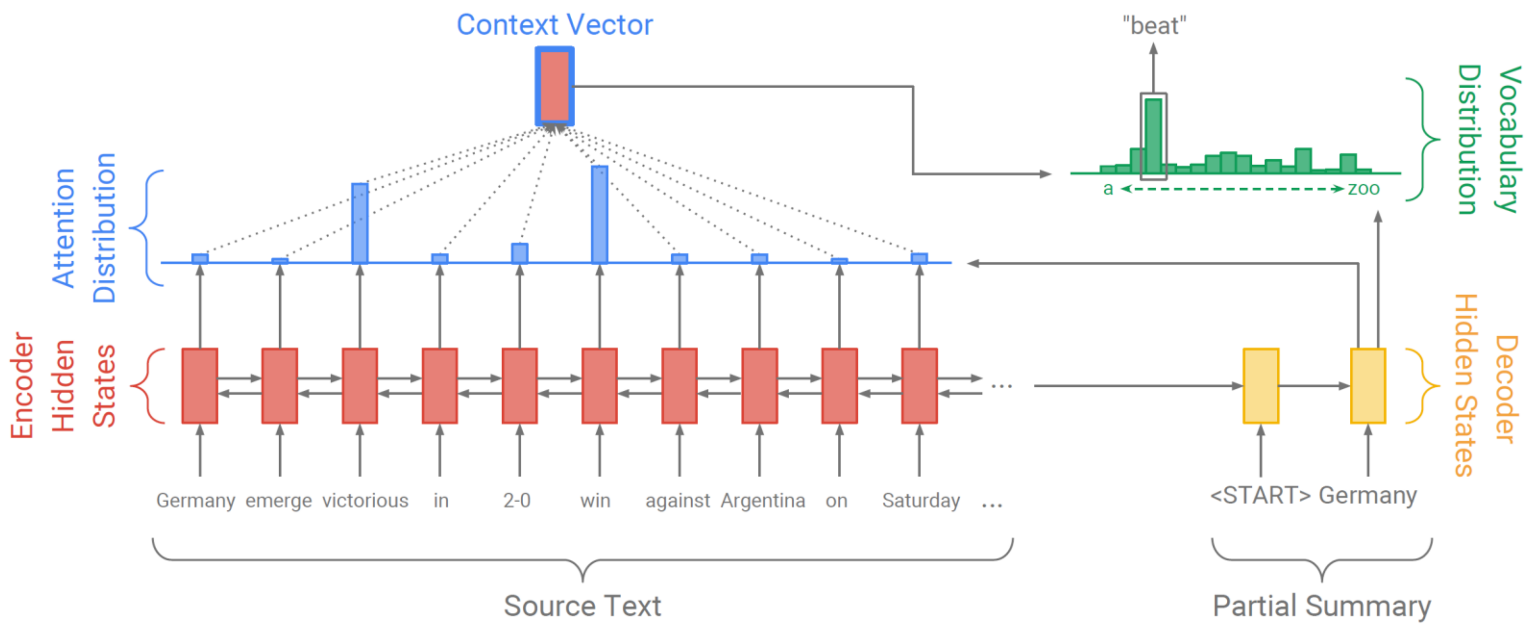


Abstractive Summarization

Pilsung Kang

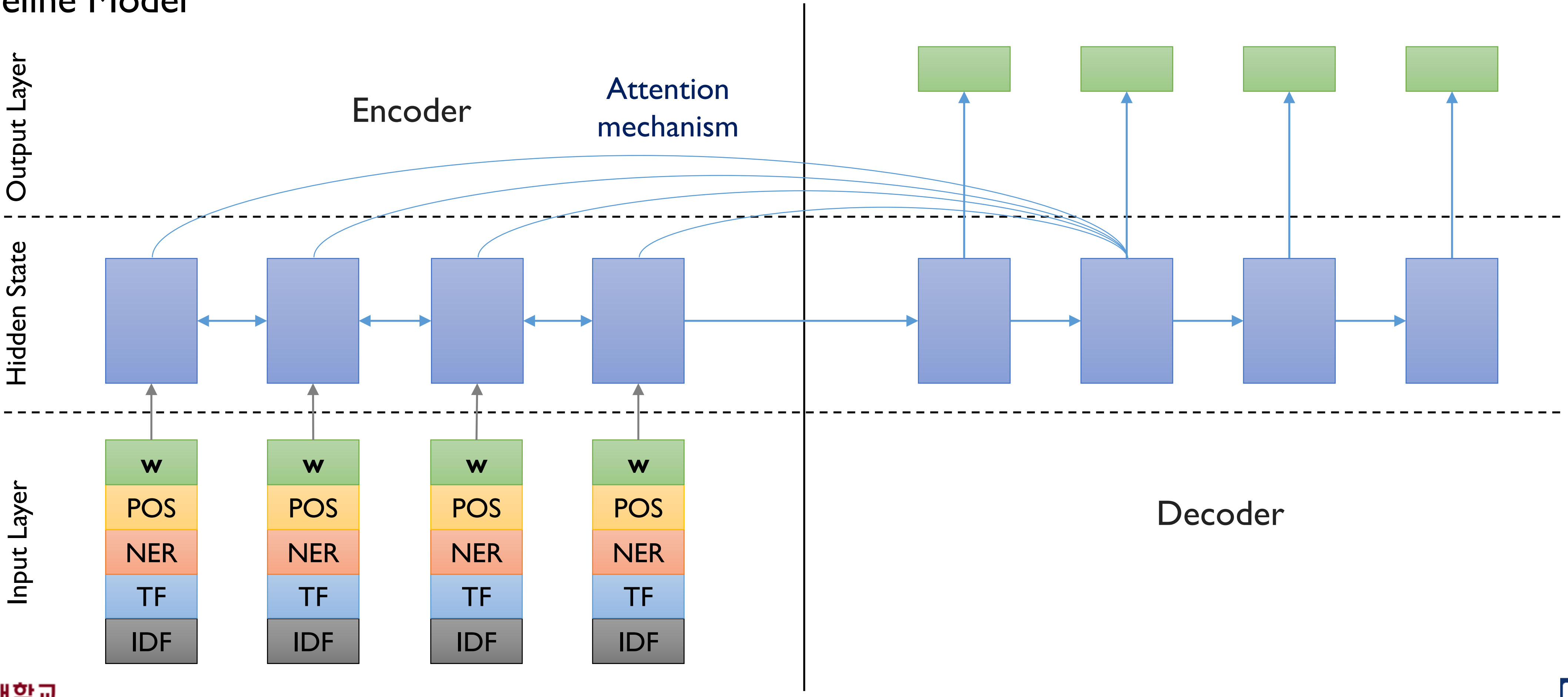
School of Industrial & Management Engineering

Korea University

Seq2Seq RNN-based Abstractive Summarization

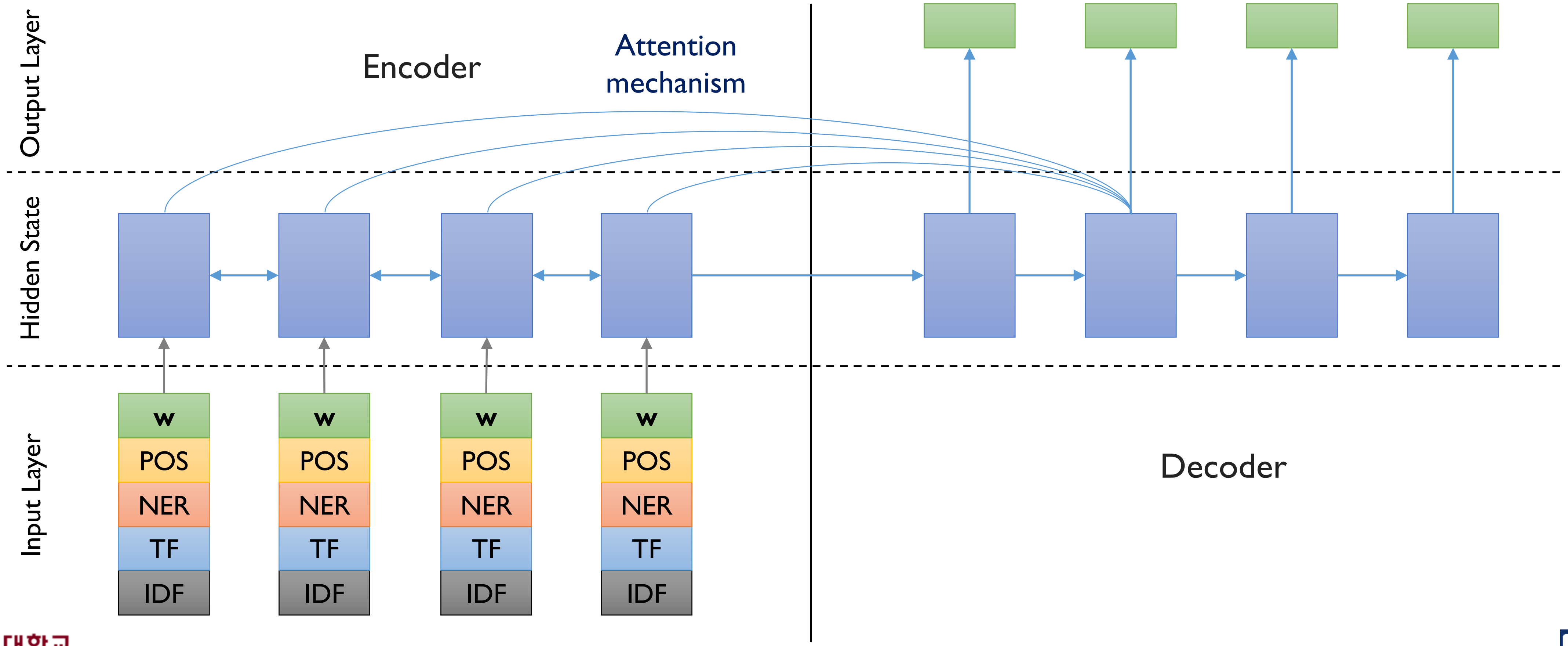
Nallapati, R., Zhou, B., Gulcehre, C., & Xiang, B. (2016). Abstractive Text Summarization using Sequence-to-sequence RNNs and beyond. *arXiv preprint arXiv:1602.06023*.

- Baseline Model



Seq2Seq RNN-based Abstractive Summarization

- Baseline Model
  - ✓ Encoder: bidirectional GRU-RNN
  - ✓ Decoder: uni-directional GRU-RNN with an attention mechanism

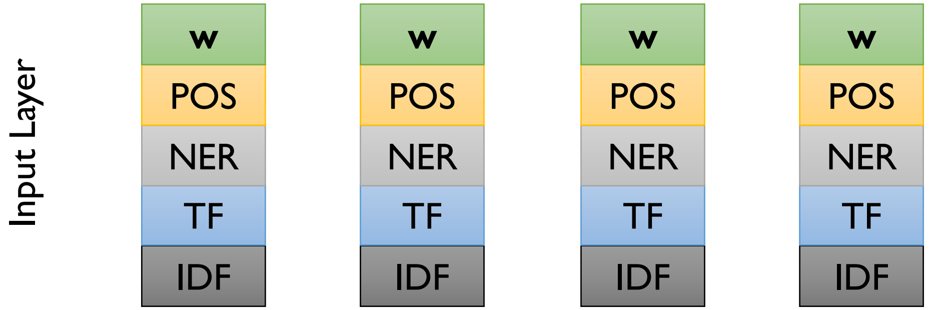


Seq2Seq RNN-based Abstractive Summarization

- Baseline Model
  - ✓ Encoder: bidirectional GRU-RNN
  - ✓ Decoder: uni-directional GRU-RNN with an attention mechanism
  - ✓ Large Vocabulary Trick (LVT)
    - ▪ The decoder-vocabulary of each mini-batch is restricted to words in the source documents of that batch.
    - ▪ The most frequent words in the target dictionary are added until the vocabulary reaches a fixed size.
    - ▪ Reduce the size of the soft-max layer of the decoder (the main computational bottleneck)
    - ▪ Well suited to summarization since a large proportion of the words in the summary comes from the source document in any case.

Seq2Seq RNN-based Abstractive Summarization

- Capturing Keywords using Feature-rich Encoder
  - ✓ Create additional look-up based embedding matrices for the vocabulary of each tag-type.
  - ✓ TF/IDF
    - Convert the value into categorical values by discretizing it into a fixed number of bins and use one-hot representations to indicate the bin number it falls into.
  - ✓ For each word in the source document, authors simply look-up its embeddings from all of its associated tags and concatenate them into a single long vector.
  - ✓ On the target side, only word-embeddings were used.



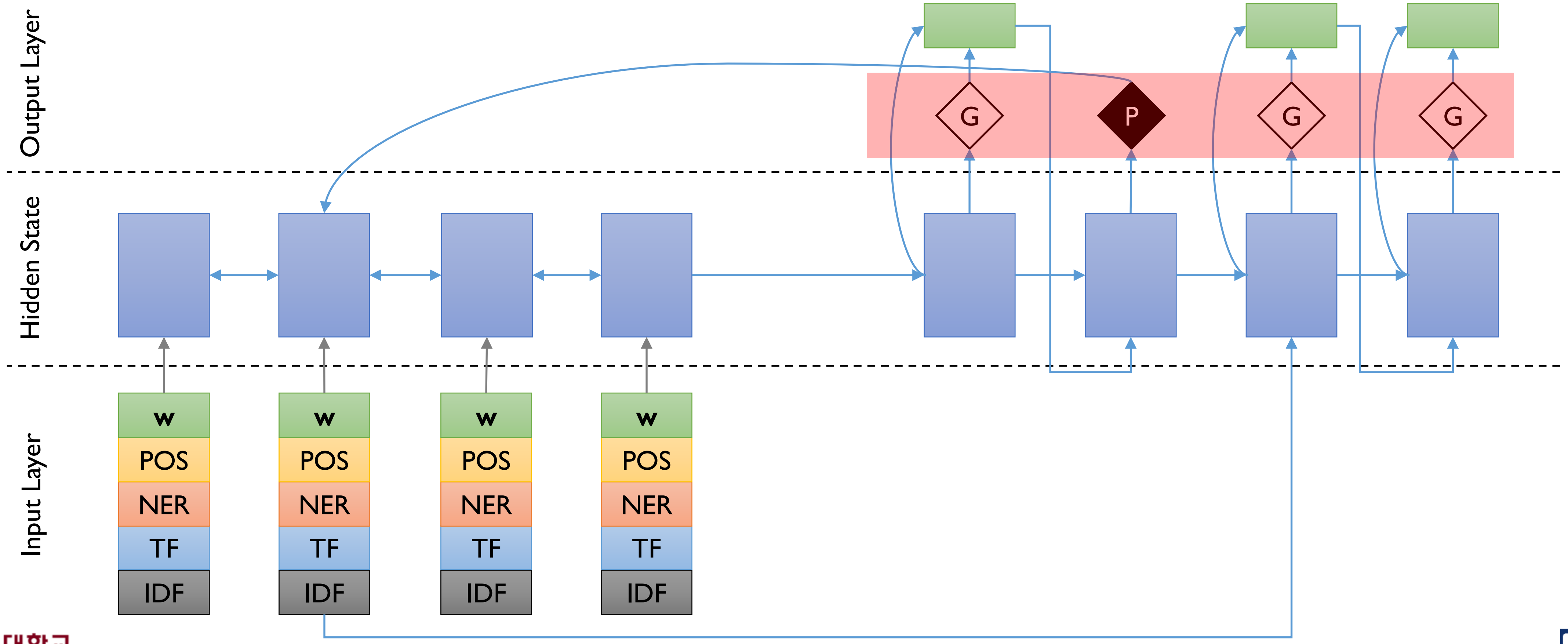
Input Layer

w	w	w	w
POS	POS	POS	POS
NER	NER	NER	NER
TF	TF	TF	TF
IDF	IDF	IDF	IDF

Seq2Seq RNN-based Abstractive Summarization

- Modeling Rare/Unseen Words using Switching Generator-Pointer

  ✓ The decoder is equipped with a **switch** that decides between using the generator or a pointer at every time step



Seq2Seq RNN-based Abstractive Summarization

- Modeling Rare/Unseen Words using Switching Generator-Pointer

  ✓ Switch: a sigmoid activation function over a linear layer based on the entire available context at each time-step.

$$\boxed{P(s_i = 1)} = \sigma(\mathbf{v}^s \cdot (\mathbf{W}^s_h \mathbf{h}_i + \mathbf{W}^s_e \boxed{\mathbf{E}[o_{i-1}]} + \mathbf{W}^s_c \boxed{\mathbf{c}_i} + \mathbf{b}^s))$$

The probability of the switch turning on at the ith time-step of the decoder.

The embedding vector of the emission from the previous time step.

The attention-weighted context vector

Seq2Seq RNN-based Abstractive Summarization

- Modeling Rare/Unseen Words using Switching Generator-Pointer
  - ✓ Attention distribution over word positions in the document is used as the distribution to sample to pointer from.

$$\boxed{P_i^a(j)} \propto \exp(\mathbf{v}^a \cdot (\mathbf{W}_h^a \mathbf{h}_{i-1} + \mathbf{W}_e^a \mathbf{E}[o_{i-1}] + \mathbf{W}_c^a \boxed{\mathbf{h}_d} + \mathbf{b}^a))$$

The probability of the i^{th} time-step in the decoder pointing to the j^{th} position in the document

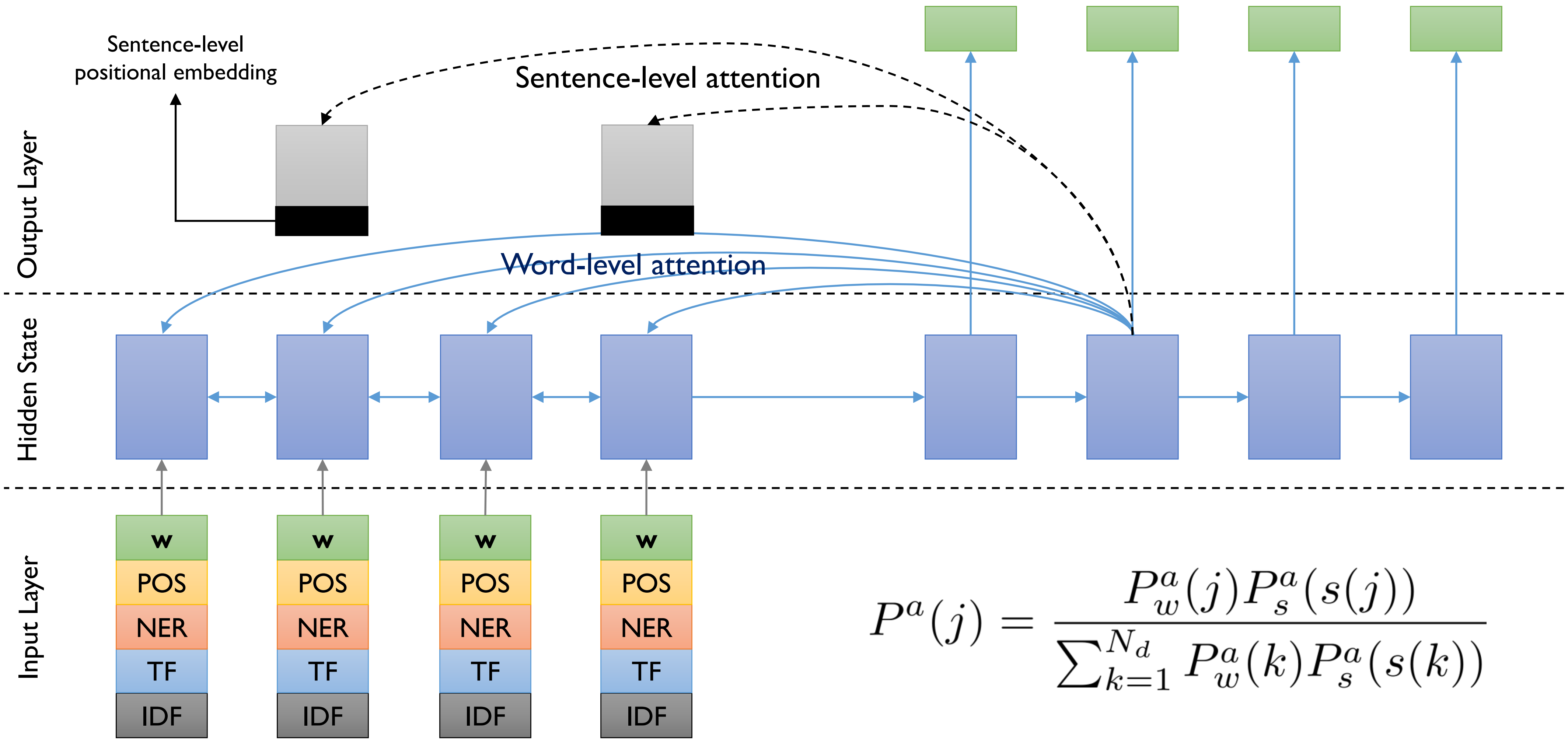
The encoder's hidden state at position j

$$\boxed{p_i} = \arg\max_j (P_i^a(j)) \quad \text{for} \quad j \in \{1, ..., N_d\}$$

Pointer value at i^{th} word-position in the summary

Seq2Seq RNN-based Abstractive Summarization

- Capturing Hierarchical Document Structure with Hierarchical Attention



$$P^a(j) = \frac{P^a_w(j)P^a_s(s(j))}{\sum_{k=1}^{N_d} P^a_w(k)P^a_s(s(k))}$$

Seq2Seq RNN-based Abstractive Summarization

- Capturing Hierarchical Document Structure with Hierarchical Attention

Word-level attention weight at j^th^ position of the source document

Sentence-level attention weight

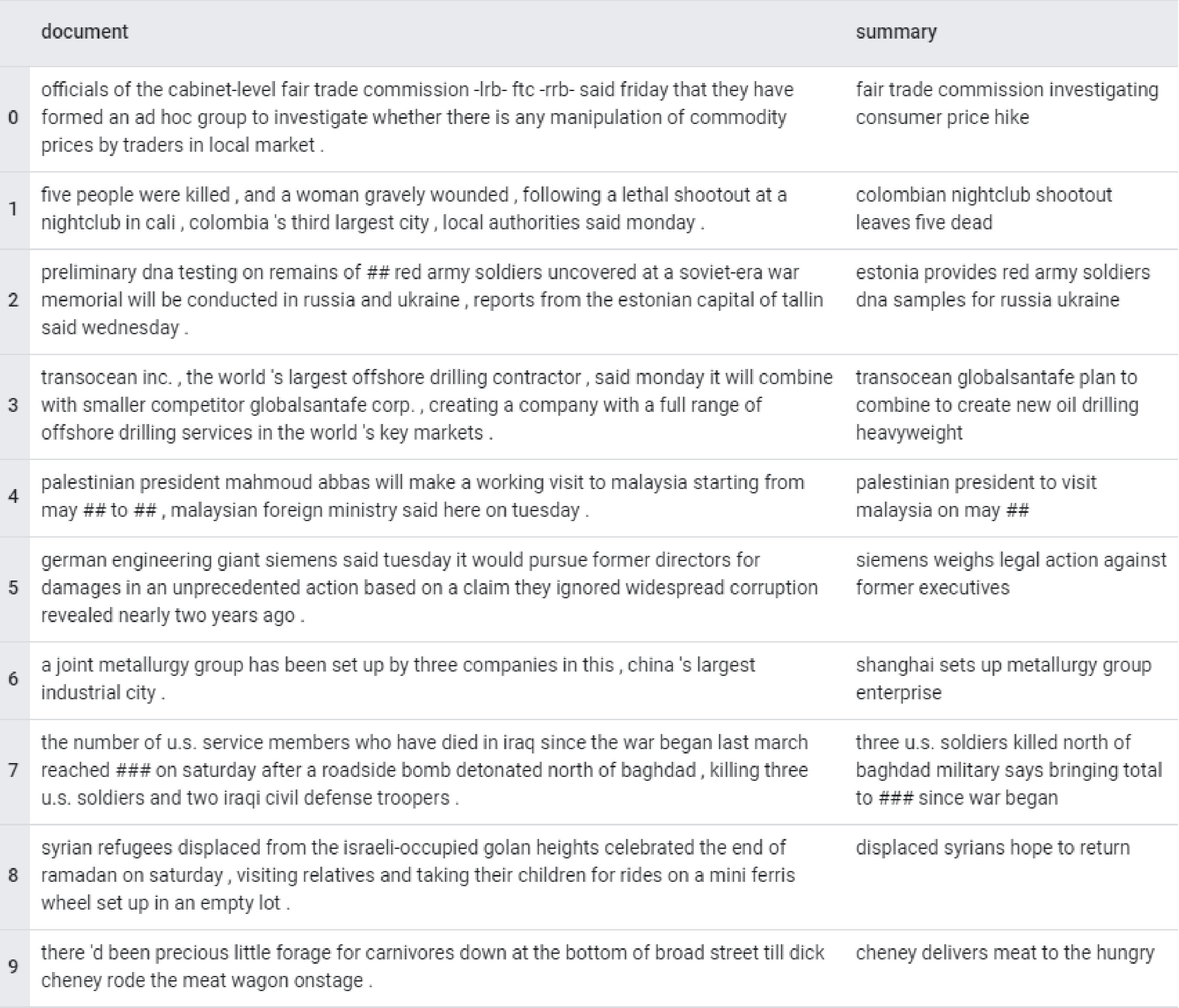
ID of the sentence at j^th^ word position

$$\boxed{P^a(j)} = \frac{\boxed{P^a_w(j)}\boxed{P^a_s(\boxed{s(j)})}}{\sum_{k=1}^{N_d} P^a_w(k) P^a_s(s(k))}$$

Re-scaled attention at the j^th^ word position

Seq2Seq RNN-based Abstractive Summarization

- Experiments and Results
  - ✓ Dataset: Gigaword
    - 3.8M training examples, 400K validation/test examples.
    - Authors randomly sampled 2000 examples for validation/test.
  - ✓ Training
    - 200-dim word2vec vectors, 400-dim hidden states for both encoder/decoder
  - ✓ Vocabulary dimension (with only first sentence)
    - Encoder(119,505), Decoder (68,885)
    - Restrict the decoder vocabulary size to 2,000 (LVT, batch size: 50)

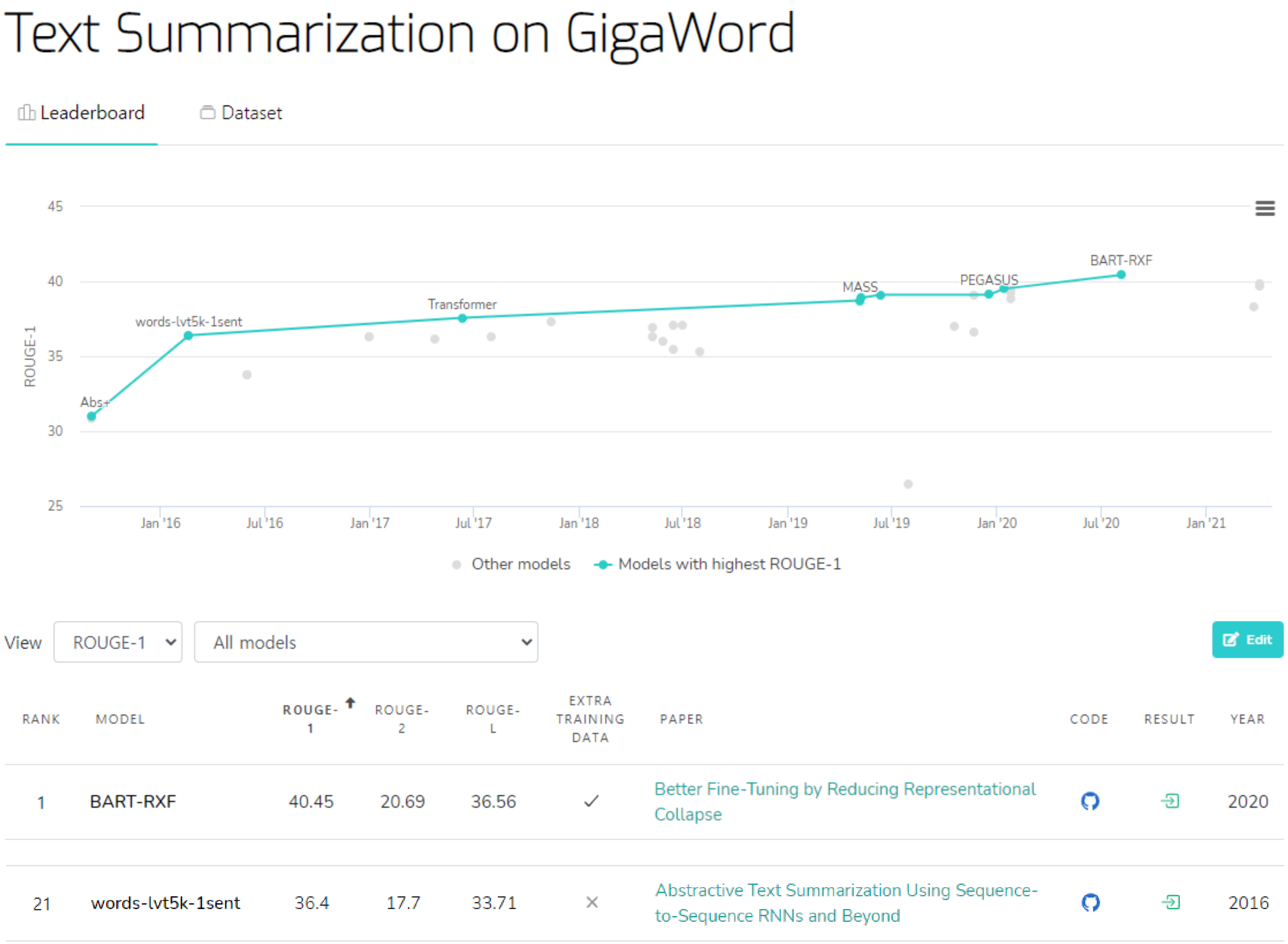


	document	summary
0	officials of the cabinet-level fair trade commission -lrb- ftc -rrb- said friday that they have formed an ad hoc group to investigate whether there is any manipulation of commodity prices by traders in local market .	fair trade commission investigating consumer price hike
1	five people were killed , and a woman gravely wounded , following a lethal shootout at a nightclub in cali , colombia 's third largest city , local authorities said monday .	colombian nightclub shootout leaves five dead
2	preliminary dna testing on remains of ## red army soldiers uncovered at a soviet-era war memorial will be conducted in russia and ukraine , reports from the estonian capital of tallin said wednesday .	estonia provides red army soldiers dna samples for russia ukraine
3	transocean inc. , the world 's largest offshore drilling contractor , said monday it will combine with smaller competitor globalsantafe corp. , creating a company with a full range of offshore drilling services in the world 's key markets .	transocean globalsantafe plan to combine to create new oil drilling heavyweight
4	palestinian president mahmoud abbas will make a working visit to malaysia starting from may ## to ## , malaysian foreign ministry said here on tuesday .	palestinian president to visit malaysia on may ##
5	german engineering giant siemens said tuesday it would pursue former directors for damages in an unprecedented action based on a claim they ignored widespread corruption revealed nearly two years ago .	siemens weighs legal action against former executives
6	a joint metallurgy group has been set up by three companies in this , china 's largest industrial city .	shanghai sets up metallurgy group enterprise
7	the number of u.s. service members who have died in iraq since the war began last march reached ### on saturday after a roadside bomb detonated north of baghdad , killing three u.s. soldiers and two iraqi civil defense troopers .	three u.s. soldiers killed north of baghdad military says bringing total to ### since war began
8	syrian refugees displaced from the israeli-occupied golan heights celebrated the end of ramadan on saturday , visiting relatives and taking their children for rides on a mini ferris wheel set up in an empty lot .	displaced syrians hope to return
9	there 'd been precious little forage for carnivores down at the bottom of broad street till dick cheney rode the meat wagon onstage .	cheney delivers meat to the hungry

https://www.tensorflow.org/datasets/catalog/gigaword

Seq2Seq RNN-based Abstractive Summarization

- Experiments and Results
  - ✓ Decoding
    - Beam search of size 5
    - Limit the size of summary to 30 words (model average: 7.8, ground truth 8.7 on average)



Text Summarization on GigaWord

RANK	MODEL	ROUGE-1 ↑	ROUGE-2	ROUGE-L	EXTRA TRAINING DATA	PAPER	CODE	RESULT	YEAR
1	BART-RXF	40.45	20.69	36.56	✓	Better Fine-Tuning by Reducing Representational Collapse			2020
21	words-lvt5k-1sent	36.4	17.7	33.71	✕	Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond			2016

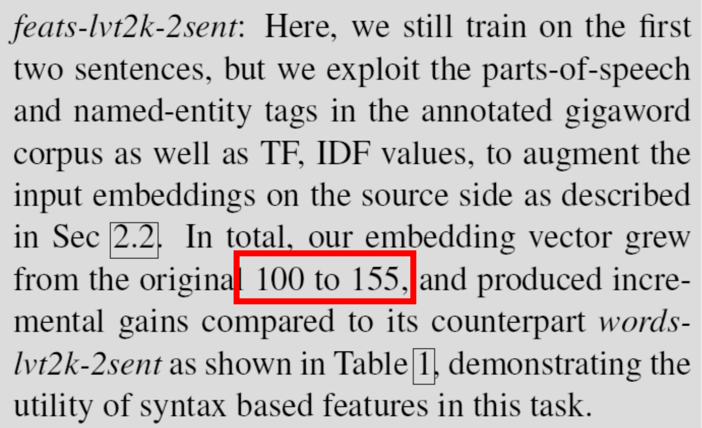
https://paperswithcode.com/sota/text-summarization-on-gigaword

Seq2Seq RNN-based Abstractive Summarization

- Experiments and Results

✓ Benchmark models

	Use additional features?	Use LVK?/ Dimension	Input length	Use hierarchical attention?	Use pointer?
word-lvt2k-1sent	No	Yes/2,000	1 sentence	No	No
word-lvt2k-2sent	No	Yes/2,000	2 sentences	No	No
word-lvt2k-2sent-hieratt	No	Yes/2,000	2 sentences	Yes	No
feat-lvt2k-2sent	Yes	Yes/2,000	2 sentences	No	No
feat-lvt2k-2sent	Yes	Yes/2,000	2 sentences	No	Yes

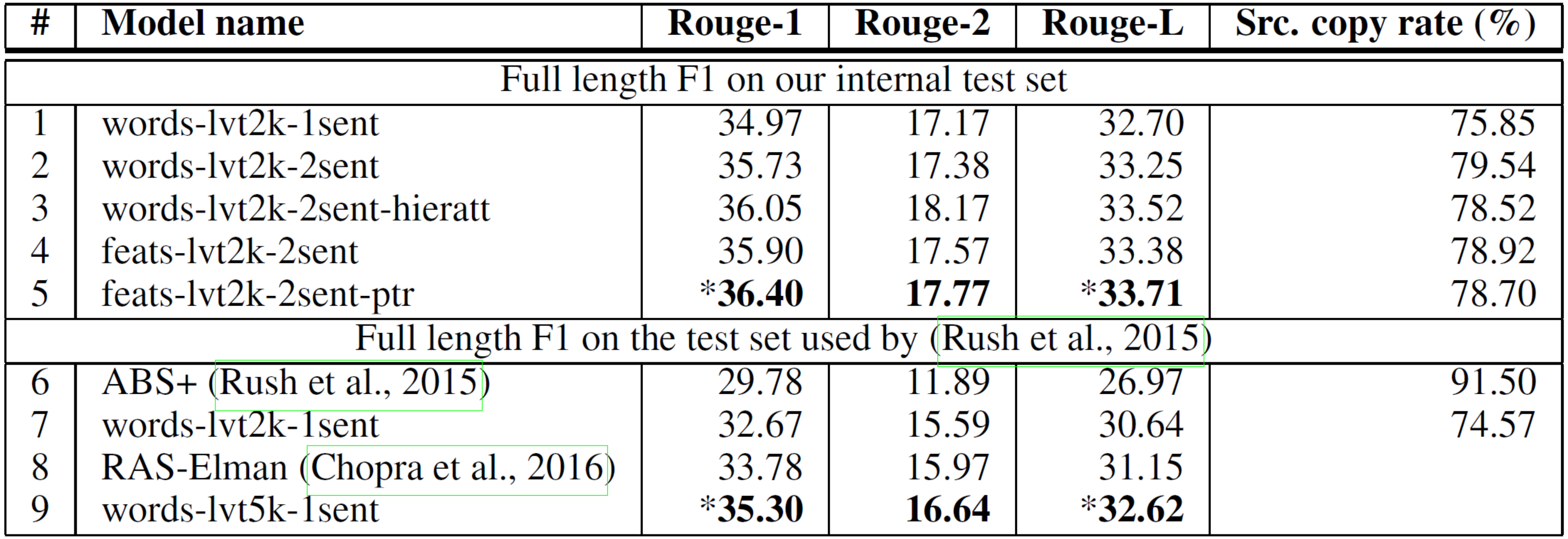


feats-lvt2k-2sent: Here, we still train on the first two sentences, but we exploit the parts-of-speech and named-entity tags in the annotated gigaword corpus as well as TF, IDF values, to augment the input embeddings on the source side as described in Sec 2.2. In total, our embedding vector grew from the original 100 to 155, and produced incremental gains compared to its counterpart *words-lvt2k-2sent* as shown in Table 1, demonstrating the utility of syntax based features in this task.

?

Seq2Seq RNN-based Abstractive Summarization

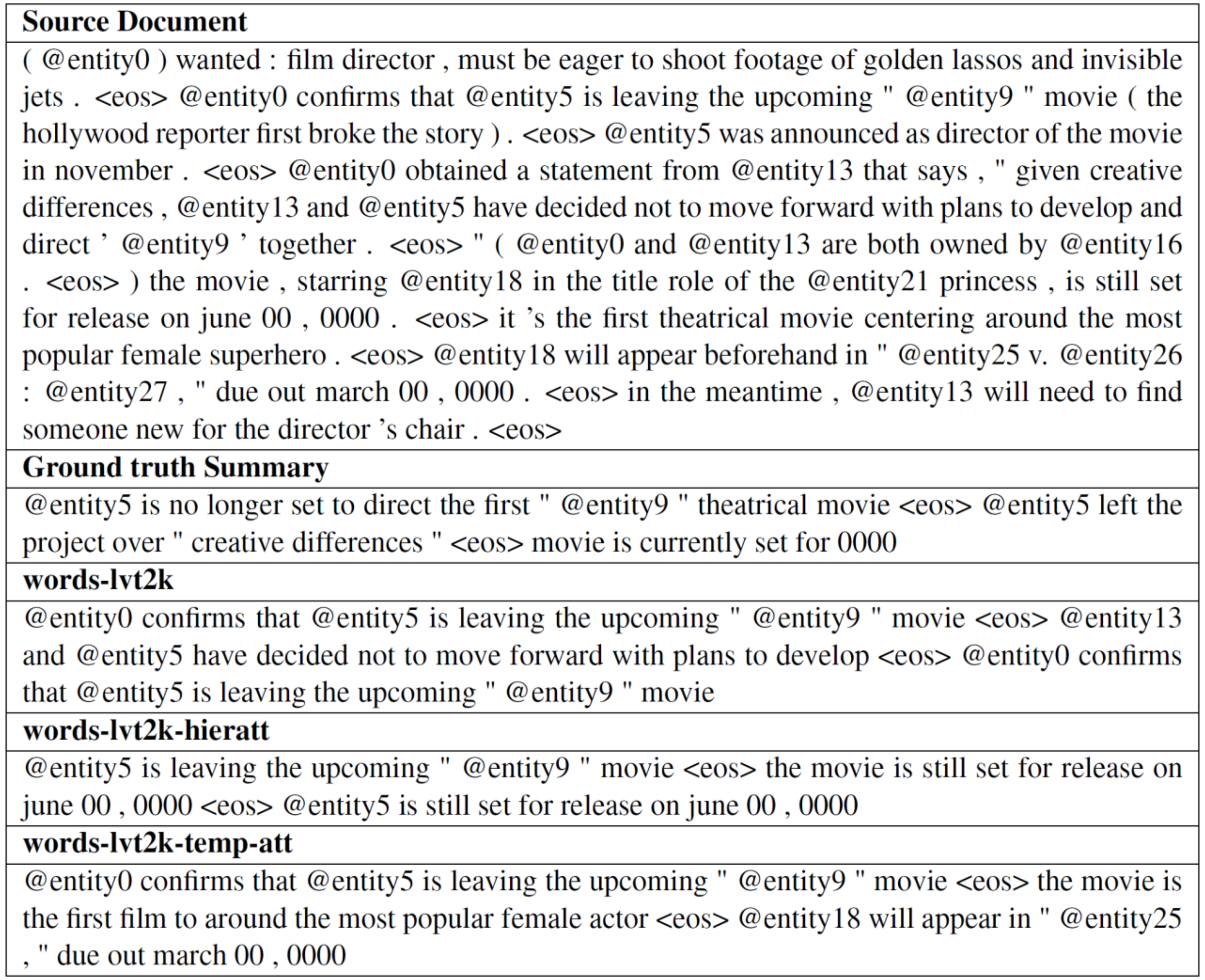
- Experiments and Results



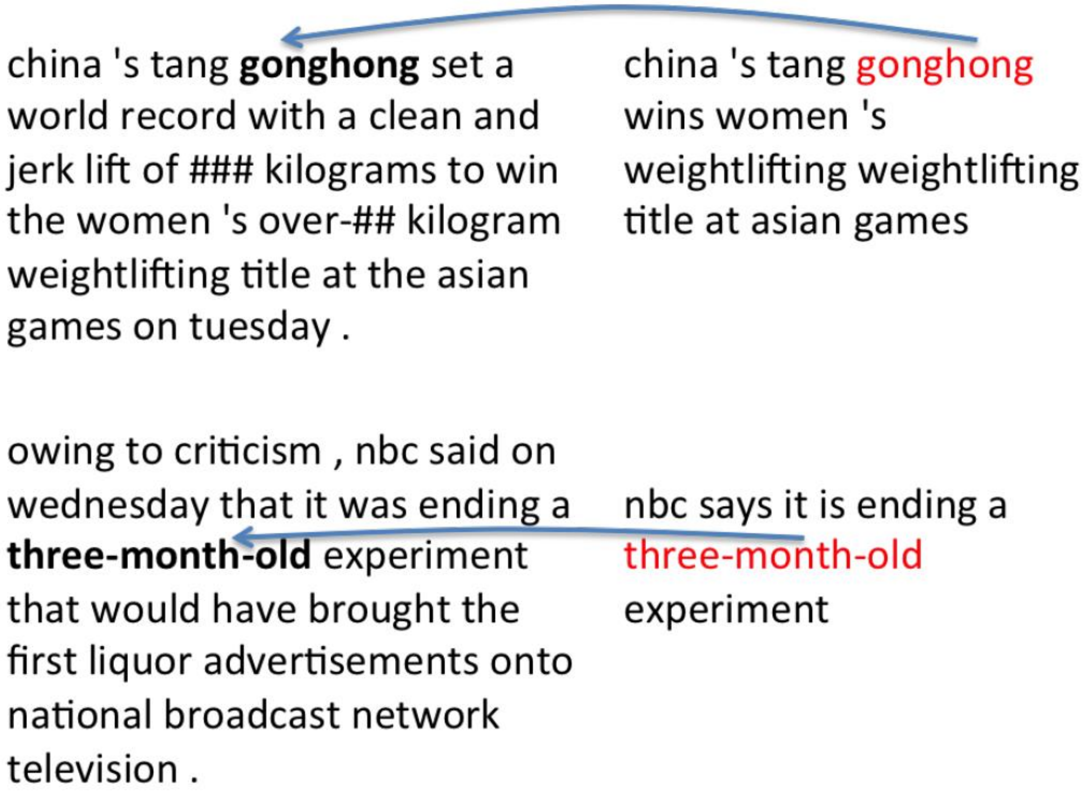
#	Model name	Rouge-1	Rouge-2	Rouge-L	Src. copy rate (%)
Full length F1 on our internal test set					
1	words-lvt2k-1sent	34.97	17.17	32.70	75.85
2	words-lvt2k-2sent	35.73	17.38	33.25	79.54
3	words-lvt2k-2sent-hieratt	36.05	18.17	33.52	78.52
4	feats-lvt2k-2sent	35.90	17.57	33.38	78.92
5	feats-lvt2k-2sent-ptr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

Seq2Seq RNN-based Abstractive Summarization

- Experiments and Results



Source Document
(@entity0) wanted : film director , must be eager to shoot footage of golden lassos and invisible jets . <eos> @entity0 confirms that @entity5 is leaving the upcoming " @entity9 " movie (the hollywood reporter first broke the story) . <eos> @entity5 was announced as director of the movie in november . <eos> @entity0 obtained a statement from @entity13 that says , " given creative differences , @entity13 and @entity5 have decided not to move forward with plans to develop and direct ' @entity9 ' together . <eos> " (@entity0 and @entity13 are both owned by @entity16 . <eos>) the movie , starring @entity18 in the title role of the @entity21 princess , is still set for release on june 00 , 0000 . <eos> it 's the first theatrical movie centering around the most popular female superhero . <eos> @entity18 will appear beforehand in " @entity25 v. @entity26 : @entity27 , " due out march 00 , 0000 . <eos> in the meantime , @entity13 will need to find someone new for the director 's chair . <eos>
Ground truth Summary
@entity5 is no longer set to direct the first " @entity9 " theatrical movie <eos> @entity5 left the project over " creative differences " <eos> movie is currently set for 0000
words-lvt2k
@entity0 confirms that @entity5 is leaving the upcoming " @entity9 " movie <eos> @entity13 and @entity5 have decided not to move forward with plans to develop <eos> @entity0 confirms that @entity5 is leaving the upcoming " @entity9 " movie
words-lvt2k-hieratt
@entity5 is leaving the upcoming " @entity9 " movie <eos> the movie is still set for release on june 00 , 0000 <eos> @entity5 is still set for release on june 00 , 0000
words-lvt2k-temp-att
@entity0 confirms that @entity5 is leaving the upcoming " @entity9 " movie <eos> the movie is the first film to around the most popular female actor <eos> @entity18 will appear in " @entity25 , " due out march 00 , 0000



china 's tang **gonghong** set a world record with a clean and jerk lift of ### kilograms to win the women 's over-## kilogram weightlifting title at the asian games on tuesday .

china 's tang gonghong wins women 's weightlifting weightlifting title at asian games

owing to criticism , nbc said on wednesday that it was ending a **three-month-old** experiment that would have brought the first liquor advertisements onto national broadcast network television .

nbc says it is ending a three-month-old experiment

Pointer-Generator Networks

See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

- Advanced version of Seq2Seq RNN-based abstractive Summarization
  - ✓ Hybrid pointer-generator network
    - Can copy words from the source text via pointing, which aids accurate reproduction of information
    - Retraining the ability to produce novel words through the generator
  - ✓ Coverage
    - To Keep track of what has been summarized

Pointer-Generator Networks

$$\text{loss}_t = -\log P(\mathbf{w}^*_t)$$

$$\text{loss} = \frac{1}{T}\sum_{i=1}^{T}\text{loss}_t$$

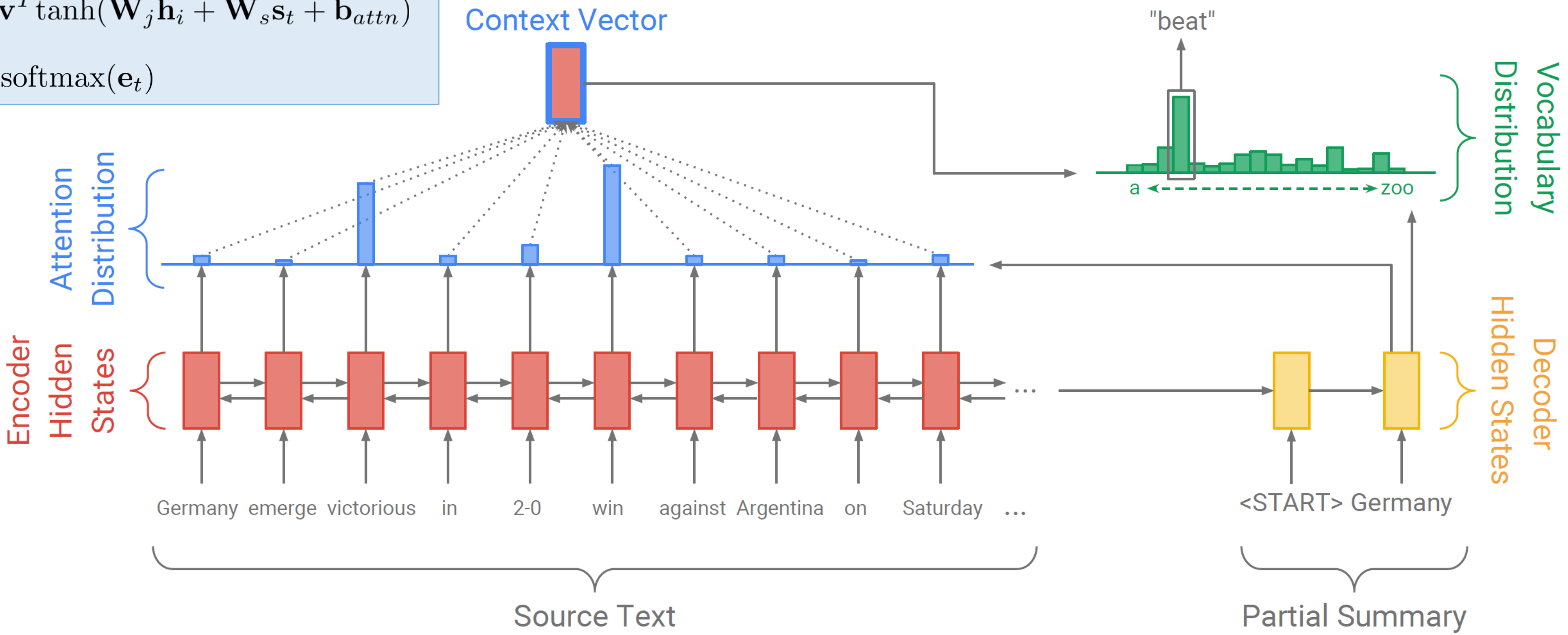
- Baseline Model

$$e^t_i = \mathbf{v}^T \tanh(\mathbf{W}_j \mathbf{h}_i + \mathbf{W}_s \mathbf{s}_t + \mathbf{b}_{attn})$$

$$\mathbf{a}^t = \text{softmax}(\mathbf{e}_t)$$

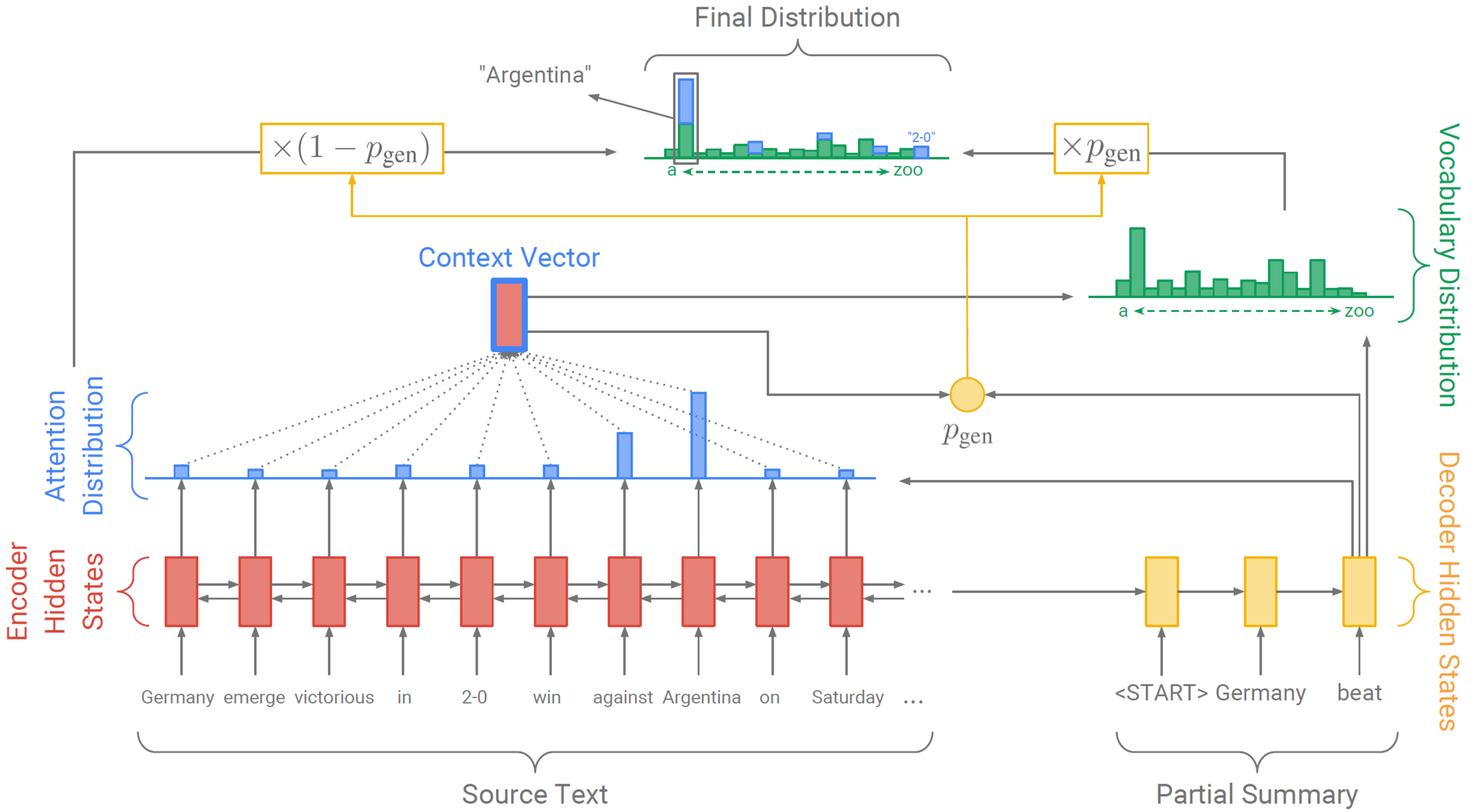
$$\mathbf{h}^*_t = \sum_i \mathbf{a}^t_i \mathbf{h}_i$$

$$P_{vocab} = \text{softmax}(\mathbf{V}'(\mathbf{V}[\mathbf{s}_t, \mathbf{h}^*_t] + \mathbf{b}) + \mathbf{b}')$$



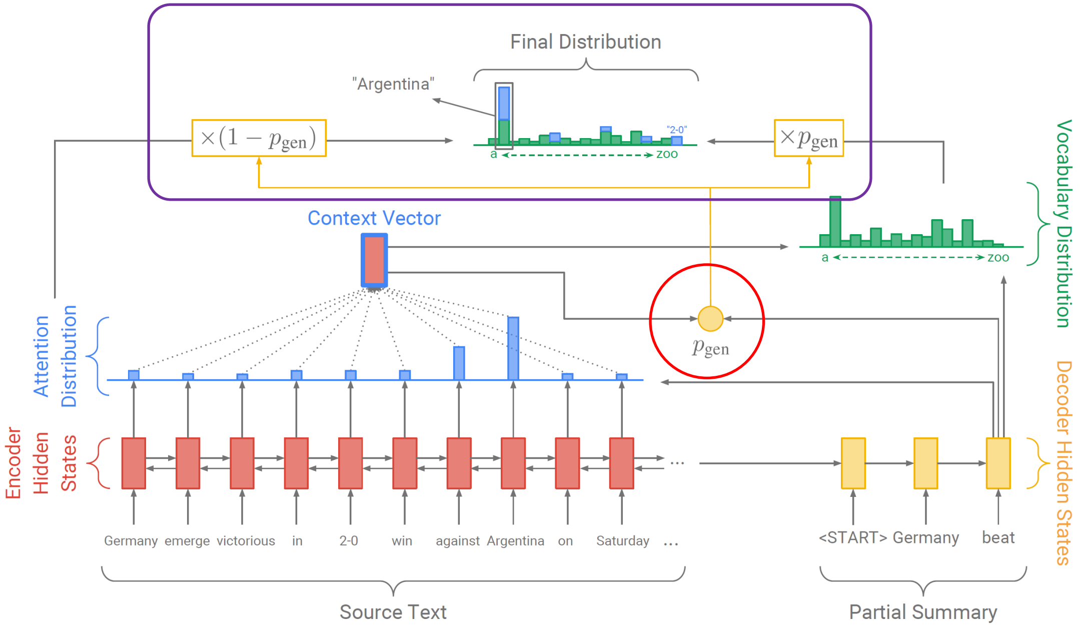
Pointer-Generator Networks

- Pointer-Generator Network



Pointer-Generator Networks

- Pointer-Generator Network



Generation probability

$$p_{gen} = \sigma(\mathbf{w}_{h^*}^T \mathbf{h}_t^* + \mathbf{w}_s^T \mathbf{s}_t + \mathbf{w}_x^T \mathbf{x}_t + b_{ptr})$$

p_{gen} is used as a soft switch to choose between generating a word from the vocabulary by sampling from P_{vocab} or copying a word from the input sequence by sampling from the attention distribution.

$$P(w) = p_{gen} P_{vocab}(w) + (1 - p_{gen}) \sum_{i:w_i=w} \mathbf{a}_i^T$$

Able to produce OOV words

Pointer-Generator Networks

- Coverage Mechanism
  - ✓ Repetition is a common problem for sequence-to-sequence models, and is pronounced when generating multi-sentence text
  - ✓ Maintain a coverage vector c^t, which is the sum of attention distributions over all previous decoder timesteps

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

  - ▪ Can be understood as a distribution over the source document words that represents the degree of coverage that those words have received from the attention mechanism so far
  - ▪ The coverage vector is used as an extra input to the attention mechanism

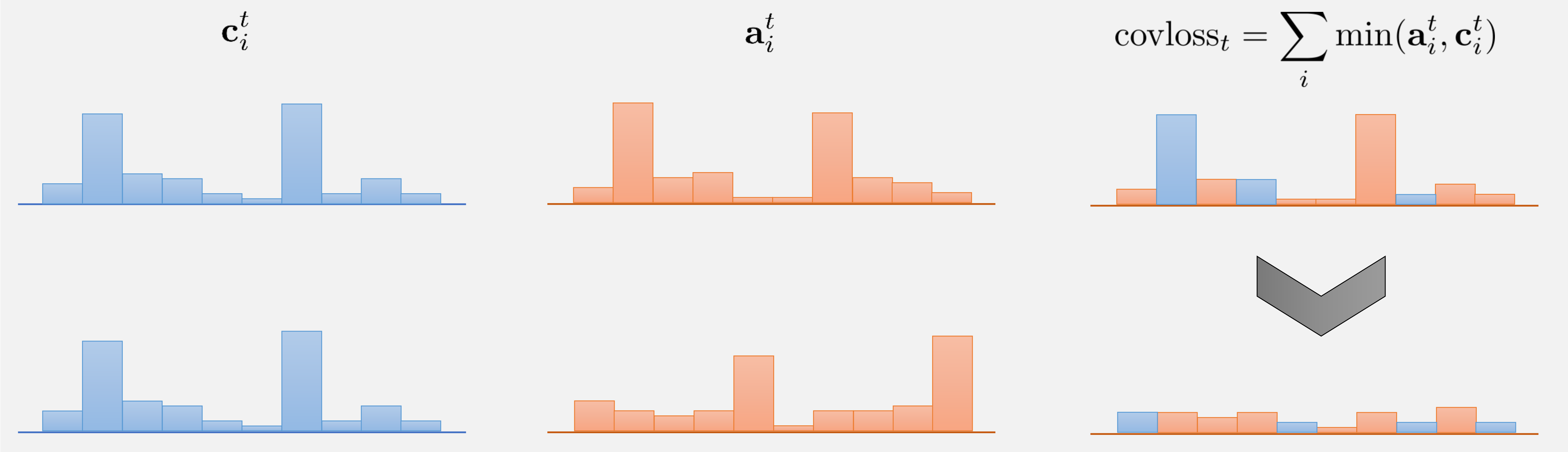
$$e_i^t = \mathbf{v}^T \tanh(\mathbf{W}_j \mathbf{h}_i + \mathbf{W}_s \mathbf{s}_t + \mathbf{w}_c \mathbf{c}_i^t + \mathbf{b}_{attn})$$

$$\mathbf{a}^t = \mathrm{softmax}(\mathbf{e}_t)$$

Pointer-Generator Networks

- Coverage Mechanism (cont')
  - ✓ Coverage loss to penalize repeatedly attending to the same locations

$$\text{covloss}_t = \sum_i \min(\mathbf{a}_i^t, \mathbf{c}_i^t) \qquad \text{loss}_t = -\log P(\mathbf{w}_t^*) + \lambda \sum_i \min(\mathbf{a}_i^t, \mathbf{c}_i^t)$$

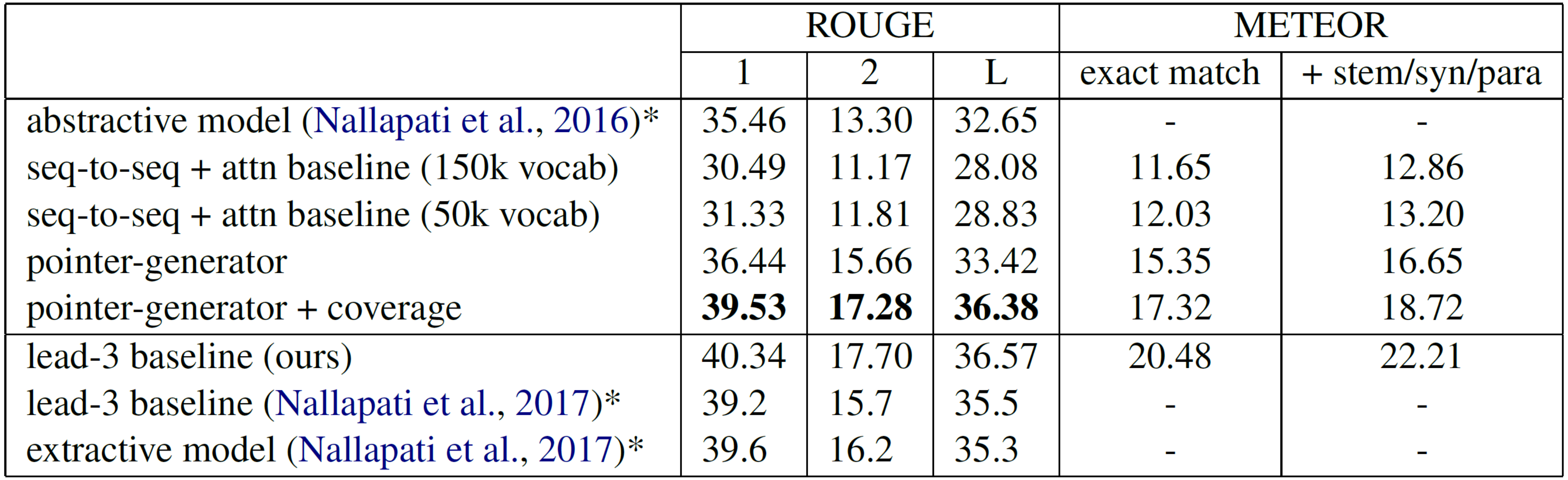


Pointer-Generator Networks

- Experiments
  - ✓ Dataset: CNN/Daily Mail
    - 287,226 training pairs, 13,368 validation pairs, and 11,490 test pairs
    - Source documents: 781 tokens on average
    - Summaries: 3.75 sentences (56 tokens on average)
  - ✓ Experimental settings
    - 128-dim word embeddings, 256-dim hidden states, 50K words for both source and target (word embeddings were not pre-trained, LVT was not used)
    - Each article is truncated to 400 tokens and the summary is limited to 100 tokens for training and 120 for test

Pointer-Generator Networks

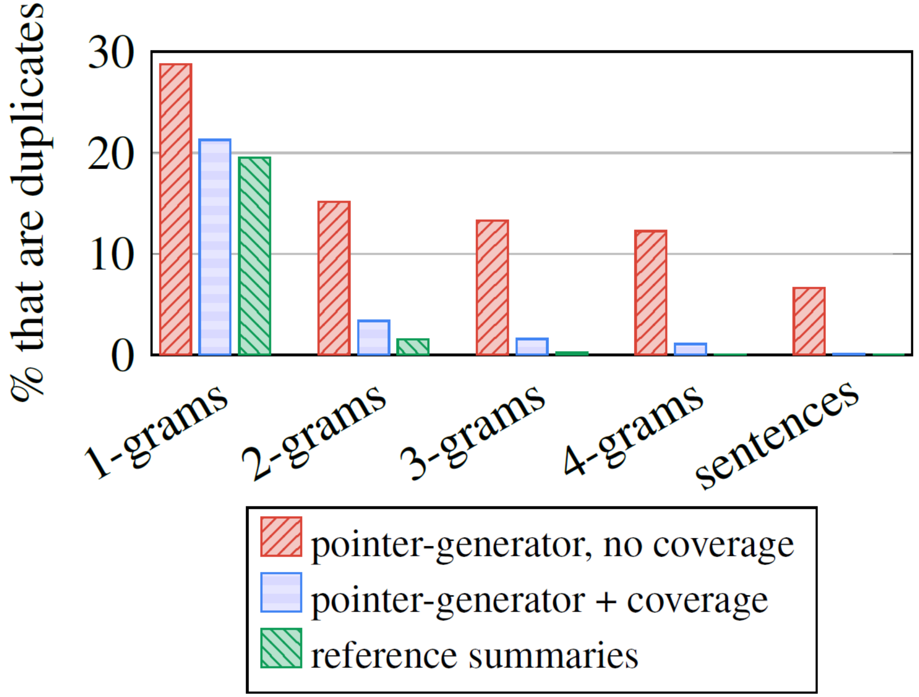
- Experimental Results: Performance



	ROUGE			METEOR	
	1	2	L	exact match	+ stem/syn/para
abstractive model (Nallapati et al., 2016)*	35.46	13.30	32.65	-	-
seq-to-seq + attn baseline (150k vocab)	30.49	11.17	28.08	11.65	12.86
seq-to-seq + attn baseline (50k vocab)	31.33	11.81	28.83	12.03	13.20
pointer-generator	36.44	15.66	33.42	15.35	16.65
pointer-generator + coverage	**39.53**	**17.28**	**36.38**	17.32	18.72
lead-3 baseline (ours)	40.34	17.70	36.57	20.48	22.21
lead-3 baseline (Nallapati et al., 2017)*	39.2	15.7	35.5	-	-
extractive model (Nallapati et al., 2017)*	39.6	16.2	35.3	-	-

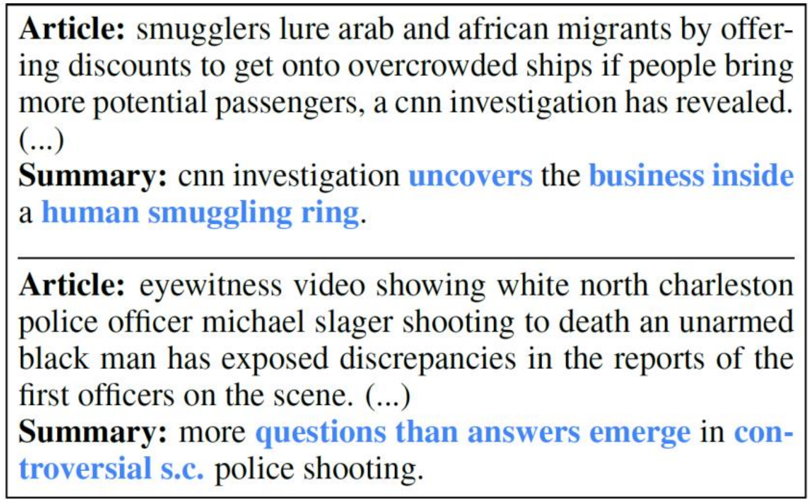
Pointer-Generator Networks

- Experimental Results: Effect of coverage vector



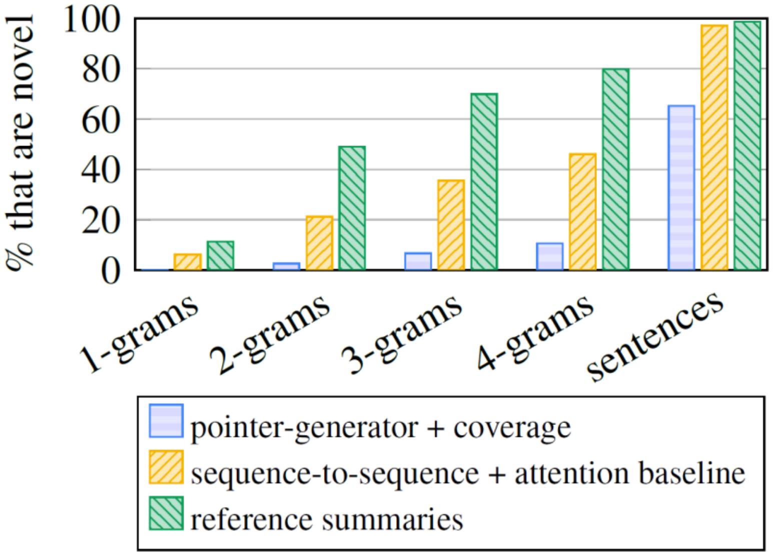
Pointer-Generator Networks

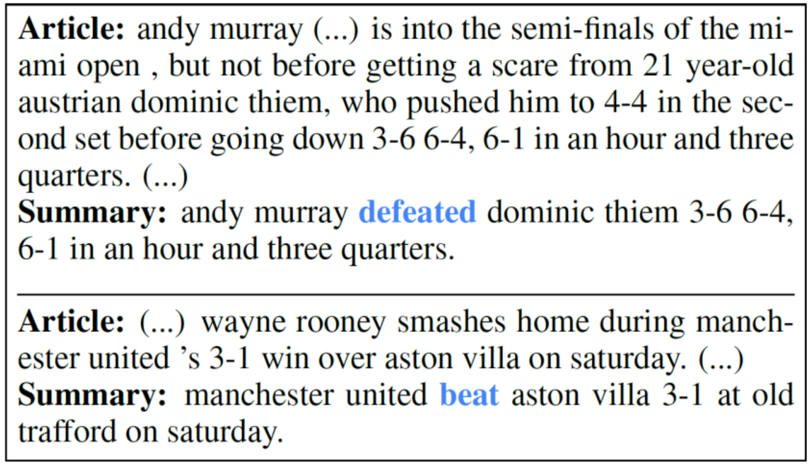
- Discussions
  - ✓ Limitation of ROUGE-based metrics for highly abstractive summaries
  - ✓ More novel but erroneous vs. less novel but correct



Article: smugglers lure arab and african migrants by offering discounts to get onto overcrowded ships if people bring more potential passengers, a cnn investigation has revealed. (...)
Summary: cnn investigation **uncovers** the **business inside** a **human smuggling ring**.

Article: eyewitness video showing white north charleston police officer michael slager shooting to death an unarmed black man has exposed discrepancies in the reports of the first officers on the scene. (...)
Summary: more **questions than answers emerge** in **controversial s.c.** police shooting.





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

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

ANY
questions?