Stanford University

Generating Arabic News Headlines

Omar Alhadlaq

Problem

Text summarization is a common natural language task of producing a coherent, informative, accurate and brief summary of a document. This task is a cornerstone of NLP, since the ability to accurately identify the most important ideas in a document, as well as producing text that can communicate these ideas is a great step towards the understanding of natural language. Many of the recent work tackled text summarization as a conditioned language generation or a sequence-to-sequence mapping, much like what is used in neural machine translation. These approaches have shown great potential, however, they tend to inaccurately reproduce factual details, and show an inability to handle out-of-vocabulary (OOV) words. In addition, Arabic imposes a new level of complexity to the task. In specific, two main challenges in Arabic are: the variation in writing forms and the lack of word boundaries.

Dataset

We collected our own dataset for the purpose of this project and this class. The dataset is the first Arabic news articles dataset of this scale. It is composed of 326K article and title pairs collected from 3 different prominent Saudi news websites (see breakdown in table 1) using 3 different web scrappers that we developed.

	Agency	Website	#Articles
1	Alarabiya	www.alarabiya.net	54K
2	Sabq	www.sabq.org	110K
3	Alriyadh	www.alriyadh.com	162K

Table 1: The sources of the collected dataset.

The dataset has a total number of 40 million words composed of 450K unique words. The average title length is 10.46 words, while the average article length is 250.81 words. Most words are rare, which means they are seen less than 5 times in the entire corpus as can be seen in figure 3.

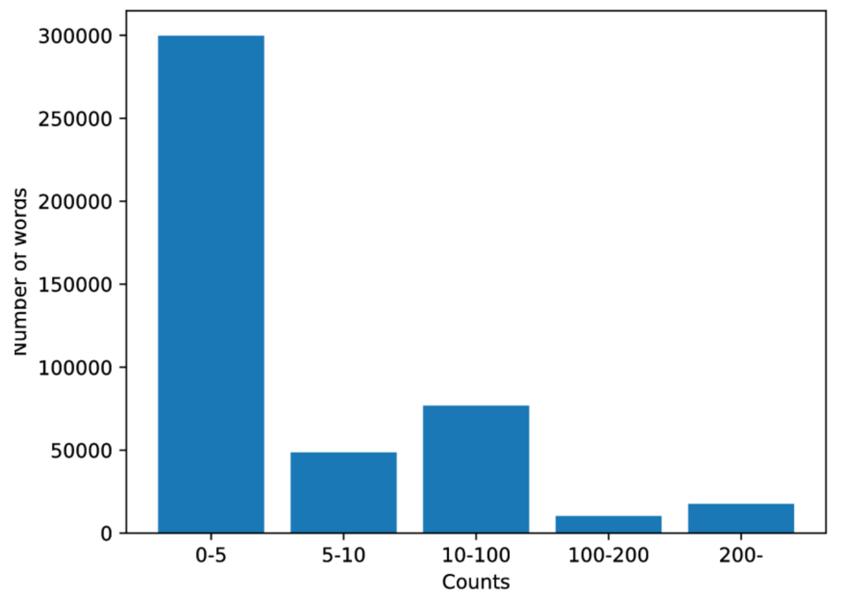

Figure 1: Word frequencies in the Arabic news dataset.

Methods

Model Architectures:

Sequence-to-Sequence (Baseline Model):

Our baseline model uses a 2-layer bidirectional Long Short-Term Memory (LSTM) encoder to map the input article to an article embedding of a fixed dimensionality. Then, it uses another 2-layer LSTM decoder to generate the title sequence from the article embedding.

Sequence-to-Sequence with Attention:

We use the a modified version of the Luong attention such that we score a decoder hidden state and a corresponding encoder state, normalized over all encoder states to get attention scores summing to 1 as in equation (1).

$$a_d(e) = \frac{exp(\text{score}(h_d, h_e))}{\sum_e exp(\text{score}(h_d, h_e))} \quad (1) \qquad \text{score}(h_d, h_e) = \frac{h_d^\top h_e}{\sqrt{n}} \quad (2)$$

The score function can have a variety of forms. The specific form that we use is a dot product between the two vectors, with the addition of a scaling factor inspired by the normalized form of Bahdanau attention as in equation (2).

Sequence-to-Sequence with Copying:

In CopyNet, the decoder switches between generate-mode and copy-mode based on a mixed probabilistic model. In generate-mode, it predicts words from the vocabulary, while in copy-mode, it picks words from the input sequence.

$$p(y_t, g|\cdot) = \begin{cases} \frac{1}{Z} e^{\psi_g(y_t)}, & y_t \in \mathcal{V} \\ 0, & y_t \in \mathcal{X} \cap \mathcal{V} \\ \frac{1}{Z} e^{\psi_g(\text{unk})}, & y_t \notin \mathcal{X} \cup \mathcal{V} \end{cases}$$
$$p(y_t, c|\cdot) = \begin{cases} \frac{1}{Z} \sum_{j:x_j=y_t} e^{\psi_c(x_j)}, & y_t \in \mathcal{V} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Sequence-to-Sequence with Byte-Pair-Encoding:

We process the input by a SetnencePiece model (SPM). This model lets us learn a vocabulary that provides a good compression rate of the text. At worst, the vocabulary can include all characters in the language, and therefore, we are guaranteed not to encounter any out-of-vocabulary (OOV) tokens in the input or the output. The segmentation algorithm used is byte-pair-encoding (BPE). Originally, BPE is a data compression algorithm that was invented in 1994. It works by iteratively replaceing the most frequent pair of bytes in a sequence with a single unused byte. For word segmentation, we merge characters or character sequences instead of merging frequent bytes.

The Transformer:

The encoder is a stack of N = 6 encoding units. All of these are identical in structure but don't share weights. Each unit has two layers: a multi-head self-attention mechanism, and a fully connected feed-forward network. It also includes a residual connection around each of the two layers, followed by layer normalization.

The decoder is also a stack of N = 6 decoding units. Similarly to the encoder unit, the decoding unit has the two multi-head self-attention mechanism, and a fully connected feed-forward network layers. However, it adds a third layer of multi-head attention over the output of the encoder stack. The decoder also includes a residual connection around each of the layers, followed by layer normalization.

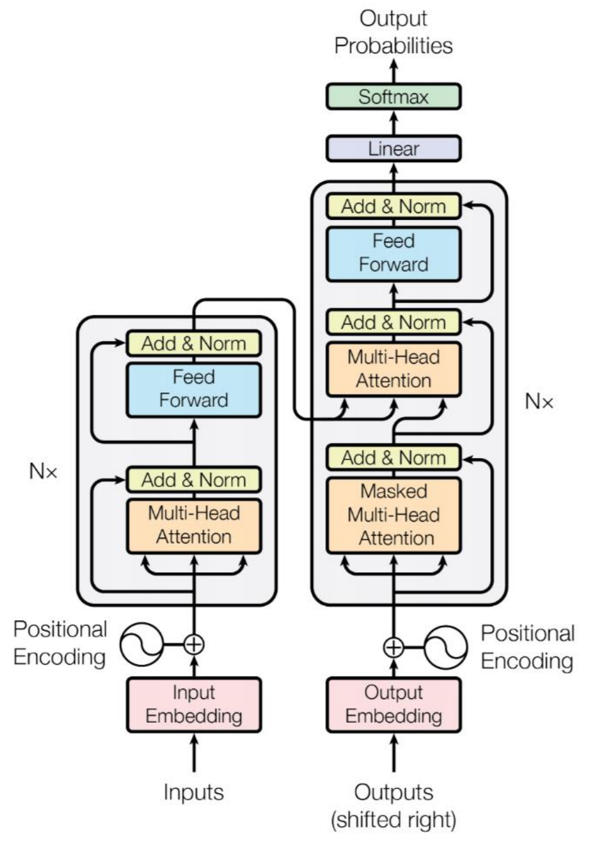

Figure 2: Transformer model.

Evaluation:

The primary evaluation metric used for this task is ROUGE, which measures recall. ROUGE indicates how much of the human produced titles appears in the model generated titles. Moreover, we also use BLEU as a secondary metric, which measures precision. In specific, how much of the model generated titles appears in the human produced titles.

Results

	Baseline	Attention	CopyNet	SPM	Transformer
ROUGE	26.91	33.75	23.80	**37.23**	36.74
BLEU	9.58	14.70	7.54	17.58	**17.60**

Table 2: Model performance on the test dataset.

Analysis

- We have seen that UNK tokens are very common for word-level encoding;
- Most generated titles are related to article which indicates some understanding of the articles;
- Models with attention seem to do a much better job at identifying the key ideas of the article;
- Subword models are able to predict numbers and rare words correctly;
- Generated headlines are straightforward and missing creativity.

Source	Headline
Human	The Custodian of the Two Holy Mosques leaves Japan.
Baseline	The Custodian of the Two Holy Mosques receives the chairman of Japan.
Attention	The Custodian of the Two Holy Mosques leaves Japan after an official visit.
CopyNet	The Custodian of the Two Holy Mosques leaves <unk><unk><unk>.
SPM	The Custodian of the Two Holy Mosques leaves Japan after an official visit.
Transformer	The Custodian of the Two Holy Mosques leaves Japan after an official visit.

Table 3: Sample (1): generated headlines (translated).

Source	Headline
Human	Aramco sets propane price in January to 435 dollars per ton.
Baseline	"Aramco" sets the price price <unk>- euro to 5.33 million dollars.
Attention	Aramco sets <unk> price in January contract to <unk> dollars.
CopyNet	Aramco: <unk><unk><unk><unk><unk><unk>Aramco.
SPM	Aramco sets propane price in January to 435 dollars per ton.
Transformer	Aramco sets propane price in January to 435 dollars per ton.

Table 4: Sample (2): generated headlines (translated).

Source	Headline
Human	More that 5515 citizen has applied to 2962 open jobs in education.
Baseline	Civil Service Ministry: jobs in education for men tonight.
Attention	Civil Service Ministry: <unk> applicatns to jobs in education.
CopyNet	Civil Service Ministry: <unk> to jobs in education <unk>.
SPM	Civil Service Ministry: more than 5515 applicant to jobs in education.
Transformer	Civil Service Ministry: we will announce new jobs in education at the 10th of Ramadan

Table 5: Sample (3): generated headlines (translated).

Conclusion

- We found that attention mechanism is improves the accuracy of factual details;
- Moreover, we found that operating on a subword level helps in dealing with out-of-vocabulary words;
- Additionally, we found that our models aim for direct reproduction of the main ideas in a piece of text. While that is acceptable in regular summarization tasks, it doesn't do as well in news headline generation since these require more creativity to hook the reader.