

Get To The Point



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

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

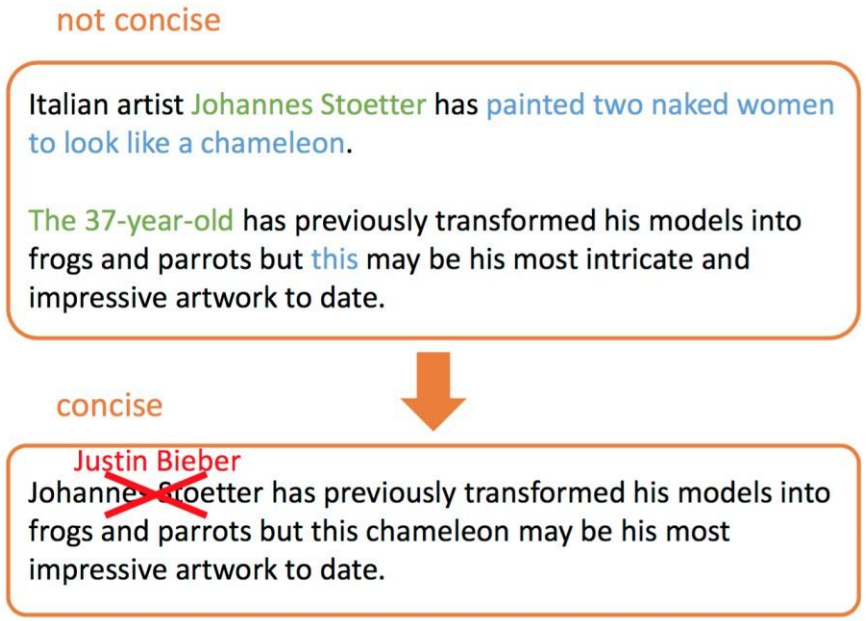
Our improvements

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

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

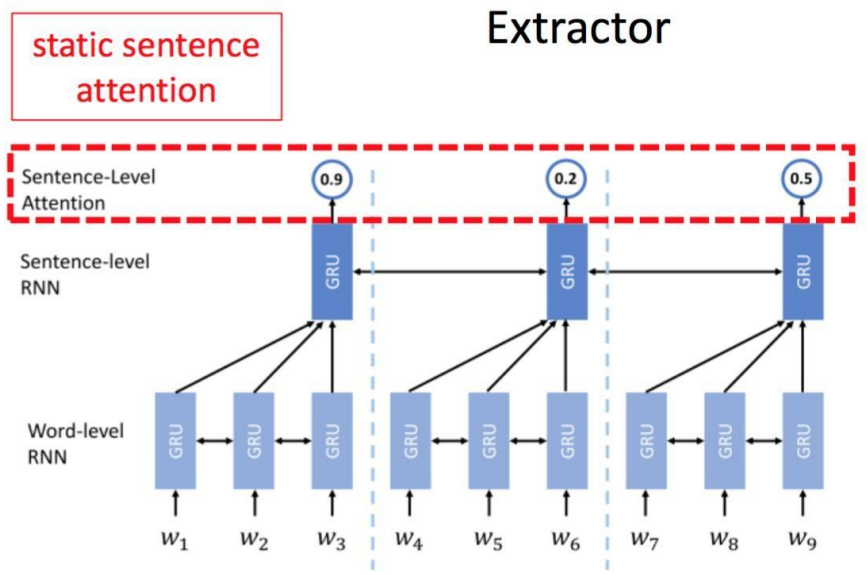


From the authors' slides: http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

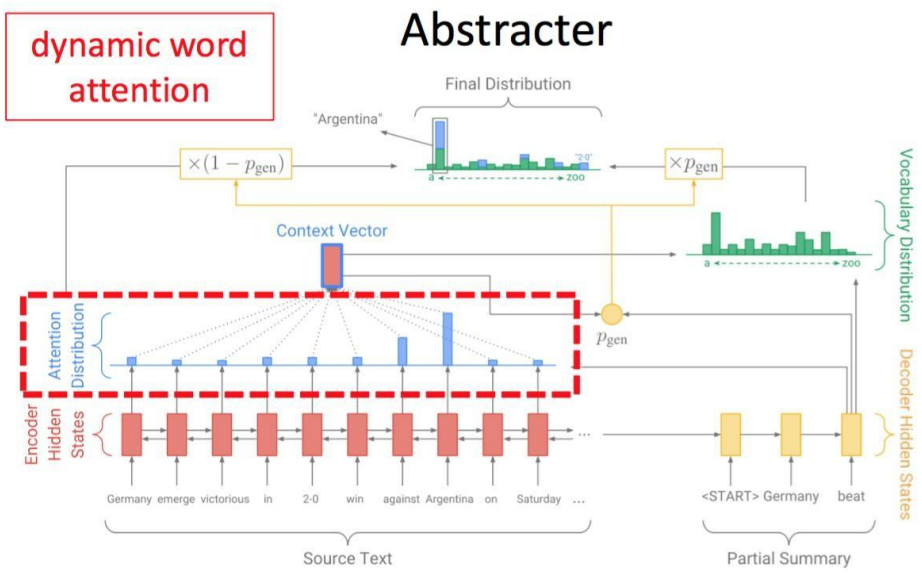
Hsu et al, ACL 2018, http://aclweb.org/anthology/P18-1013

A Unified Model for Extractive and Abstractive Summarization

Models



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



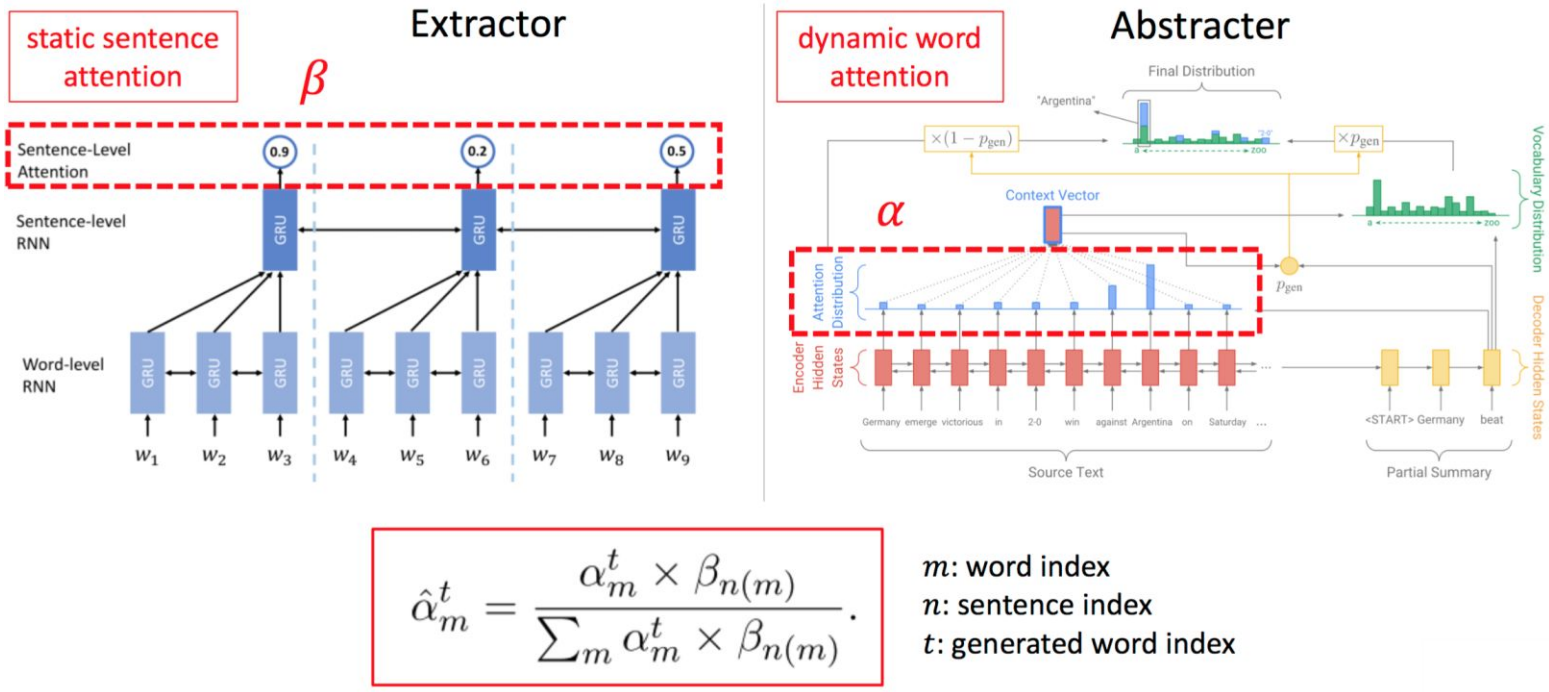
Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. ACL 2017

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

Combined Attention



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

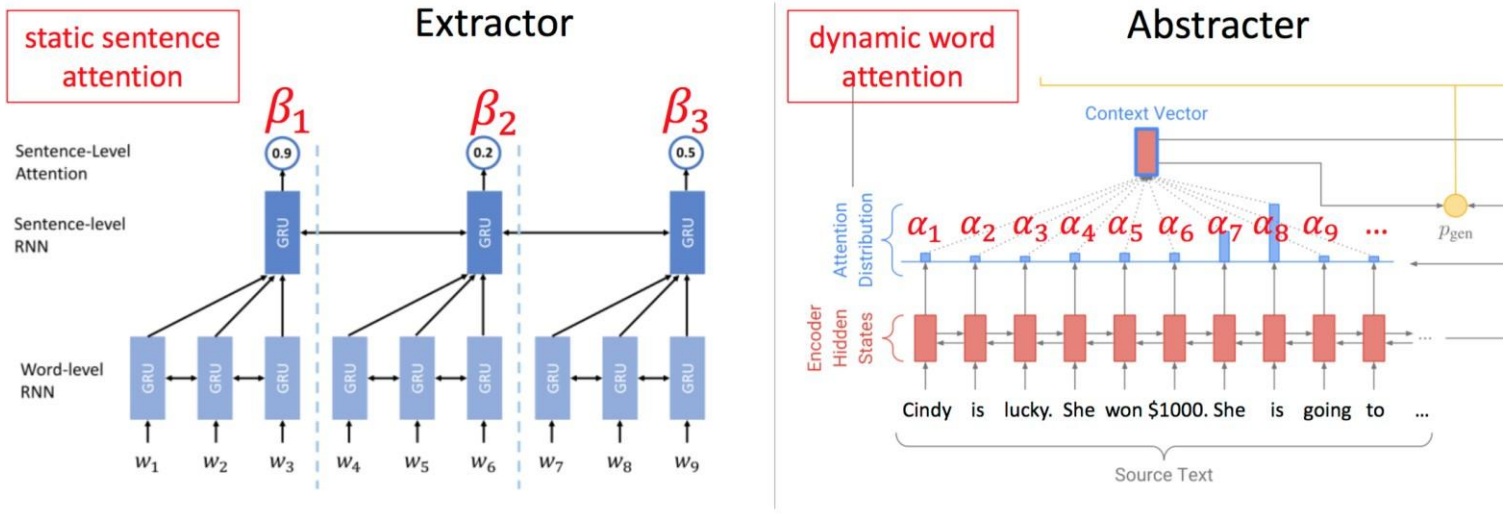
m: word index
n: sentence index
t: generated word index

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

Combined Attention



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

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

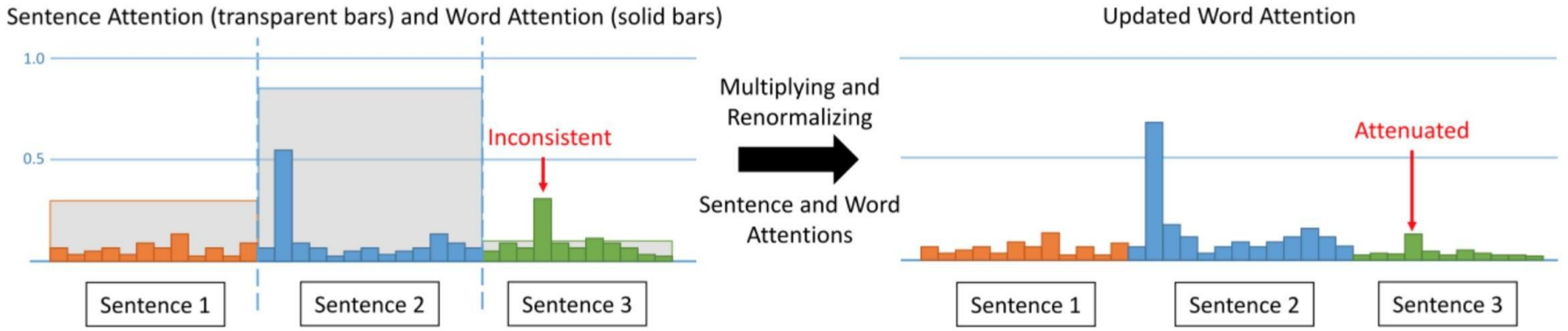
Hsu et al, ACL 2018, http://aclweb.org/anthology/P18-1013

A Unified Model for Extractive and Abstractive Summarization

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

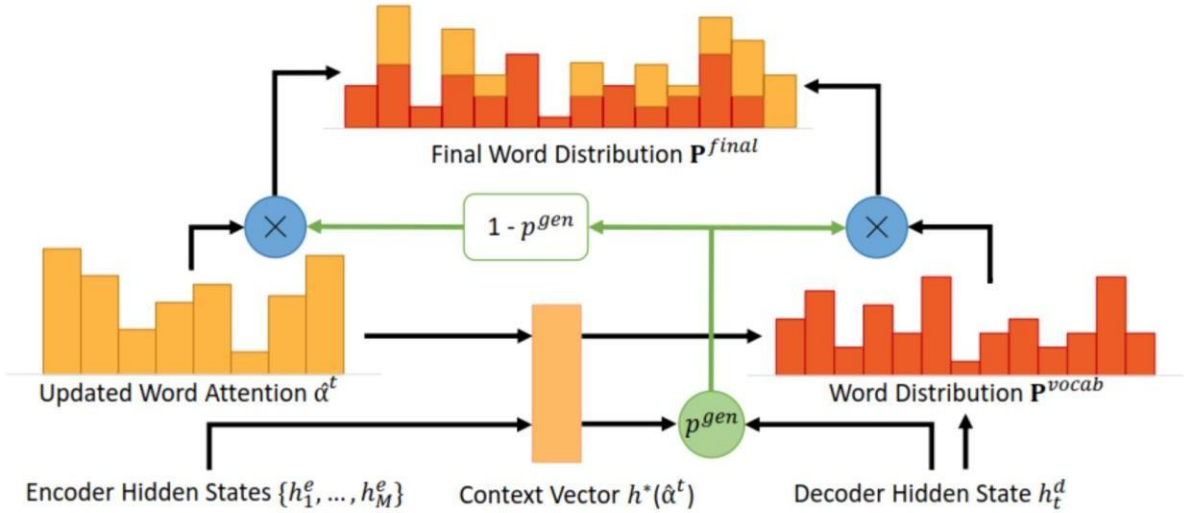
Hsu et al, ACL 2018, http://aclweb.org/anthology/P18-1013

A Unified Model for Extractive and Abstractive Summarization

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

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

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

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

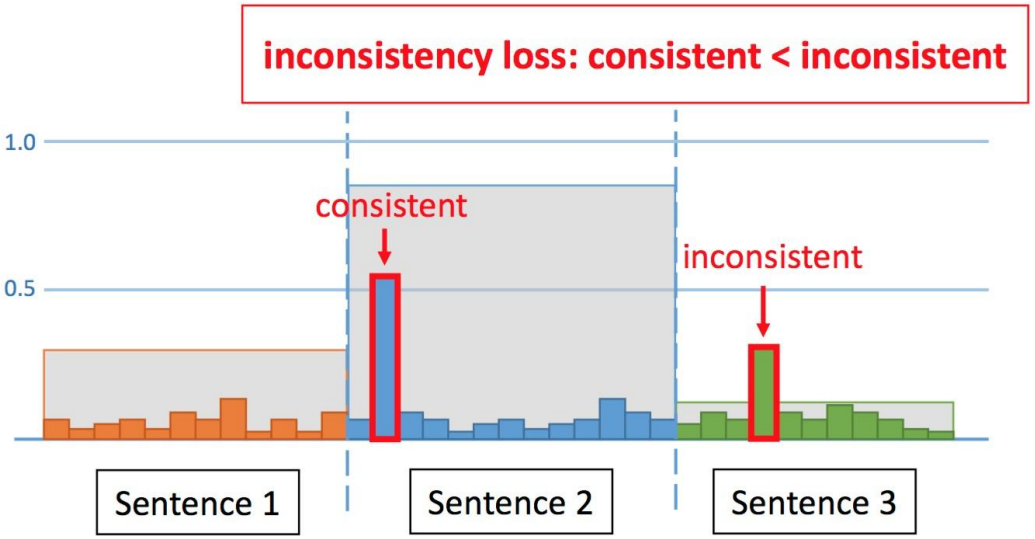
Inconsistency Loss

Encourage Consistency

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

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

K = 2



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

- 3 types of loss functions:

  1. extractor loss

  2. abstracter loss
     + coverage loss

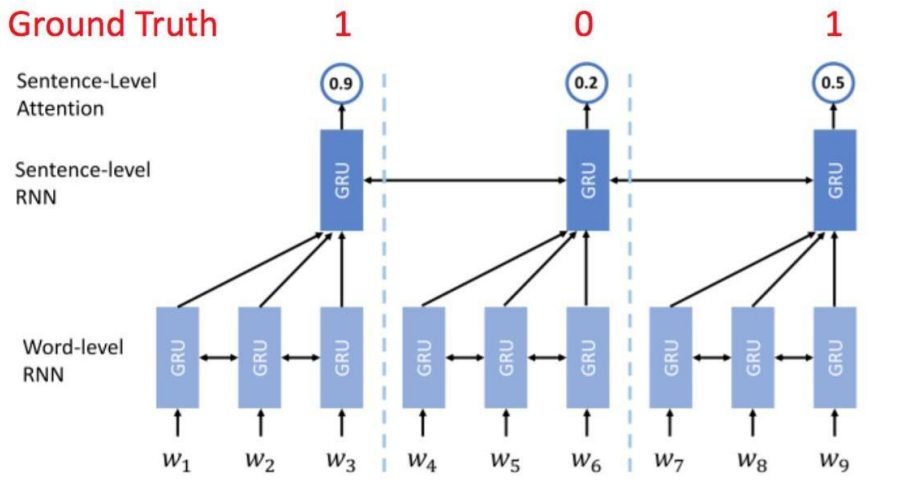
  3. inconsistency loss

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

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



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

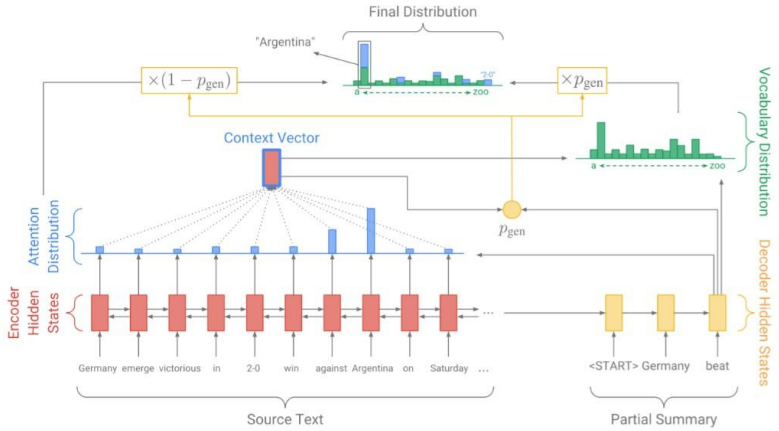
where $g_n \in \{0, 1\}$ is the ground-truth label for the n^{th} sentence and N is the number of sentences.

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

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



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

$$L_{cov} = \frac{1}{T}\sum_{t=1}^{T}\sum_{m=1}^{M} \min(\hat{\alpha}^t_m, c^t_m) \qquad \mathbf{c}^t = \sum_{t'=1}^{t-1} \hat{\boldsymbol{\alpha}}^{t'}$$

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

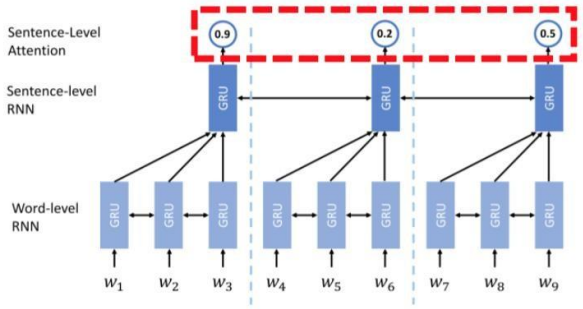
A Unified Model for Extractive and Abstractive Summarization

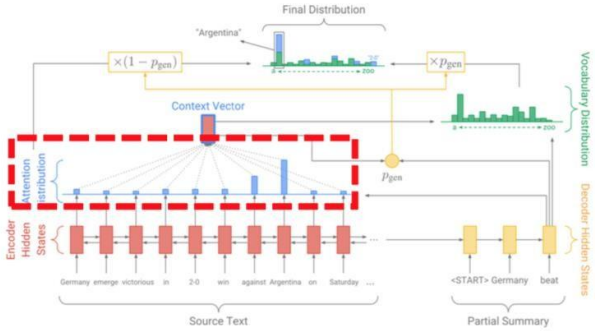
- 3 types of loss functions:

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

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

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

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

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.

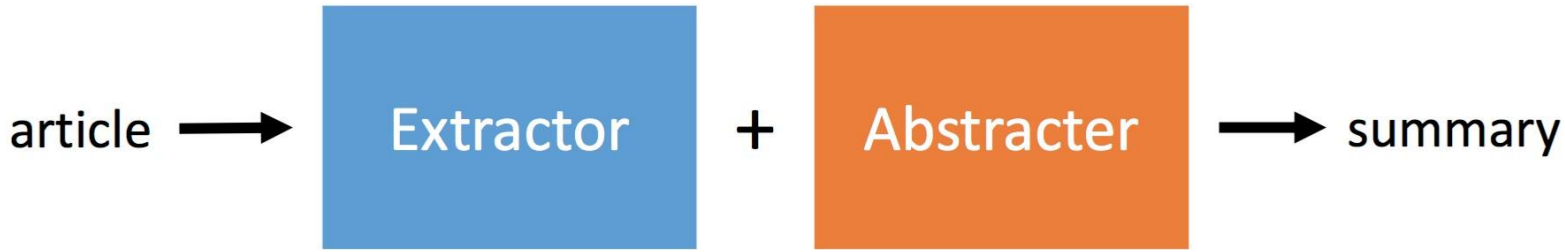
From the authors' slides:
http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

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

A Unified Model for Extractive and Abstractive Summarization

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

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



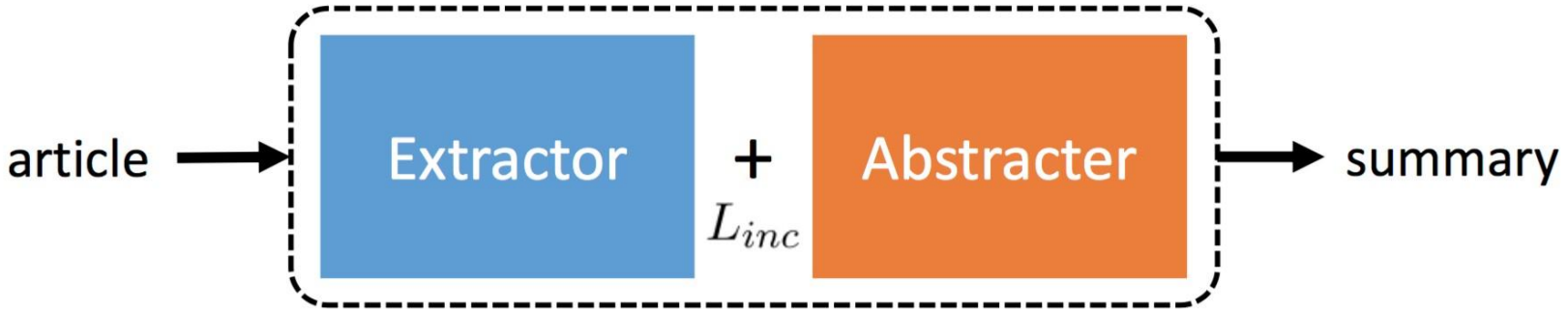
From the authors' slides:
http://anthology.aclweb.org/attachments/P/P18/P18-1013.Presentation.pdf

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

A Unified Model for Extractive and Abstractive Summarization

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:

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



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

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

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

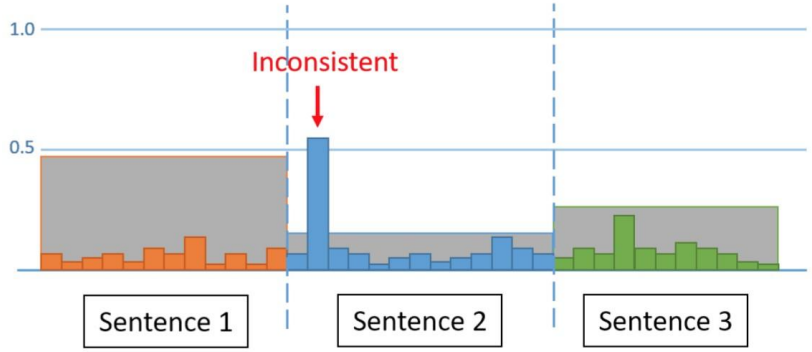
Hsu et al, ACL 2018, http://aclweb.org/anthology/P18-1013

A Unified Model for Extractive and Abstractive Summarization

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.

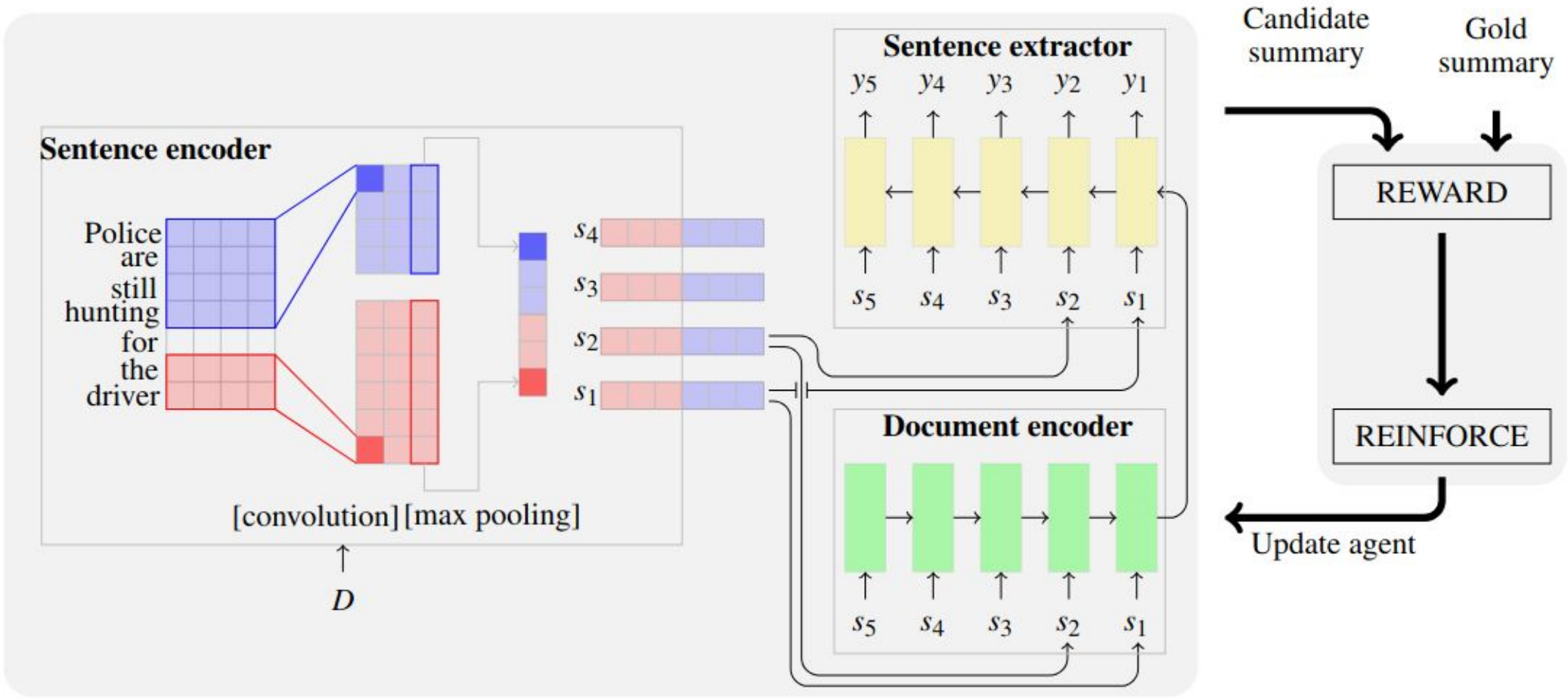
Method	avg. R_{inc}
w/o incon. loss	0.198
w/ incon. loss	0.042

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

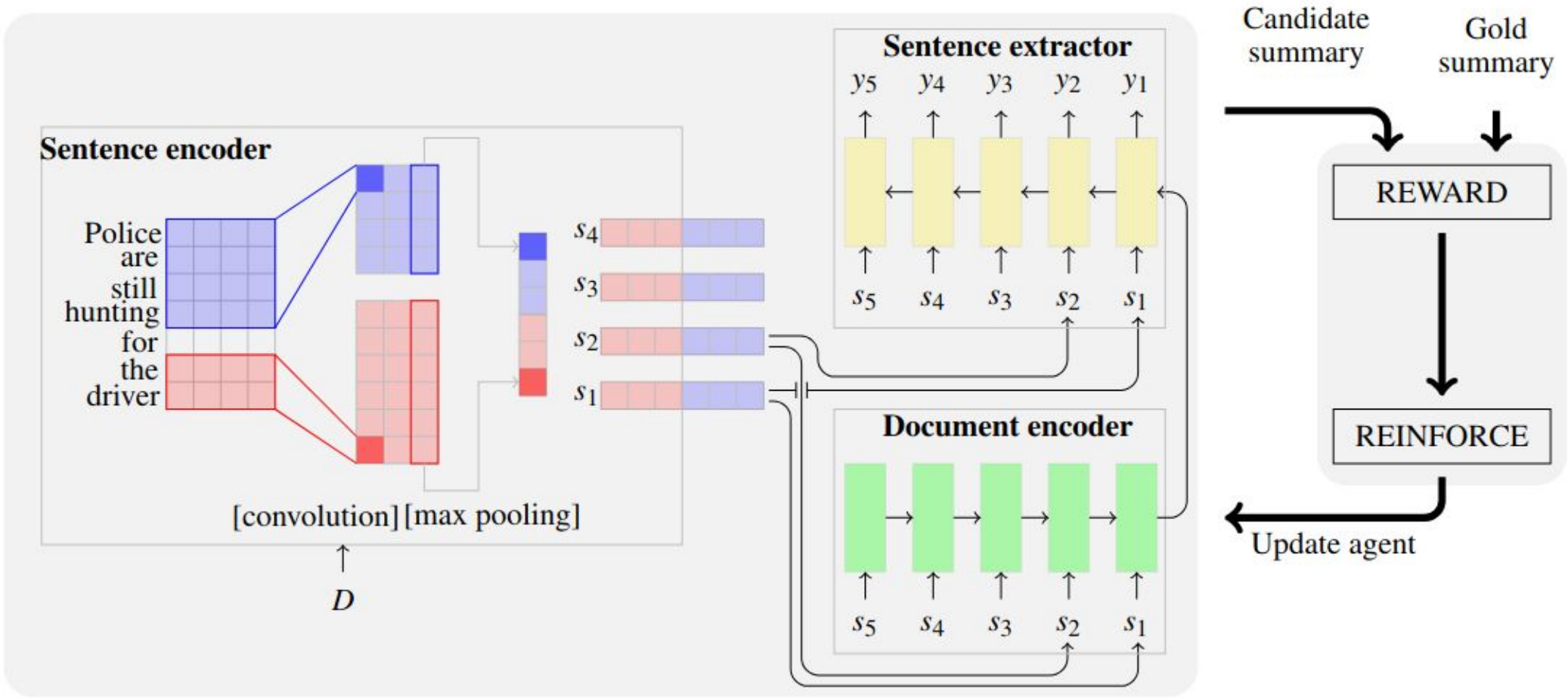
REFRESH



Optimizes the evaluation metric instead of maximizing the likelihood

https://www.aclweb.org/anthology/N18-1158.pdf

REFRESH



$$L(\theta) = -\sum_{i=1}^{n} \log p(y_i|s_i, D, \theta) \longrightarrow L(\theta) = -\mathbb{E}_{\hat{y}\sim\, p_\theta}[r(\hat{y})]$$

https://www.aclweb.org/anthology/N18-1158.pdf