Abstractive Text Summarization

using Sequence-to-sequence RNNs and Beyond

Written by: Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çaglar Gulçehre, and Bing Xiang

From IBM and Université de Montréal

Presented by: **Ehsan Amjadian @RBC**

Summary

- Objective:
    - Abstractive summarization using attentional encoder-decoder RNNs
- Main Contributions:
    - Modeling keywords
    - Explicitly capturing syntactic hierarchy
    - Decoder able to emit rare or unseen words
- State-of-the-art performance (2016)
- Creating and benchmarking dataset for *multi-sentence summary* summarization

What is abstractive text summarization

- Generating summaries (a few short sentences) that capture the salient ideas of the text
- Abstractive: not a mere selection of existing sentences but a compressed rephrasing with potentially unseen words
- Why not just use MT seq2seq?
  - Short target summary
  - Target summary does not necessarily depend on source length
  - Summarization is lossy (optimally compress) vs. lossless in MT
  - Strong almost one to one word alignment in MT and not in summarizaton

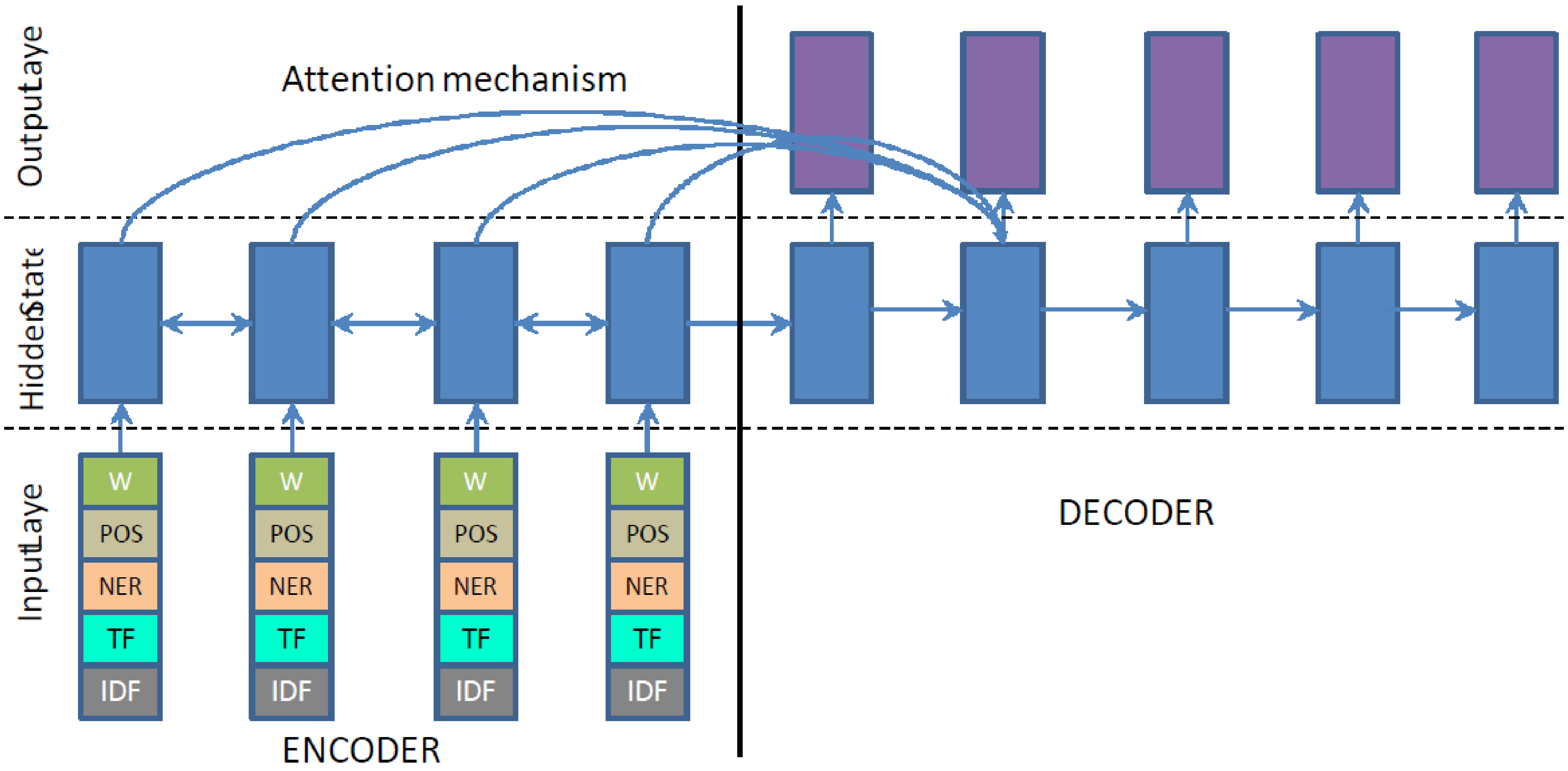
Attentional Encoder-Decoder RNN with LVT

- They adopt the *Bahdanau* (2014) MT seq2seq model
- Encoder: bidirectional GRU
- Decoder:
  - unidirectional GRU
  - attention over source hid.
  - Softmax over target vocab
- Same hidden size for encoder and decoder
- *Jean* (2014) LVT, target vocab:
  - Source words of each batch + high frequency target words until cap

Feature rich decoder

- Identifying key concepts an entities
- POS tags
- NER tags
- TF and IDF stats
- Embedding dictionary for vocabulary of each tag types
- Continuous features (TF/IDF) categorical by discretized into fixed # of bins. Bin# = one-hot encoding
- Concatenate all the embeddings
- Target only word-based embeddings

Feature rich decoder



Rare/Unseen Words (generator/pointer switch)

- Keywords or named entities in the test document can be unseen or rare with respect to training data
- And dec vocab is fixed at training so dec cannot omit these rare/ unseen words
- Common solution: emit "UNK" token
- Better solution: pointer network
- Switch decides between generator/pointer at each time step
  - Switch 1: generator
  - Switch 0: pointer
- Switch: sigmoid activation over the entire available context at each time step

Generator/Pointer Switch

$$\begin{aligned} P(s_i = 1) &= \sigma(\mathbf{v}^s \cdot (\mathbf{W}_h^s \mathbf{h}_i + \mathbf{W}_e^s \mathbf{E}[o_{i-1}] \\ &+ \mathbf{W}_c^s \mathbf{c}_i + \mathbf{b}^s)), \end{aligned}$$

i = decoder timestep

h_i = hidden state

$E[O_i -1]$ = embedding vec of emission from previous time step

c_i = attention weighted context vector

Ws are switch parameters

The Pointer

- Pointer = same attention dist over source [j over source document]

$$
\begin{aligned}
P_i^a(j) &\propto \exp(\mathbf{v}^a \cdot (\mathbf{W}_h^a \mathbf{h}_{i-1} + \mathbf{W}_e^a \mathbf{E}[o_{i-1}] \\
&+ \mathbf{W}_c^a \mathbf{h}_j^d + \mathbf{b}^a)), \\
p_i &= \arg\max_j (P_i^a(j)) \text{ for } j \in \{1, \ldots, N_d\}
\end{aligned}
$$

- P_i = the pointer value at *i*th position in the summary
- h_j^d = *encoder hidden state at position j*
- At training the model will have explicit training information whenever summary word not in dec vocab
- When the same OOV in more than one position□point to 1st occurrence position.

Pointer network optimization

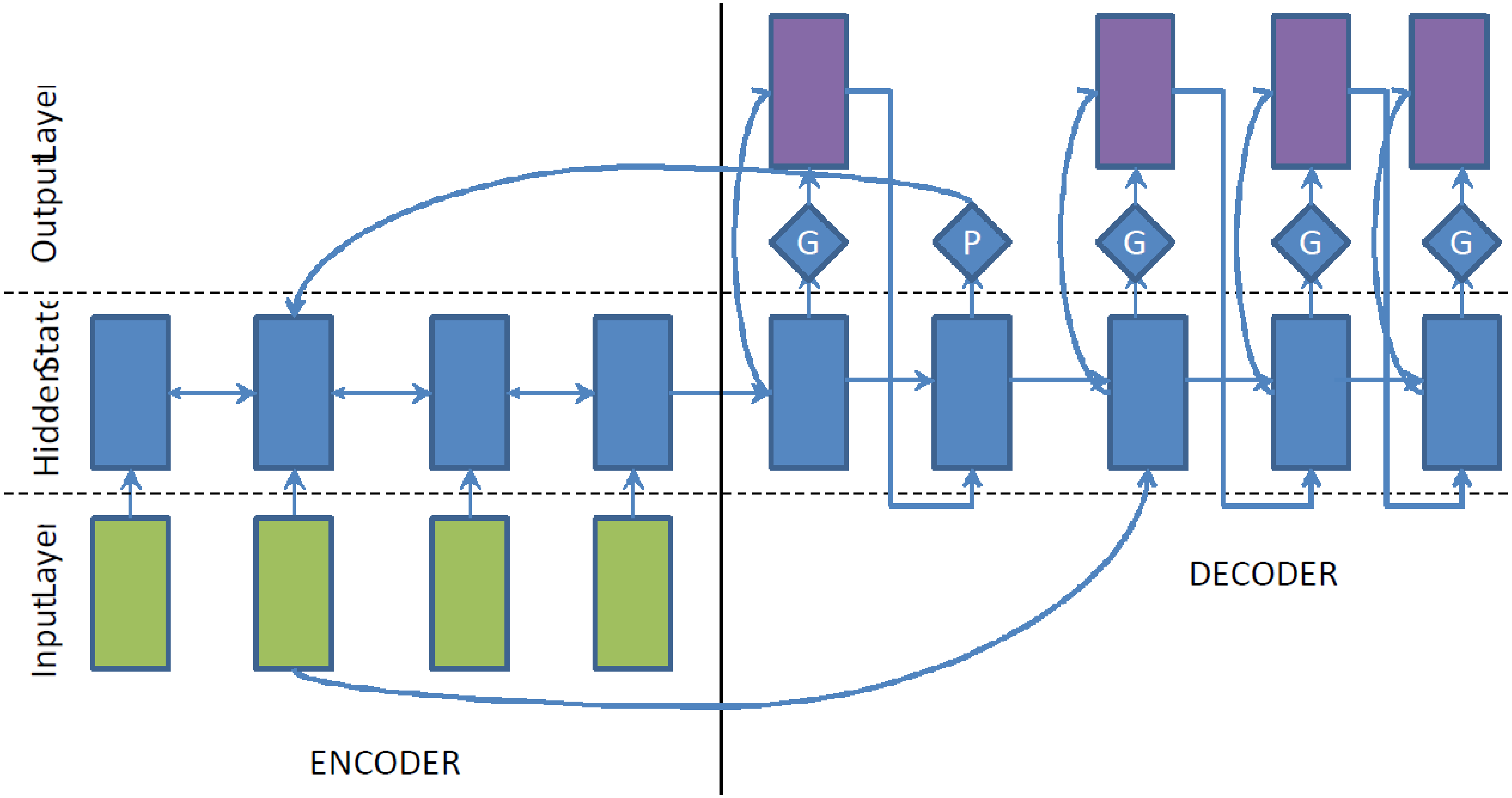
- Optimize conditional log-likelihood with additional regularization

$$\log P(\mathbf{y}|\mathbf{x}) = \sum_i (g_i \log\{P(y_i|\mathbf{y}_{-i}, \mathbf{x})P(s_i)\}$$

$$+(1 - g_i) \log\{P(p(i)|\mathbf{y}_{-i}, \mathbf{x})(1 - P(s_i))\})$$

- *y* and *x* are doc and summary words
- g_i is an indicator function = 1 for OOV at position *i and*
- At test time thy use *$P(s_i)$* to decide. (argmax of posterior of gen/point)

Pointer Illustration



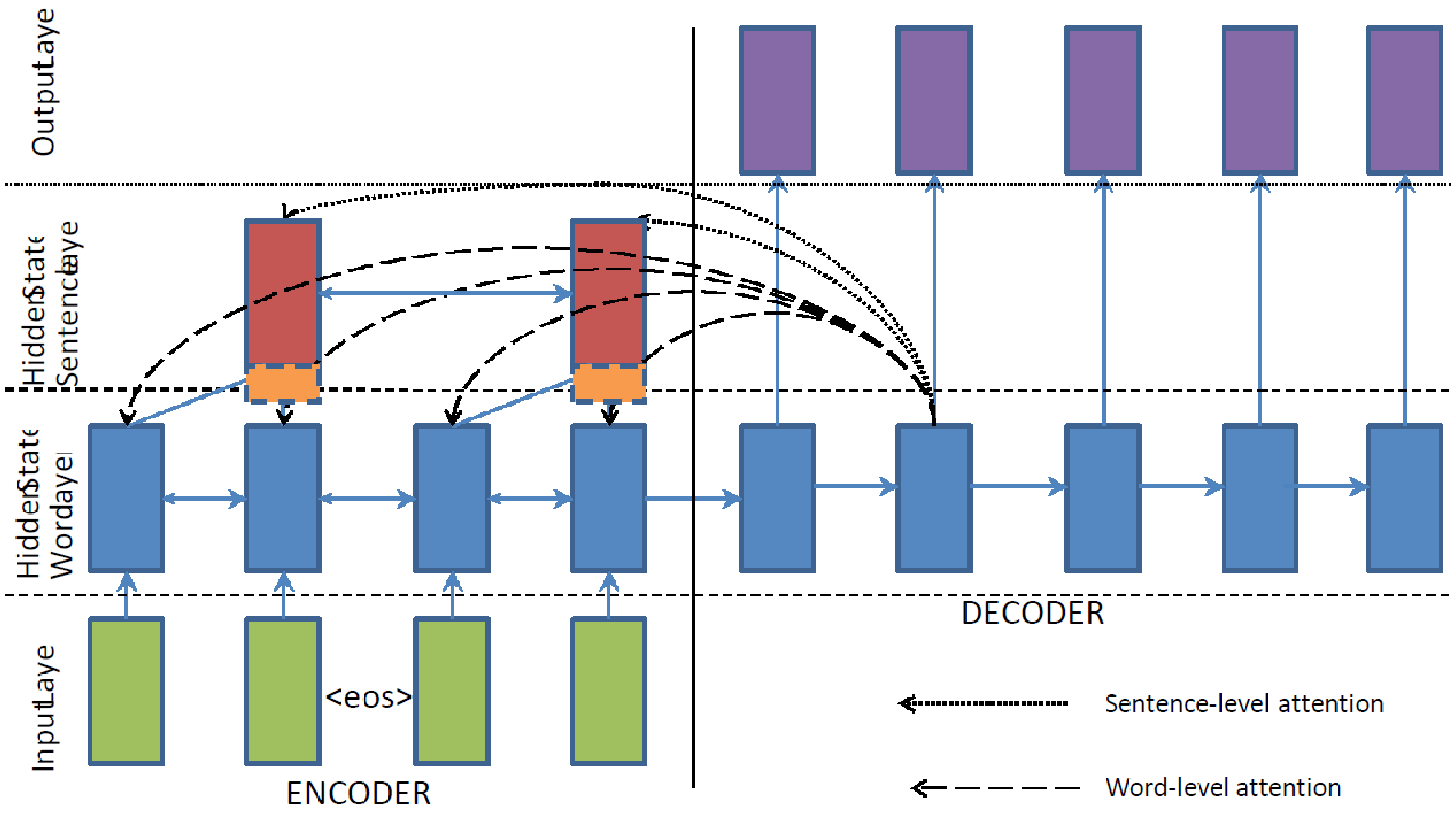
Hierarchical Doc Structure with Hierarchical Attention

- Useful datasets where the source document is long
- Two-level importance model using 2 bidirectional RNNs on source
- Word level attention re-weighted by corresponding sent level attention

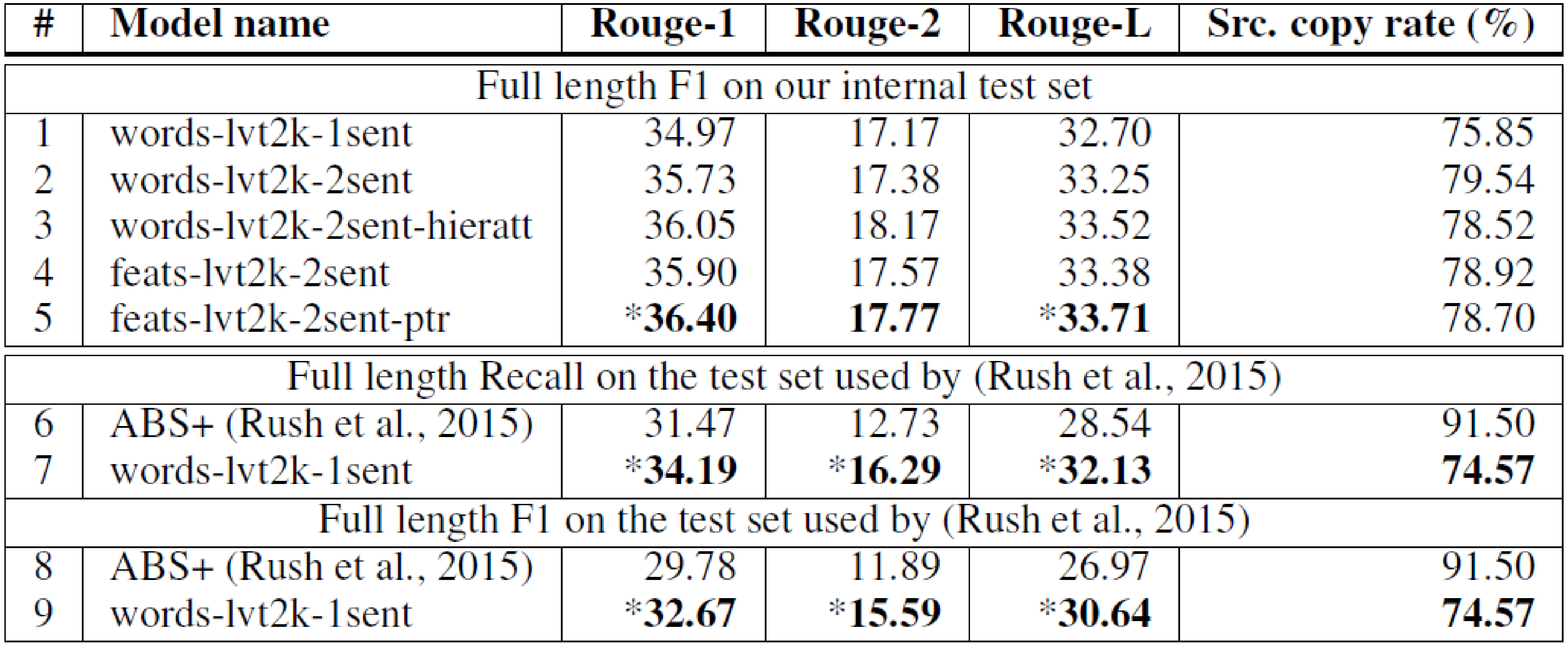
$$P^a(j) \quad = \quad \frac{P^a_w(j)P^a_s(s(j))}{\sum_{k=1}^{N_d} P^a_w(k)P^a_s(s(k))}$$

- *s(j)* is the ID of the sent at *jth* position
- Concatenate positional embeddings to the hid state of sent RNN

Hierarchical Attention Mechanism



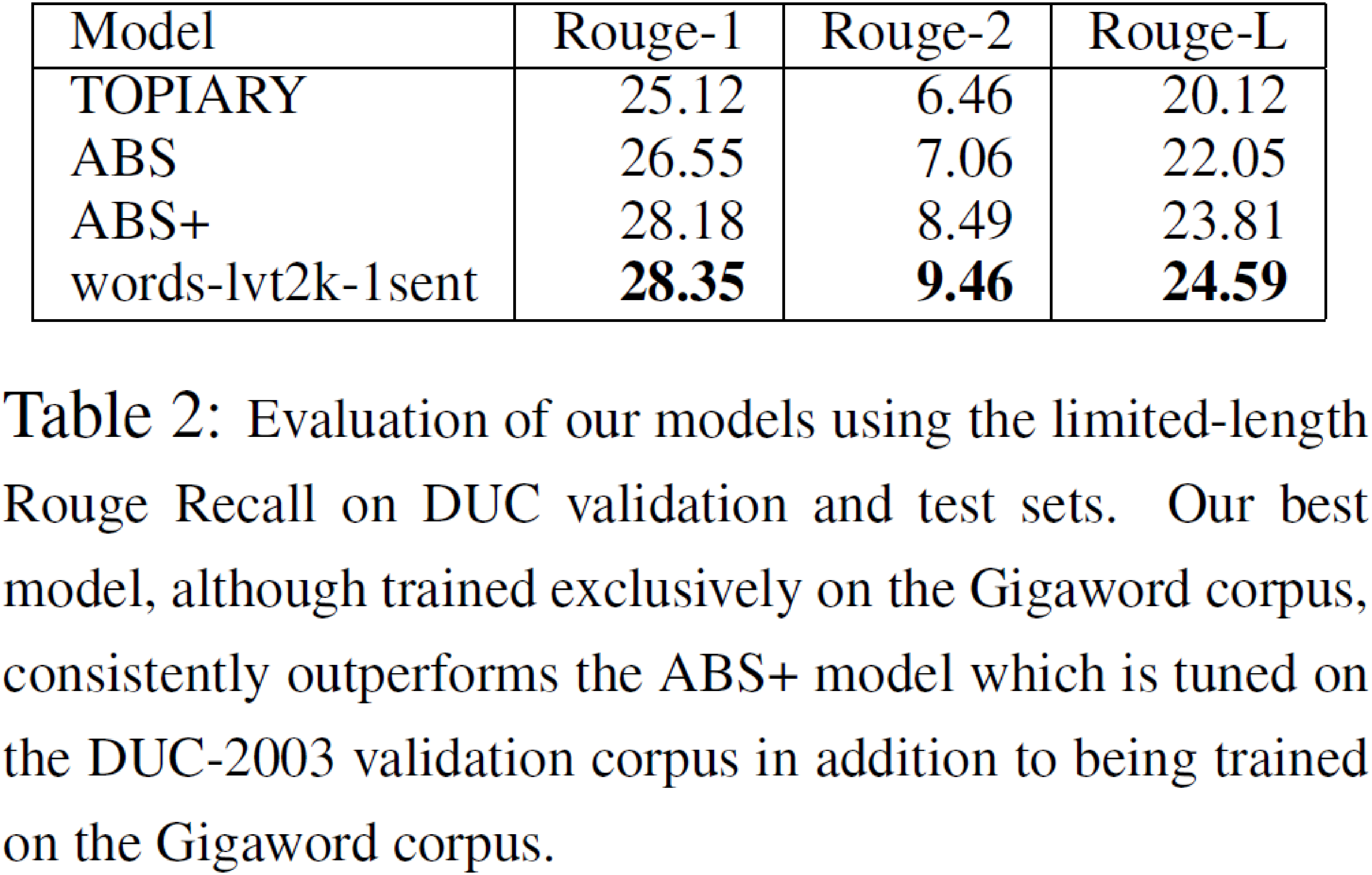
Results on Gigaword corpus



#	Model name	Rouge-1	Rouge-2	Rouge-L	Src. copy rate (%)
Full length F1 on our internal test set					
1	words-lvt2k-1sent	34.97	17.17	32.70	75.85
2	words-lvt2k-2sent	35.73	17.38	33.25	79.54
3	words-lvt2k-2sent-hieratt	36.05	18.17	33.52	78.52
4	feats-lvt2k-2sent	35.90	17.57	33.38	78.92
5	feats-lvt2k-2sent-ptr	*__36.40__	**17.77**	*__33.71__	78.70
Full length Recall on the test set used by (Rush et al., 2015)					
6	ABS+ (Rush et al., 2015)	31.47	12.73	28.54	91.50
7	words-lvt2k-1sent	*__34.19__	*__16.29__	*__32.13__	**74.57**
Full length F1 on the test set used by (Rush et al., 2015)					
8	ABS+ (Rush et al., 2015)	29.78	11.89	26.97	91.50
9	words-lvt2k-1sent	*__32.67__	*__15.59__	*__30.64__	**74.57**

Table 1: Performance comparison of various models. '*' indicates statistical significance of the corresponding model with respect to the baseline model on its dataset as given by the 95% confidence interval in the official Rouge script. We report statistical significance only for the best performing models. 'src. copy rate' for the reference data on our validation sample is 45%. Please refer to Section 4 for explanation of notation.

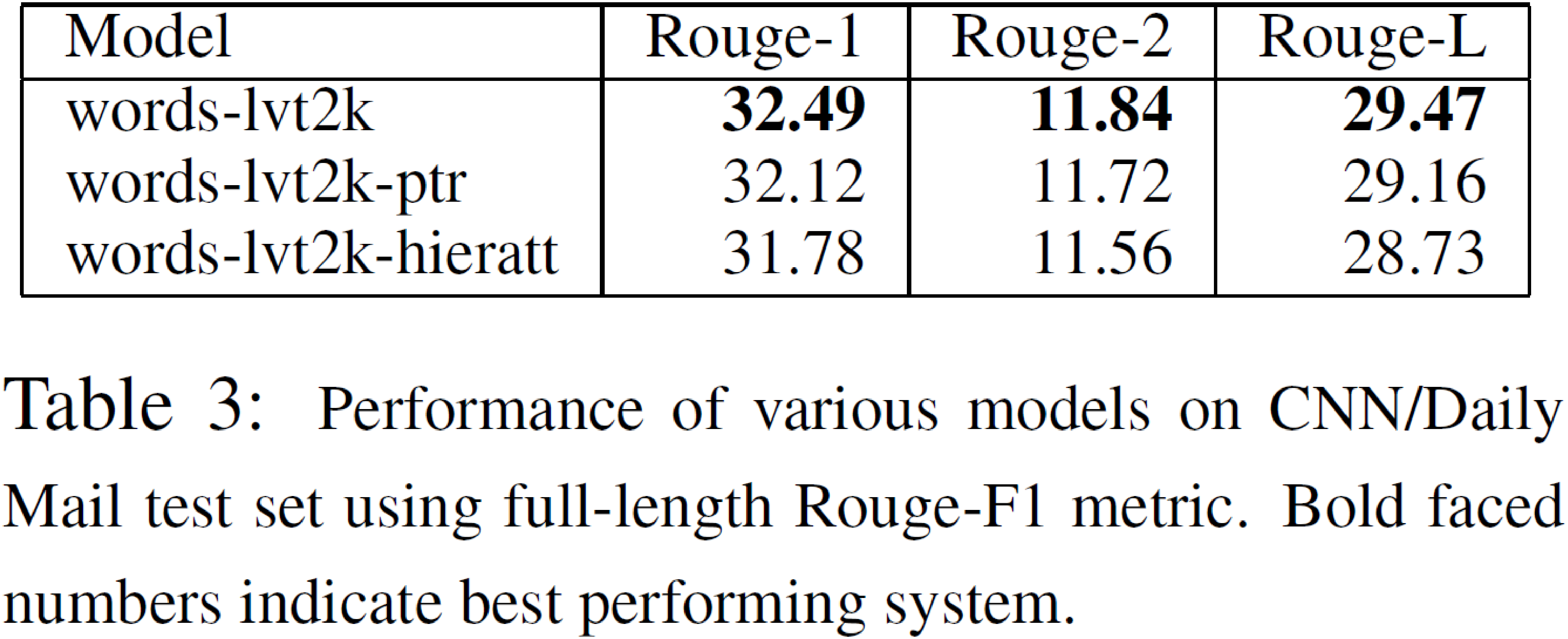
Results on DUC



Model	Rouge-1	Rouge-2	Rouge-L
TOPIARY	25.12	6.46	20.12
ABS	26.55	7.06	22.05
ABS+	28.18	8.49	23.81
words-lvt2k-1sent	**28.35**	**9.46**	**24.59**

Table 2: Evaluation of our models using the limited-length Rouge Recall on DUC validation and test sets. Our best model, although trained exclusively on the Gigaword corpus, consistently outperforms the ABS+ model which is tuned on the DUC-2003 validation corpus in addition to being trained on the Gigaword corpus.

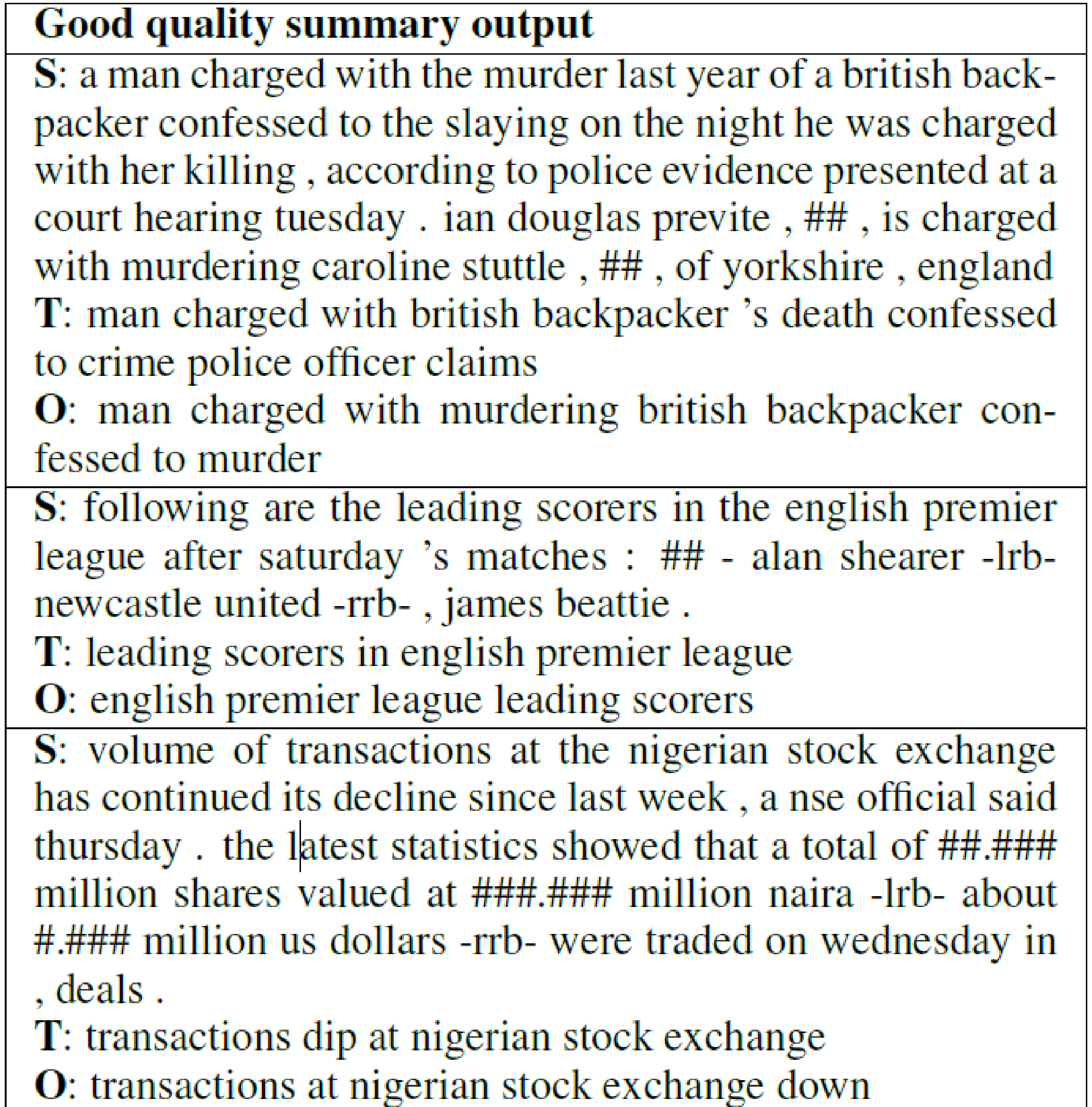
Results on CNN/DM Corpus



Model	Rouge-1	Rouge-2	Rouge-L
words-lvt2k	**32.49**	**11.84**	**29.47**
words-lvt2k-ptr	32.12	11.72	29.16
words-lvt2k-hieratt	31.78	11.56	28.73

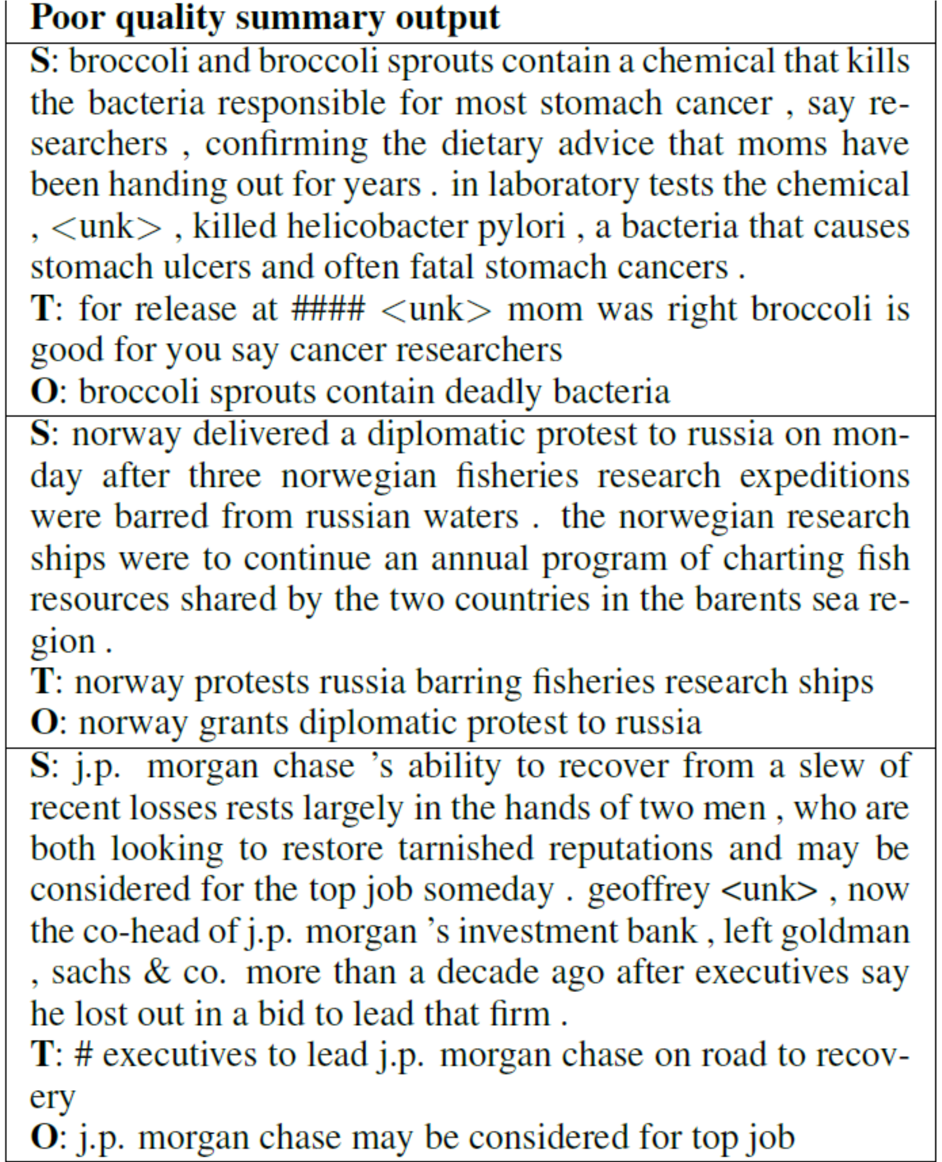
Table 3: Performance of various models on CNN/Daily Mail test set using full-length Rouge-F1 metric. Bold faced numbers indicate best performing system.

Good Quality Summarization



Good quality summary output
S: a man charged with the murder last year of a british backpacker confessed to the slaying on the night he was charged with her killing , according to police evidence presented at a court hearing tuesday . ian douglas previte , ## , is charged with murdering caroline stuttle , ## , of yorkshire , england **T**: man charged with british backpacker 's death confessed to crime police officer claims **O**: man charged with murdering british backpacker confessed to murder
S: following are the leading scorers in the english premier league after saturday 's matches : ## - alan shearer -lrb- newcastle united -rrb- , james beattie . **T**: leading scorers in english premier league **O**: english premier league leading scorers
S: volume of transactions at the nigerian stock exchange has continued its decline since last week , a nse official said thursday . the latest statistics showed that a total of ##.### million shares valued at ###.### million naira -lrb- about #.### million us dollars -rrb- were traded on wednesday in , deals . **T**: transactions dip at nigerian stock exchange **O**: transactions at nigerian stock exchange down

Poor Quality Summarization



Poor quality summary output
S: broccoli and broccoli sprouts contain a chemical that kills the bacteria responsible for most stomach cancer , say researchers , confirming the dietary advice that moms have been handing out for years . in laboratory tests the chemical , <unk> , killed helicobacter pylori , a bacteria that causes stomach ulcers and often fatal stomach cancers . **T**: for release at #### <unk> mom was right broccoli is good for you say cancer researchers **O**: broccoli sprouts contain deadly bacteria
S: norway delivered a diplomatic protest to russia on monday after three norwegian fisheries research expeditions were barred from russian waters . the norwegian research ships were to continue an annual program of charting fish resources shared by the two countries in the barents sea region . **T**: norway protests russia barring fisheries research ships **O**: norway grants diplomatic protest to russia
S: j.p. morgan chase 's ability to recover from a slew of recent losses rests largely in the hands of two men , who are both looking to restore tarnished reputations and may be considered for the top job someday . geoffrey <unk> , now the co-head of j.p. morgan 's investment bank , left goldman , sachs & co. more than a decade ago after executives say he lost out in a bid to lead that firm . **T**: # executives to lead j.p. morgan chase on road to recovery **O**: j.p. morgan chase may be considered for top job

References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. CoRR,abs/1409.0473.
- Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2014. On using very large target vocabulary for neural machine translation. CoRR, abs/1412.2007.