Extractive Summarization with SWAP-NET: Sentences and Words from Alternating Pointer Networks

Aishwarya Jadhav
Indian Institute of Science
Bangalore, India

Vaibhav Rajan
School of Computing
National University of Singapore
Extractive Summarization

Select *salient sentences* from *input document* to create a summary

**INPUT**
Document with sentences $S_1, S_2, \ldots, S_n$

**OUTPUT**
Summary $1 \leq i_k \leq n$

• **Supervised** extractive summarization for *single document* inputs
Our Contribution

A Deep Learning Architecture for training an extractive summarizer: SWAP-NET

- Unlike previous methods, SWAP-NET uses **keywords** for sentence selection
- Predicts both **important words and sentences** in document
- Two-level Encoder-Decoder Attention model
- **Outperform state of the art** extractive summarisers.

**INPUT**

Document with sentences $S_1, S_2, .., S_n$

**OUTPUT**

Summary $1 \leq i_k \leq n$
Extractive Summarization Methods

Recent extractive summarization methods
Recent extractive summarization methods

- **NN (Cheng and Lapata, 2016)**

  - Pre-trained word embeddings
  - Sentence Encoding wrt words in it
  - Sentence encodings wrt other sentences
  - Sentence Label Prediction (with decoder)
Recent extractive summarization methods

- **NN (Cheng and Lapata, 2016)**
  
  - Pre-trained word embeddings
  - Sentence Encoding wrt words in it
  - Sentence encodings wrt other sentences
  - Sentence Label Prediction (with decoder)

- **SummaRuNNer (Nallapati et al., 2017)**
  
  - Pre-trained word embeddings
  - Word Encodings wrt other words
  - Sentence Encoding wrt words in it
  - Sentence Encodings wrt other sentences
  - Document Encoding wrt its sentences

---


Recent extractive summarization methods

- **NN (Cheng and Lapata, 2016)**

  - Pre-trained word embeddings
  - Sentence Encoding wrt words in it
  - Sentence encodings wrt other sentences
  - Sentence Label Prediction (with decoder)

- **SummaRuNNer (Nallapati et al., 2017)**

  - Pre-trained word embeddings
  - Word Encodings wrt other words
  - Sentence Encoding wrt words in it
  - Sentence Encodings wrt other sentences
  - Sentence Label Prediction
  - Document Encoding wrt its sentences

- Both assume **saliency of sentence s depends on salient sentences appearing before s**


Intuition Behind Approach

**Question:** Which sentence should be considered salient (part of summary)?

- Our hypothesis: saliency of a sentence depends on both salient sentences and words appearing before that sentence in the document.
- Similar to graph based models by Wan et al. (2007).
- Along with labelling sentences we also label words to determine their saliency.
- Moreover, saliency of a word depends on previous salient words and sentences.

Intuition Behind Approach

Three types of Interactions:

- Sentence-Sentence Interaction
- Word-Word Interaction
- Sentence-Word Interaction
Intuition: Interaction Between Sentences

A sentence should be salient if it is heavily linked with other salient sentences.
Intuition: Interaction Between Words

A word should be salient if it is heavily linked with other salient words
Intuition: Words and Sentences Interaction

A sentence should be salient if it contains many salient words

A word should be salient if it appears in many salient sentences
Intuition: Words and Sentences Interaction

Generate extractive summary using both important words and sentences

Important Sentences: S3
Important Words: V2, V3
Keyword Extraction and Sentence Extraction

- Sentence to Sentence Interaction as Sentence Extraction
- Word to Word Interaction as Word Extraction
- For discrete sequences, pointer networks have been successfully used to learn how to select positions from an input sequence
- We use two pointer networks one at word-level and another at sentence-level
Pointer Network

Pointer network (Vinyals et al., 2015),

- Encoder-Decoder architecture with Attention
- Attention mechanism is used to select one of the inputs at each decoding step
- Thus, effectively pointing to an input

\[ u^j_i = v^T \tanh(W_e e^j_i + W_d d^j), \quad i \in (1, \ldots, n) \]
\[ p(r_j | r_1, \ldots, r_{j-1}, X) = \text{softmax}(u^j) \]

Three Interactions

Sentence-Level Pointer Network

Sentence - Sentence

Sentence-Word

Word-Word

Word-Level Pointer Network
Three Interactions: SWAP-NET

Sentence-Level Pointer Network

Sentence - Sentence

Sentence-Word

A Mechanism to Combine Word Level Attentions and Sentence Level Attentions

Generate Summary

Word-Word

Word-Level Pointer Network
Questions

Q1: *How can the two attentions be combined?*

Q2: *How can the summaries be generated considering both the attentions?*

![Diagram showing the relationship between Sentence-Word, A Mechanism to Combine Word Level Attentions and Sentence Level Attentions, and Generate Summary questions.](image)
Three Interactions: SWAP-NET

Sentence-Level Pointer Network
Sentence - Sentence

Word-Level Pointer Network
Word-Word

Sentence-Word

Sentence - Sentence

Word-Word

Word-Level Pointer Network
SWAP-NET Architecture: Word-Level Pointer Network

Similar to Pointer Network,
- The word encoder is bi-directional LSTM
- Word-level decoder learns to point to important words
SWAP-NET Architecture: Word-Level Pointer Network

- Purple line: attention vector given as input to each decoding step
- Sum of word encodings weighted by attention probabilities generated in previous step

Probability of word $i$, at decoding step $j$

$$\alpha_{ij}^w = p(t_j = i \mid v_{<j}, Q_j = 0, D)$$

$$a_j = \sum_{i=0}^{n} \alpha_{ij}^w e_i$$

Word Attention

Word Attention Vector
Three Interactions: SWAP-NET

Sentence-Level Pointer Network

Sentence - Sentence

Sentence-Word

Word-Word

Word-Level Pointer Network

Sentence - Sentence

Sentence-Word

Word-Word
SWAP-NET Architecture:
Sentence-Level Hierarchical Pointer Network

Sentence is represented by encoding of last word of that sentence
SWAP-NET Architecture:
Sentence-Level Hierarchical Pointer Network

Attention vectors are sum of sentence encodings weighted by attention probabilities by previous decoding step

\[ A_j = \sum_{k=0}^{N} \alpha_{kj}^s E_k \]

\[ \alpha_{kj}^s = p(T_j = k | v_{<j}, Q_j = 1, D) \]

Probability of sentence k, at decoding step j
Combining Sentence Attention and Word Attention

Q1: How can the two attentions be combined?

A document with three sentences and corresponding words is shown.
Sentence and Word Interactions

Possible Solution:
Step 1: Hold sentence processing. Then group all words and determine their saliency sequentially.
Sentence and Word Interactions

Possible Solution:
Step 2: Using output of step 1, i.e., using keywords, process sentences to determine salient sentences

INCOMPLETE SOLUTION: This method processes sentence depending on words but does not use sentences for processing words.
Sentence and Word Interactions

Solution:
Group each sentence and its words separately and process them sequentially.
Sentence and Word Interactions

Step 1: Hold sentence processing. Determine saliency of words in S1
Sentence and Word Interactions

Step 2:
Using information about saliency of words in S1
- Hold word processing and resume sentence processing.
- Determine saliency of S1
Sentence and Word Interactions

Step3:
Using information about saliency of both S1 and its words
• Hold sentence processing and resume word processing.
• Determine saliency of words in next sentence S2
Step 4:
Using information about saliency of words in S2 and saliency of previous sentence S1
- Hold word processing and resume sentence processing.
- Determine saliency of sentence S2
Sentence and Word Interactions

Solution:
And so on.

This method ensures that saliency of word and sentence is determined from previously predicted both salient sentences and words.
Sentence and Word Interactions

Using previously predicted salient word and sentences

- **Synchronising Decoding Steps**: Decide when to turn off and on word processing and sentence processing to synchronise word and sentence prediction

- **Sharing Attention Vectors**: Determine salient words and sentences
Three Interaction: SWAP-NET

Sentence-Level Pointer Network

Switch Mechanism

Sentence - Sentence

Sentence-Word

Sentence - Sentence

Word-Word

Word-Level Pointer Network
SWAP-NET: Switch Mechanism

Sharing both attention vectors (purple and orange lines) between the two decoder

Synchronising decoding steps of the two decoders by allowing only one decoder output
at a step

Sentence Decoder Hidden State

\[ H_j = \text{LSTM}(H_{j-1}, A_{j-1}, \phi(a_{j-1})) \]

Switch Probability

\[
p(Q_j = 1 | v_{<j}, D) = \sigma(w_Q^T H_{j-1}, A_{j-1}, \phi(h_{j-1}, a_{j-1}))
\]

Word Decoder Hidden State

\[ h_j = \text{LSTM}(h_{j-1}, a_{j-1}, \phi(A_{j-1})) \]
SWAP-NET: Switch Mechanism

Output is selected with maximum of final word and sentence probabilities.
Prediction with SWAP-NET: Encoding

Sentence Encoder

Sentence Encodings

Word Encoder

Word Encodings

Input Document

w1 w2 w3 w4 w5
Prediction with SWAP-NET: Decoding Step 1

Switch has two states,
Q = 0 : word selection and
Q = 1 : sentence selection
Prediction with SWAP-NET: Decoding Step 2

Word Attention

Sentence Attention

Switch

Q=1

Output

S1

W2
Prediction with SWAP-NET: Decoding Step 2
Questions

Q1: How can the two attentions be combined?

Q2: How can the summaries be generated considering both the attentions?
House prices across the UK will rise at a fraction of last year’s frenetic pace, forecasts show.
House prices across the UK will rise at a fraction of last year’s frenetic pace, forecasts show.

Score of Given Sentence = (Sentence Probability) + (Sum of its keyword Probabilities)

\[ \text{Score of Given Sentence} = P_s + \sum_{i=1}^{k} P_i \]

where k is number of keywords in sentence S

Top 3 sentences with maximum scores are chosen as summary.
Extractive Summarization Methods

- **NN (Cheng and Lapata, 2016)**

  - Pre-trained word embeddings → Sentence Encoding wrt words in it → Sentence encodings wrt other sentences → Sentence Label Prediction (with decoder)

- **SummaRuNNer (Nallapati et al., 2017)**

  - Pre-trained word embeddings → Word Encodings wrt other words → Sentence Encoding wrt words in it → Sentence Encodings wrt other sentences → Document Encoding wrt its sentences

- **SWAP-NET**

  - Pre-trained word embeddings → Word Encodings wrt other words → Sentence Encoding wrt words in it → Sentence Encodings wrt other sentences → Sentence Label Prediction (with decoder)
Dataset and Evaluation

• Large Benchmark Dataset CNN/DailyMail News Corpus
  News articles from CNN/DailyMail along with human generated summary (gold summary) for each article

• GroundTruth Binary Labels For Training
  
  | Sentences: Anonymised version of dataset given by (Cheng and Lapata, 2016) |
  | Words: Extract keywords from each gold summary using RAKE |

• Number Labeled Documents

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Training</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>83568</td>
<td>1220</td>
<td>1093</td>
</tr>
<tr>
<td>Dailymail</td>
<td>193986</td>
<td>12147</td>
<td>10346</td>
</tr>
</tbody>
</table>

• Standard Evaluation Metric: Three Variates of Rouge Score
  Comparing generated summaries and gold summaries for matching:

  ROUGE-1 (R1): Unigrams
  ROUGE-2 (R2): Bigrams
  ROUGE-L (RL): Longest Common Subsequences

## Results

Performance on *DailyMail Dataset* using limited length recall of Rouge

<table>
<thead>
<tr>
<th>Models</th>
<th>275 Bytes</th>
<th></th>
<th>75 Bytes</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R1</td>
<td>R2</td>
<td>RL</td>
<td>R1</td>
</tr>
<tr>
<td>Lead-3</td>
<td>40.5</td>
<td>14.9</td>
<td>32.6</td>
<td>21.9</td>
</tr>
<tr>
<td>NN</td>
<td>42.2</td>
<td>17.3</td>
<td>34.8</td>
<td>22.7</td>
</tr>
<tr>
<td>SummaRuNNner-abs</td>
<td>40.4</td>
<td>15.5</td>
<td>32.0</td>
<td>23.8</td>
</tr>
<tr>
<td>SummaRuNNner</td>
<td>42.0</td>
<td>16.9</td>
<td>34.1</td>
<td>26.2</td>
</tr>
<tr>
<td>SWAP-NET</td>
<td><strong>43.6</strong></td>
<td><strong>17.7</strong></td>
<td><strong>35.5</strong></td>
<td><strong>26.4</strong></td>
</tr>
</tbody>
</table>
## Results

Performance on **CNN and Daily-Mail test set** using the full length **Rouge F score**

<table>
<thead>
<tr>
<th>Models</th>
<th>R1</th>
<th>R2</th>
<th>RL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lead-3</td>
<td>39.2</td>
<td>15.7</td>
<td>35.5</td>
</tr>
<tr>
<td>ABS</td>
<td>35.4</td>
<td>13.3</td>
<td>32.6</td>
</tr>
<tr>
<td>SummaRuNNer-abs</td>
<td>37.5</td>
<td>14.5</td>
<td>33.4</td>
</tr>
<tr>
<td>SummaRuNNer</td>
<td>39.6</td>
<td>16.2</td>
<td>35.3</td>
</tr>
<tr>
<td>SWAP-NET</td>
<td><strong>41.6</strong></td>
<td><strong>18.3</strong></td>
<td><strong>37.7</strong></td>
</tr>
</tbody>
</table>
Meet the four immigrant students each accepted to ALL EIGHT Ivy League schools who want to pay back their parents who moved to the U.S. to give them a better future.

Their parents came to the U.S. for opportunities and now these four teens have them in abundance.

The high-achieving high schoolers have each been accepted to all eight Ivy League schools: Brown University, Columbia University, Cornell University, Dartmouth College, Harvard University, University of Pennsylvania, Princeton University, and Yale University.

And as well as the Ivy League colleges, each of them has also been accepted to other top schools.

While they all grew up in different cities, the students are the offspring of immigrant parents who moved to America - from Bulgaria, Somalia, or Nigeria.

And all four - Munira Khalif from Minnesota, Stefan Stoykov from Indiana, Victor Agbafe from North Carolina, and Harold Ekeh from New York - say they have their parents' hard work to thank.

Now they hope to use the opportunities for good - whether it's effecting positive social change, improving education across the world or becoming a neurosurgeon.

The teens have one more thing in common: they do not know which school they're going to pick yet.

The daughter of Somali immigrants who has already received a U.N. award and wants to improve education across the world

Star pupil: Munira Khalif, from St. Paul, Minnesota, says she has always been driven by the thought that her parents, who left Somalia during the civil war, fled to the U.S. so she would have better opportunities.

Munira Khalif, who attends Mounds Park Academy in St. Paul, Minnesota, was shocked when she was accepted by eight Ivy Schools and three others - but her teachers were not.

'She is composed and she is just articulate all the time,' Randy Comfort, an upper school director at the private school, told KMSP. 'She's pretty remarkable.'

The 18-year-old student, who was born and raised in Minnesota after her parents fled Somalia during the civil war, she said she was inspired to work hard because of the opportunities her family and the U.S. had given her.

'The thing is, when you come here as an immigrant, you're hoping to have opportunities not only for yourself, but for your kids,' she told the channel. 'And that's always been at the back of my mind.'

As well as achieving top grades, Khalif has immersed herself in other activities both in and out of school - particularly those aimed at doing good.

She was one of nine youngsters in the world to receive the UN Special Envoy for Global Education's Youth Courage Award for her education activism, which she started when she was just 13.

Gold Summary

Munira_Khalif from Minnesota, Stefan_Stoykov from Indiana, Victor_Agbafe from North_Carolina, and Harold_Ekeh from New_York got multiple offers.

All have immigrant parents - from Somalia, Bulgaria, or Nigeria - and say they have their parents' hard work to thank for their successes.

They hope to use the opportunities for good, from improving education across the world to becoming neurosurgeons.
While they all grew up in different cities, the students are the offspring of immigrant parents who moved to America - from Bulgaria, Somalia or Nigeria. And all four - Munira_Khalif from Minnesota, Stefan_Stoykov from Indiana, Victor_Agbafe from North_Carolina, and Harold_Ekeh from New_York - say they have their parents' hard work to thank. Now they hope to use the opportunities for good - whether it's effecting positive social change, improving education across the world or becoming a neurosurgeon.
While they all grew up in different cities, the *students* are the offspring of *immigrant parents* who moved to America - from *Bulgaria*, *Somalia* or *Nigeria*. And all four - *Munira_Khalif* from *Minnesota*, *Stefan_Stoykov* from *Indiana*, *Victor_Agbafe* from *North_Carolina*, and *Harold_Ekeh* from *New_York* - say they have their *parents' hard work* to thank. *Now* they hope to use the *opportunities* for *good* - whether its effecting *positive social change*, improving education across the *world* or becoming a *neurosurgeon*.
While they all grew up in different cities, the students are the offspring of immigrant parents who moved to America - from Bulgaria, Somalia or Nigeria. And all four - Munira_Khalif from Minnesota, Stefan_Stoykov from Indiana, Victor_Agbafe from North_Carolina, and Harold_Ekeh from New_York - say they have their parents’ hard work to thank. Now they hope to use the opportunities for good - whether it's effecting positive social change, improving education across the world or becoming a neurosurgeon.
Experiments

- **Key word coverage** measures the proportion of key words from those in the gold summary present in the generated summary.

- **Sentences with key words** measures the proportion of sentences containing at least one key word.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Lead-3</th>
<th>SWAP-NET</th>
</tr>
</thead>
<tbody>
<tr>
<td>KW coverage</td>
<td>61.6%</td>
<td>73.8%</td>
</tr>
<tr>
<td>Sentences with KW</td>
<td>92.2%</td>
<td>98%</td>
</tr>
</tbody>
</table>

- **Average pairwise cosine distance** between paragraph vector representations of sentences in summaries to measure semantic redundancy in summaries.

<table>
<thead>
<tr>
<th>Summary Type</th>
<th>Lead-3</th>
<th>SWAP-NET</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold summary</td>
<td>0.81</td>
<td>0.8</td>
</tr>
<tr>
<td>Lead-3</td>
<td>0.553</td>
<td></td>
</tr>
<tr>
<td>SWAP-NET</td>
<td>0.8</td>
<td></td>
</tr>
</tbody>
</table>

SWAP-NET summaries are similar in redundancy to the Gold summary.

Highlights the importance of key words in finding salient sentences for extractive summaries.
Conclusion

- We develop SWAP-NET, a neural sequence-to-sequence model for extractive summarization

- By effective modelling of interactions between sentences and key words, SWAP-NET outperforms state-of-the-art extractive single-document summarizers

- SWAP-NET models these interactions using a new two-level pointer network based architecture with a switching mechanism

- Experiments suggest that modelling sentence-keyword interaction has the desirable property of less semantic redundancy in summaries generated by SWAP-NET

An implementation of SWAP-NET and generated summaries from the test sets are available online: https://github.com/aishj10/swap-net