Classify of Select: Neural Architectures for Extractive Document Summarization (Ramesh Nallapati, Bowen Zhou, Mingbo Ma)

Presented by Jongjin Lee.

Seoul National University

ga0408@snu.ac.kr

April 26, 2018

Architecture for extractive summarization of documents

- Two novel and contrasting RNN baseed architectures for extractive summarization of documents.
  - Classifer Architecture
  - Selector Architecture
- Two models imitate two human's strategies for extracting salient sentences in a document
- Deduce the conditions under which one architecture is superior to the other based on experimental evidence

Shared Building Blocks : Bidirectional GRU

- Bidirectional Gated Recurrent Unit(GRU)
  - Similiar to LSTM
  - Two gates(update gate(z), reset gate(r))
- GRU

$$z_t = \sigma(x_t U^z + s_{t-1} W^z),\ h_t = tanh(x_t U^h + (s_{t-1} \circ r_t) W^h)$$
$$r_t = \sigma(x_t U^r + s_{t-1} W^z),\ s_t = (1 - z) \circ h_t + z_t \circ s_{t-1}$$

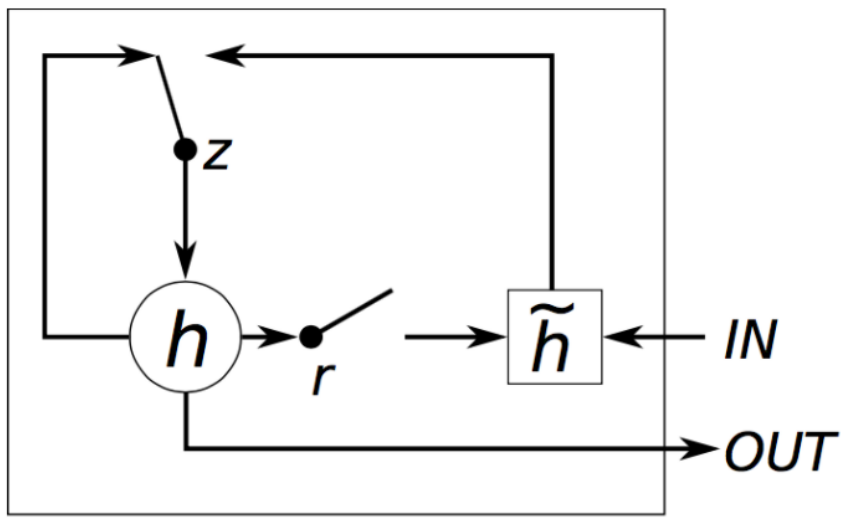
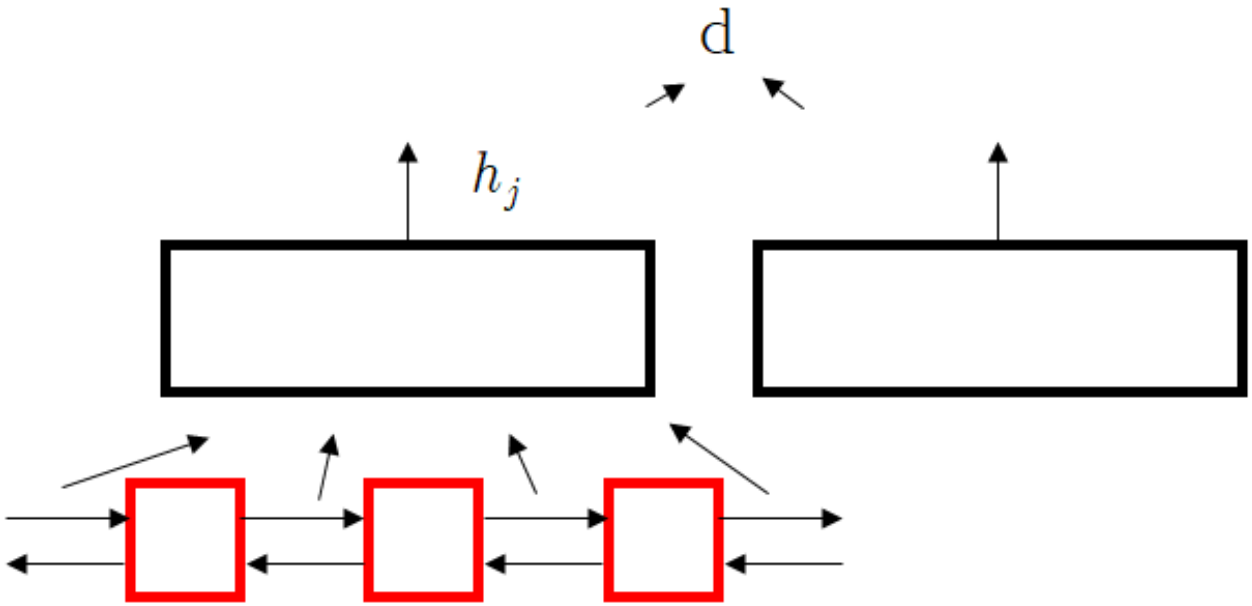


Figure: GRU

Shared Building Blocks

In this work, we propose two neural architectures for extractive summarization. Our proposed models under these architectures are not only very interpretable, but also achieve state-of-the-art performance on two different data sets. We also empirically compare our two frameworks and suggest conditions under which each of them can deliver optimal performance.



Shared Building Blocks

- Both architectures begin with word-level bidirectional GRU run independently over each sentence in the document.
  - The average pooling of the concatenated hidden states of this bi-GRU is used as an input to another bi-GRU for sentences
- The concatenated hidden states 'h' from the forward and backward layers of this second GRU are used as sentence representaion
- The average pooling of the sentence representations as the document representation 'd'
- Dynamic summary representation is 's' whose estimation is architecture dependent

Shared Building Blocks : Score

- For interpretation, explicitly model abstract features such as salience, novelty and information content.

$$score(h_j, s_j, d, p_j) = \omega_c \sigma(W_c^T h_j) + \omega_s \sigma(cos(h_j, d)) + \omega_p \sigma(W_p^T p_j) - \omega_r \sigma(cos(h_j, s_j)) + b$$

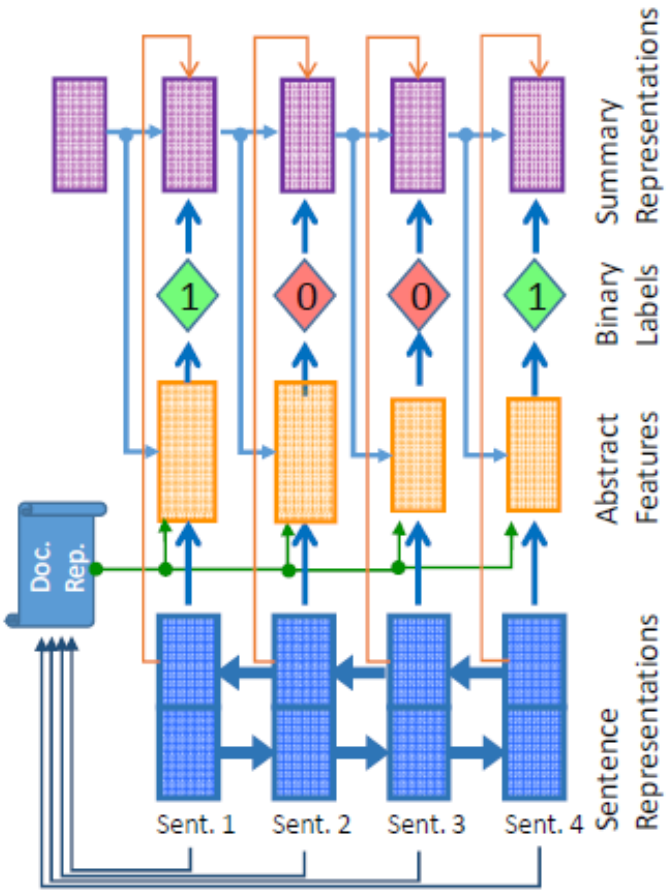
- d: document representation.
- j is index of sentences in document
  - s_j is j-th dynamic summary representation.
  - h_j is j-th sentence representation.
  - p_j is j-th positional embedding of the sentence computed by concatenation of embeddingds to forward and backward position indices of the sentence in the document.
- cos(a,b) is the cosine similiarity between two vector a,b.

Shared Building Blocks : Score

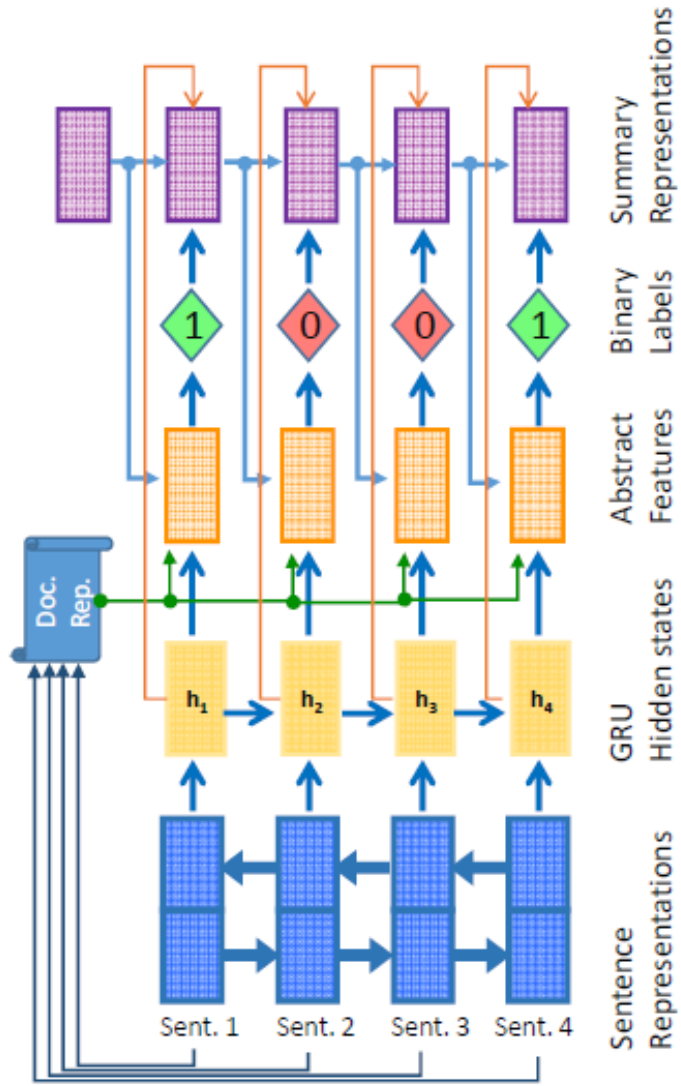
$$score(h_j, s_j, d, p_j) = \omega_c \sigma(W_c^T h_j) + \omega_s \sigma(cos(h_j, d)) + \omega_p \sigma(W_p^T p_j) - \omega_r \sigma(cos(h_j, s_j)) + b$$

- (#content richness) + (#salience w.r.t. document) +(#positional importance) + (# redundancy w.r.t. summary) + (# bias)
- The differences between two architecture are the estimation of dynamic summary representation (s_j) and the cost function

Two models : Classifer Architecture(shallow, deep)



(a) Shallow Classifier Model

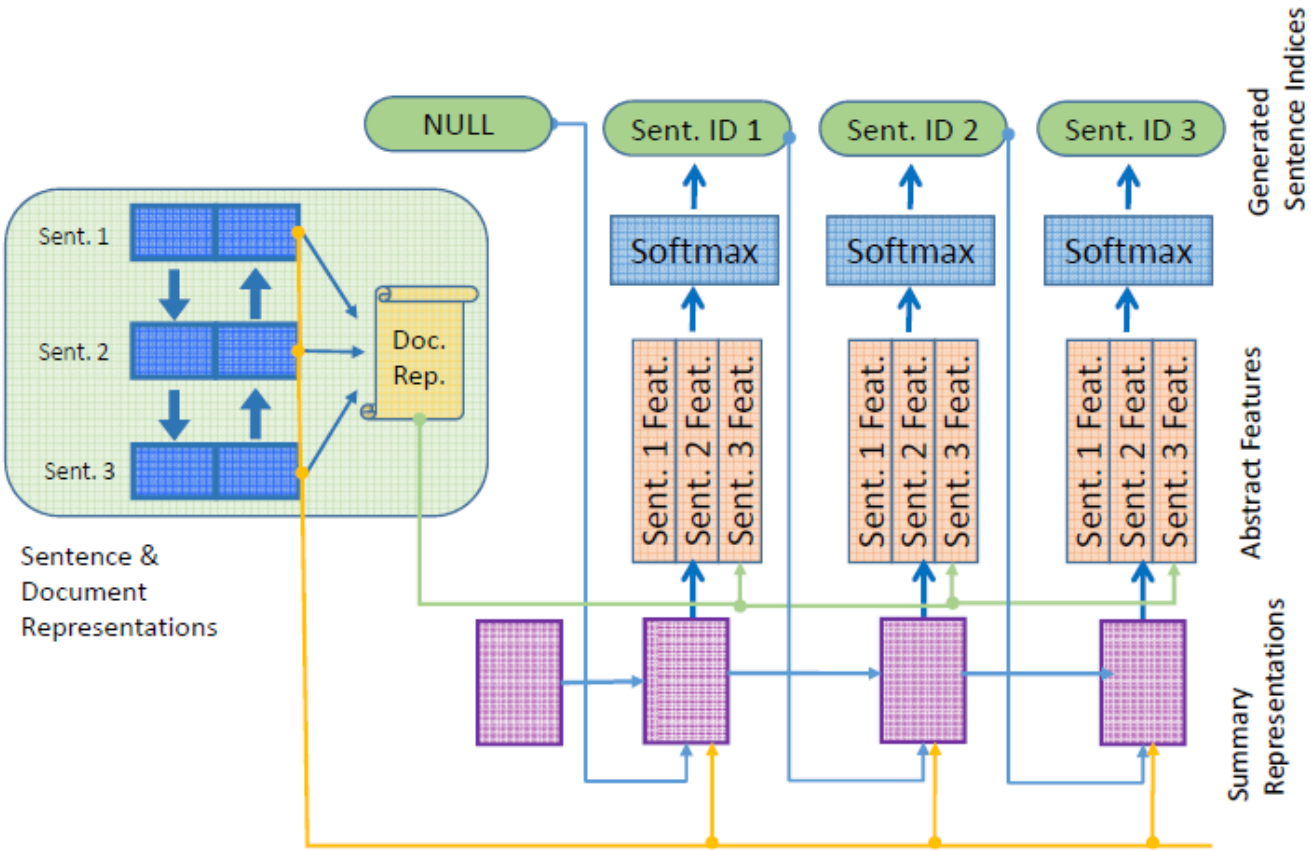


(b) Deep Classifier Model

Two models : Classifer Architecture

- Sequentially visit each sentence in the original document.
- Binary classify the sentence in terms of whether it belongs to the summary
- $P(y_j = 1|h_j, s_j, d, p_j) = \sigma(score(h_j, s_j, d, p_j))$
- $L(W, w, b) =$
  $-\sum_{d=1}^{N}\sum_{j=1}^{N_d}(y_j^d log P(y_j = 1|\bullet) + (1 - y_j) log(1 - P(y_j = 1|\bullet))$
- $s_j = \sum_{i=1}^{j-1} h_i y_i$, #(training time)
  $s_j = \sum_{i=1}^{j-1} h_i P(y_i = 1|h_i, s_i, d, p_i)$ #(test time)
- At deep model, use additional GRU-RNN that takes h_j as input
- (When computing score) Replace h_j as $\hat{h}_j = GRU(h_j)$

Two models : Selector architecture(shallow, deep)



The simple vector representaion for summary representaion in the shallow version is replaced with a gated recurrent unit in the deep version

Two models : Selector Architecture

- Do no make decisions in the sequence of sentence ordering.
- Pick one sentence that maximizes the score at a time.
- $P(\mathrm{I}(j) = k | s_j, h_k, d, p_k) = \frac{exp(score(h_k, s_j, d, p_k))}{\sum_{l \in \{1, \ldots N_d\}} exp(score(h_l, s_j, d, p_l))}$
- $L(W, w, b) = -\sum_{d=1}^{N} \sum_{j=1}^{M_d} log(P(\mathrm{I}(j)^{(d)} | h_{\mathrm{I}(j)^{(d)}}, s_j^d, d_d)$
  (M_d is number of sentences selected in the ground truth of document d)
- $\mathrm{I}(j) = argmax_{k \in \{1, \ldots N_d\}} score(h_k, s_j, d, p_k)$
- $s_j = \sum_{i=1}^{j-1} h_{\mathrm{I}(i)}$ (# for both training and test time)
- $\hat{h_j} = GRU(h_{\mathrm{I}(j-1)})$, use $\hat{h_j}$ as the summary representaion s_j
  $\rightarrow$ GRU can capture a non-linear aggregation of the sentences selected until time step j-1

Experiments and Results

- Evaluation : using differernt variants of the Rough metric computed with respect to the gold abstractive summaries.
  - Rouge-1 : refers to the overlap of 1-gram(each word) between the system and reference summaries.
  - Rouge-2 : refers to the overlap of bigrams between the system and reference summaries.
  - Rouge-L : Longest Common subsequence.
- Experimental Settings
  - 100 dimensional word2vec
  - Limit the vocabulary size to 150K and maximum sentence length to 50 words.
  - Fix model's hidden state size at 200
- Two datasets : Daily Mail corpus, Out-of-Domain DUC 2002 corpus.

Experiments and Results(Daily Mail Corpus)

Model	Recall at 75 bytes			Recall at 275 bytes		
	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L
Lead-3	21.9	7.2	11.6	40.5	14.9	32.6
LReg(500)	18.5	6.9	10.2	N/A	N/A	N/A
Cheng '16	22.7	8.5	12.5	**42.2**	**17.3***	34.8
Shal.-Select	25.6	10.3	14.0	41.3	16.8	34.9
Deep-Select	26.1	10.7	14.4	41.3	15.3	33.5
Shal.-Cls.	26.0	10.5	14.23	42.1	16.8	34.8
Deep-Cls.	**26.2*** ±0.4	**10.7*** ±0.4	**14.4*** ±0.4	**42.2** ±0.2	16.8 ±0.2	**35.0** ±0.2

- Two Models achieve state-of-art performances.
- Classifier architecture is better than Selector architechture.

Experiments and Results(Out-of-Domain DUC 2002 corpus)

	Rouge-1	Rouge-2	Rouge-L
Lead-3	43.6	21.0	40.2
LReg	43.8	20.7	40.3
ILP	45.4	21.3	42.8
TGRAPH	48.1	**24.3***	-
URANK	**48.5***	21.5	-
Cheng *et al* '16	47.4	23.0	**43.5**
Shallow-Selector	44.6	20.0	41.1
Deep-Selector	45.9	21.5	42.4
Shallow-Classifier	45.9	21.5	42.3
Deep-Classifier	46.8 $\pm$0.9	22.6 $\pm$0.9	43.1 $\pm$0.9

- Two Models achieve state-of-art performances.
- Classifier architecture is better than Selector architechture.

Experiments and Results

	Trained on original data			Trained on shuffled sentences		
	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L
Shallow-Selector	41.3	16.8	34.9	**40.6**	**15.6**	**33.0**
Shallow-Classifier	**42.1**	16.8	**35.0**	40.1	15.3	32.9
Deep-Selector	41.3	15.3	33.5	**40.5**	**15.3**	32.5
Deep-Classifier	**42.2**	**16.8**	**35.0**	40.1	15.1	**32.9**

- The original sentence ordering is perhaps advantageous in document summarization since there is a smooth sequential discourse structure in news stories starting.
- If it is true, in scenarios where sentence ordering is less sturcture $\rightarrow$ the selector architecture would be better.

Experiments and Results : Interpretability

Training condition	Salience	Content	Position	Redundancy
Original data	42.75	14.83	-31.09	40.99
Shuffled data	9.69	2.85	0.20	16.08

- Proposed models are not only very interpretable, but also achieve state-of-the-art performence.
- Above table shows the learned importance weights corresponding to various abstract features for deep sentence selector.
- It learns very small weight for the positional features, which is exactly whay one expects.

Experiments and Results : Interpretability

	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

- A representative document along with normalized scores from the deep classifier model

Experiments and Results : Interpretability

	Deep Classifier			Deep Selector		
Features	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L
All	42.43	17.32	34.07	41.55	16.52	32.41
-Salience	42.40	17.27	34.09	40.82	15.99	31.45
-Position	41.78	16.76	33.58	**39.06**	**14.32**	**29.85**
-Content	**41.12**	**15.78**	33.23	40.68	15.83	31.13
-Redundancy	41.67	16.86	**32.93**	41.46	16.50	32.31

- Removing any of the features results in a small loss in performance.
- For the deep classifer, content and redundacy seem to matter the most.
- For the deep selector, dropping positional features hurts the most.