ATTENTION BASED SEQ2SEQ MODEL FOR ABSTRACTIVE TEXT SUMMARIZATION

May 9, 2020

XIN SUN
University of Heidelberg
Computer Linguistic & IWR

Contents

0.1 Abstract

In this mini project, I propose to build abstractive text summarization system using Recurrent Neural Networks based Encoder-Decoder seq2seq architecture with attention mechanism, and do further experiment on dataset Amazon Review corpus. After that I analyze the summary result with ground truth by ROUGE score and also manually as evaluation.

0.2 Abstractive Text Summarization

Text Summarization (TS) aims at composing a concise version of an original text, retaining its salient information. Since manual TS is a demanding, time expensive and generally laborious task, automatic TS is gaining increasing popularity and therefore constitutes a strong motivation for further research.

Two main approaches to automatic Text Summarization have been reported in the relevant literature: extractive and abstractive. In the former case, those sentences of original text that convey its content are firstly identified and then extracted in order to construct the summary which means Extractive text summarization algorithms are capable of extracting key sentences from a text without modifying any word. In the latter case, new sentences are generated which concatenate the overall meaning of the initial text, rephrasing its content. Abstractive Text Summarization is a more challenging task which resembles human-written summaries, as it may contain rephrased sentences or phrases with new words (i.e. sentences, phrases and words that do not appear in the original text), thereby improving the generated summary in terms of cohesion, readability or redundancy.

A lot of algorithms for both extractive and abstractive text summarization are based on Recurrent Neural Networks(RNN). Furthermore, using RNNs in an Encoder-Decoder manner leads us to the well known Sequence-To-Sequence (Seq2Seq) architecture, which is one of the most used and best-performing approaches in text generation tasks. Most of the current advancements have been performed on very short summaries and documents: a lot of algorithms tend to perform worse when a big document has to be summarized in more than a few words. The majority of state-of-the-art algorithms use pre-trained word embeddings, for a better understanding of the concepts expressed in a text.

0.3 Attention based seq2seq model

0.3.1 Seq2seq model

Sequence-to-sequence (Seq2Seq) models are deep learning models that have achieved a lot of success in tasks like machine translation, text summarization, and image captioning.

And what is the Seq2Seq Model? More specifically, Seq2Seq model is a model that takes a sequence of items (words, letters, time series, etc) as inputs and outputs another sequence of items.

In the case of Neural Machine Translation, the input is a series of words, and the output is the translated series of words.

And the internal seq2seq model is composed of an encoder and a decoder. The encoder captures the context of the input sequence in the form of a hidden state vector and sends it to the decoder, which then produces the output sequence. Since the task is sequence based, both the encoder and decoder tend to use some form of RNNs, LSTMs, GRUs, etc. The hidden state vector can be of any size, though in most cases, it's taken as a power of 2 and a large number (256, 512, 1024) which can in some way represent the complexity of the complete sequence as well as the domain.

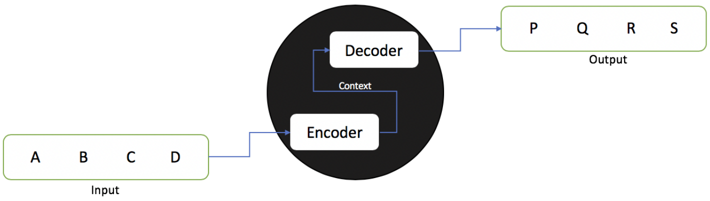

Figure 1: The Seq2seq model

RNNs by design, take two inputs, the current example they see, and a representation of the previous input. Thus, the output at time step t depends on the current input as well as the input at time t-1. This is the reason they perform better when posed with sequence related tasks. The sequential information is preserved in a hidden state of the network and used in the next instance. The Encoder, consisting of RNNs, takes the sequence as an input and generates a final embedding at the end of the sequence. This is then sent to the Decoder, which then uses it to predict a sequence, and after every successive prediction, it uses the previous hidden state to predict the next instance of the sequence.

0.3.2 Encoder-Decoder Architecture

The Seq2Seq framework relies on the encoder-decoder paradigm. The encoder encodes the input sequence, while the decoder produces the target sequence.

The encoder-decoder architecture is a neural network design pattern. As shown in Figure, the architecture is partitioned into two parts, the encoder and the decoder. The encoder's role is to encode the inputs into state, which often contains several tensors. Then the state is passed into the decoder to generate the outputs.

0.3.2.1 Encoder

Our input sequence is "how are you". Each word from the input sequence is associated to a vector wR (via a lookup table). In our case, we have 3 words, thus our input will be transformed into R. Then, we simply run an LSTM over this sequence of vectors and store the last hidden state outputed by the LSTM: this will be our encoder representation $\boldsymbol{e}$. Let's write the hidden states.

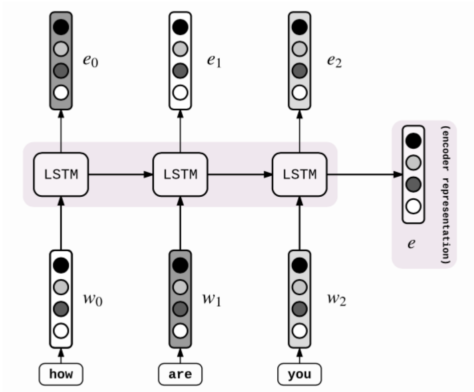

Figure 2: Encoder

0.3.2.2 Decoder

Now that we have a vector that captures the meaning of the input sequence, we'll use it to generate the target sequence word by word. Feed to another LSTM cell. Like the following figure.

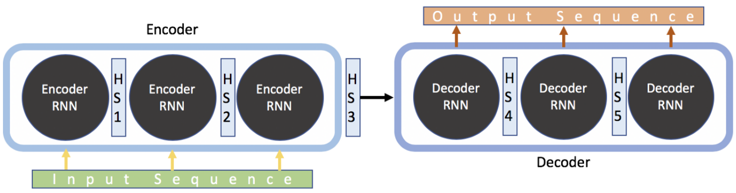

Figure 3: Decoder

0.3.3 Attention mechanism

Drawback of normal encoder-decoder mechanism: The output sequence relies heavily on the context defined by the hidden state in the final output of the encoder, making it challenging for the model to deal with long sentences. In the case of long sequences, there is a high probability that the initial context has been lost by the end of the sequence.
Solution: Bahdanau et al., 2014 and Luong et al., 2015 papers introduced and a technique called

"Attention" which allows the model to focus on different parts of the input sequence at every stage of the output sequence allowing the context to be preserved from beginning to end.

Therefore we can realize that Attention is the mechanism that forces the model to learn to focus (=to attend) on specific parts of the input sequence when decoding, instead of relying only on the hidden vector of the decoder's RNNs. So here is the new representation.

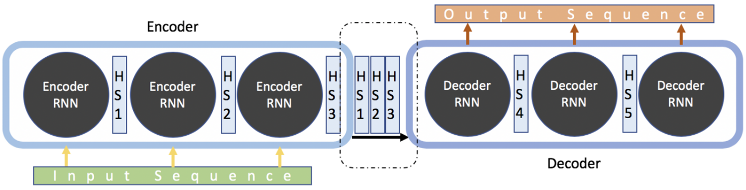

Figure 4: Attention

At every step, the context vector is a weighted sum of the input hidden states as given in figure 5. The generated context vector is combined with the hidden state vector by concatenation and this new attention hidden vector is used for predicting the output at that time instance. Note that this attention vector is generated for every time instance in the output sequence.

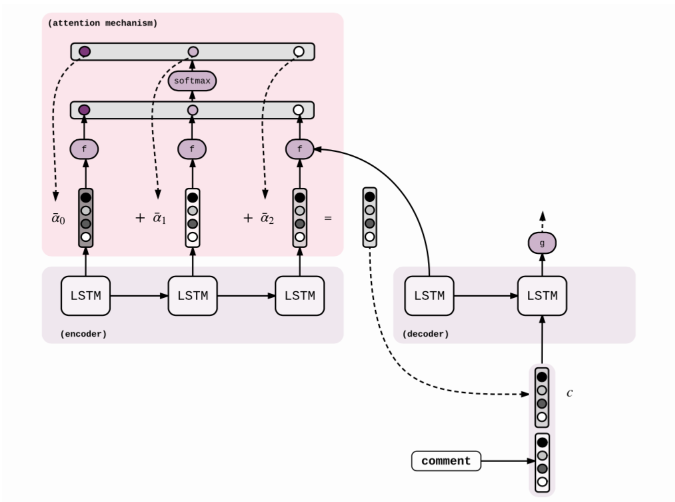

Figure 5: Attention Details

0.4 Experiment

In this experiment section, I propose to build abstractive summarization system by seq2seq model and make comparison and analysis on model with or without attention mechanism. The abstractive summarization system will take each review comment text as input of seq2seq encoder and output a simplified summary with fixed length as the generated summary sequence of seq2seq decoder. Therefore I mainly did three things below:

1. Build a very basic seq2seq based abstractive summarization system by initially word to index word2vec embeddings;

2. Add attention mechanism to build an attention based seq2seq model for abstractive summarization by initially word to index word2vec embeddings;

3. Used pre-trained word embeddings on Wiki corpus to build an attention based seq2seq model for abstractive summarization.

0.4.1 Dataset

I tried to build this seq2seq model that can create relevant summaries for reviews written about fine foods sold on Amazon. This dataset contains above 40,0000 reviews of amazon products, and is hosted on Kaggle which size is over 300MB. The following figure 6 shows us how this dataset looks like.

	Summary	Text	Summary_len	Text_len
0	Good Quality Dog Food	I have bought several of the Vitality canned d...	4	48
1	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...	3	31
2	"Delight" says it all	This is a confection that has been around a fe...	4	94
3	Cough Medicine	If you are looking for the secret ingredient i...	2	41
4	Great taffy	Great taffy at a great price. There was a wid...	2	27

Figure 6: The overview of Amazon Reviews dataset

For each review comment, the text has average 159 words and the length of summary ground truth is average 12. We can see the following dataset words count distribution below.

```
Text length 90 percentile: 159.0
Summary length 99 percentile: 12.0
```

0.4.2 Initial word2Vec Embeddings

In the first two parts experiments: the very basic seq2seq model with and without attention mechanism for abstractive text summarization. I used a very basic word to index embeddings as initial embedding approach. While in the third part of this experiment, I used a pre-trained embeddings matrix from fasttext as initial embedding approach which was trained on Wiki English Corpus with embedding matrix dimension 300. After that, I used single layer GRU as internal feature extractor in both seq2seq encoder and decoder, which represent review comment

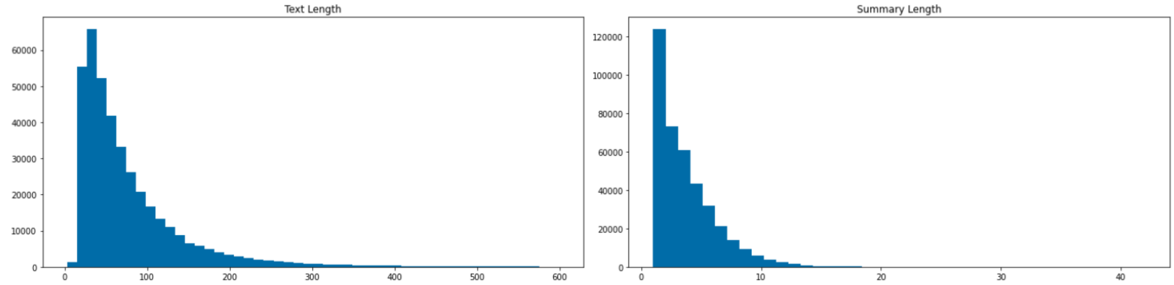

Figure 7: The words count distribution of Amazon Reviews dataset

in encoder or generate the summaries in decoder.
And these are two different approaches I did to compare final generated summary quality in the end.

0.4.3 Model construction

Now we come to the part of constructing seq2seq model. I will make a simplified explanation how I build this seq2seq model with / without attention mechanism for abstractive summarization. More detials please check the source code in .ipynb file on Github.

0.4.3.1 The basic seq2seq model without attention

In the first part of this mini project experiment, I plan to build a very basic seq2seq model by encoder-decoder architecture for abstractive text summarization. And we can know how basic encoder decoder works by following figure 8.

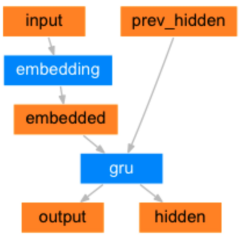

Figure 8: How basic encoder / decoder works

0.4.3.2 The basic seq2seq model with attention

In this part of experiment, I built a seq2seq model for abstractive text summarization with attention mechanism. And we can know this attention based encoder decoder works by following figure 9.

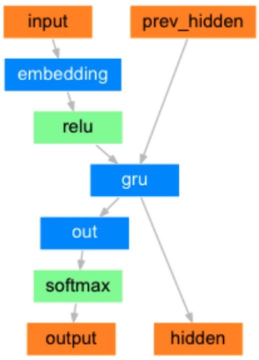

Figure 9: How attention based encoder / decoder works

Figure 10: How attention core looks like

0.4.3.3 The seq2seq model with attention using pre-trained word embeddings

In this part, I used a pre-trained embeddings matrix from fasttext as initial embedding approach instead of basic word to index approach which was trained on Wiki English Corpus with totally 200,0000 words in dictionary and embedding matrix dimension is 300.

0.4.3.4 Training / Inference

Duaring training and inference period, the encoder architecture keeps fix, while decoder core is different. The input of decoder RNNs in training phrase is: previous hidden states, contextual vectors and current input word of ground truth summaries. While the input of decoder RNNs in inference phrase is: previous hidden states, contextual vectors and predicted output word from last step during inference, because we do not have ground truth summaries and can not feed RNNs ground truth summaries in each inference steps.

And the more detailed hyper-parameters setting is: Epochs is 30; Batch Size is 64: single layer GRU hidden state size is 100; Embedding Size of word2index way is 100 while with pre-trained embeddings is 300; Encoding and decoding embedding size is 100; Learning Rate is 0.001.

0.4.4 Result

I analyzed the generated summary result with ground truth by ROUGE score (ROUGE 1, 2, L) and also evaluated them manually to see whether the generated summary quality is acceptable or not. We can see the comparison result below.

From the comparison of ROUGE score and generated summaries above, we could find that:

1. The quality of all 3 different models are not quite acceptable, but in some texts, the overall meaning of generated summary is comparable to the ground truth summary. Maybe this

summary_comparison

idx	text_sent	summ_truth	summ_pred_no_att	summ_pred_with_att	summ_pred_emb
0	BOS fastest shipping around and zsweet is the best artifical sweetener there is i use it in my sweet tea and im a s	zsweet the best	a great way to start the day !	great product	the best !
200	BOS my daughter who does not like many cereal varieties loves this so much . she has a big bowl of mike and th	kids love !	my new favorite	honey is great !	my dog loves it !
400	BOS i got some of these and some of the less intense ginger candy . i like this one better than the less intense . s	really , really good	ginger altoids	great candy	a little bit of a little more than the original
600	BOS i love buying items and groceries on amazon . the fact that you receive a discount for signing up for a subsc	best popcorn in the world !	best tasting mrp out there !	great stuff !	love this stuff !
800	BOS the intention to create recyclable plastic is great , but the construction on these is unreliable . the bottoms fa	good concept , poor construction	new recipe is terrible	poor representation	not what i expected
1000	BOS gave this as a gift to a french friend . she felt it was not like the mg 's she was used to in france . something	UNK UNK	a little bit of my first	not what i expected	not sure how to identify
1200	BOS this is a poor excuse for lychee product . it is water with sugar and a very faint lychee flavor . you can do be	did i pay for sugar and water ?	artificial flavor	horrible	not worth the money
1400	BOS i have tried several of the flavors in this oatmeal line , and this chocolate oatmeal and the blueberry muffin fl	delicious !	deeply moving culinary grade b maple syrup	the best of the best	great flavor , but not as good as the other brands
1600	BOS lovely plant , no brown spots or dead leaves : -) the only thing that bothers me is that it is all growing to one	beautiful healthy plant	a little silver dusting is deceiving	remember !	not what i expected
1800	BOS this is great coffee . i generally like weaker coffee but this can also be made quite strong without being bitte	great coffee	smooth , bold , smooth and delicious	great coffee	this is a good coffee , but not great
2000	BOS perky 's nutty , crunch cereal is quite good and not sweet . great for gluten free diets . add some fruit for kic	perky 's is good	best cereal ever !	great cereal	a good stuff
2200	BOS the crackers , fruit and nut mix , the dry roasted edamame and the dark chocolate were ok. the salmon was	the salmon is not what i hoped - everything else is not bad	a little bit of a treat	not bad , but not great	not bad
2400	BOS as a mom of a child with dairy allergies , it 's very hard to travel without this product . the small boxes are ve	worth it	a little more like a charm	great product , great packaging	best milk
2600	BOS if you 're hungry for alfredo and do n't have time to make one of those knorr packet things (or cook !) , this	pretty good	good but not great	it 's ok	a good idea
2800	BOS i have been a lipton loose tea junkie for 50 years and as lipton loose tea is getting hard to find locally i order	lipton yellow label loose tea	best tea ever	not dilmah	not worth the money
3000	BOS i 'm a huge fan of earnest eats with my favorites being the UNK butter and the cranberry-orange . i tried the	yummy	my new favorite !	great for trekking traveling	love this stuff !
3200	BOS my dog is pretty picky - she wo n't touch the c.e.t . enzymatic chews , but she gobbles these up . she also l	picky eater loves these	my dog will eat	dog loves them	my dog loves these
3400	BOS ordered 3 jars of this stuff . boy was i ever disappointed in a product . did n't even taste like beef , but totally	bad news review	not the best	not the best	do n't buy this product
3600	BOS as a business traveler , i really love this product . i wondered , at first , if the tsa would allow it through secu	great for travel	great product	great product !	great stuff !
3800	BOS when overused , they produce a gasoline-like taste . i bought these for a `` decorate your own cupcake '' pa	looks pretty but terrible taste	great for a quick and easy to make	not as good as it claims !	fun !
4000	BOS i ordered these 4 days ago and now i only have 1 bag left . it 's low in sugar , contains many healthy ingredi	best granola on this planet	best tasting mrp out there	great taste , but availability	best thing ever !
4200	BOS these strawberries and cream gummies are the best . they are so tasty . when i got them i thought they wou	the best !	yummy !	these are the best	yummy !
4400	BOS i am lucky my food store no longer carried the us produced tangerine juice i was purchasing , because it ca	the best tangerine juice .	best tasting drink ever	coconut juice	the best !
4600	BOS i am used to the powdered packets that crystal light offers , and i think i could go either way . the nice thing	compared to crystal light	best tasting , healthiest	great substitute for soda	a little weak
4800	BOS as a confirmed tea drinker , i 'm discriminating about the tea i drink . had the luck to find a really good tea in	great cup of tea !	simply the best !	black tea	best tea ever !
5000	BOS do some research on tumeric powder . the stuff is full of good things that are great for you , and it has tons	multiple uses	a little bit of a real thing	great product	great bloody mary mix
5200	BOS i was very pleased with my purchase of white chia seeds from superior nut company . normally , i grind the	great value ! !	best instant i 've found	great product	great product !
5400	BOS i absolutely love pop chips . cheddar is my favorite , but i thought i 'd try these cause i like parmesan too . ju	great chips !	best chip ever !	popchips	my favorite pop chips
5600	BOS < a href= '' http : UNK '' > frontier soups homemade in minutes oregon lakes wild rice & mushroom soup , 4	i did not like the item and could not return	`` instant ''	misleading	not the best
5800	BOS this thing is absolutely amazing ! i had to go through the company website to get the right size for the jars w	greatest thing since sliced bread !	amazing !	awesome !	amazing !
6000	BOS i love this tea . it is , hands down , my favorite . and the ingredients are very `` clean '' and it 's low calorie --	best tea on earth ! ! !	wow ! ! ! ! ! ! ! ! ! !	love this tea !	my favorite tea !
6200	BOS it 's been really hot (over 100 degrees) so i have keep myself hydrated by drinking lots of water . but i hate	not bad but would not have ordered had i known about its ingredi	i like it !	it 's alright	it 's hot
6400	BOS i have a waring belgian waffle maker and was looking for a good batter to use . i tried this based on reviews	excellent choice for belgian waffles	best instant pancake mix	great mix	great stuff !
6600	BOS i ate these as a kid , and that was a long time ago , and loved them . as i got older and forgetful this great sn	garlic sensation	a great alternative to traditional popsicles/ice	great snack	a great snack
6800	BOS this oil was one of 3 or 4 recommended in mario batali 's book , `` simple italian food '' from the late 90 's . t	great daily extra virgin olive oil	simply the best	great oil !	great for stir-frys and seborrheic dermatitis)
7000	BOS i have been a regular of celestial seasonings for many years , and had no idea there was such a great value	best bargain in the best tea going	best tasting tea ever	great for upset stomach	great tea , great price

Figure 11: Generated summaries comparison

ROUGE SCORE

	Seq2seq without attention	Seq2seq with attention	Seq2seq with Pre-trained embedding
ROUGE 1 F1	0.095	0.120	0.125
ROUGE 1 Precision	0.112	0.143	0.151
ROUGE 1 Recall	0.097	0.119	0.130
ROUGE L F1	0.097	0.122	0.131
ROUGE L Precision	0.114	0.147	0.156
ROUGE L Recall	0.099	0.121	0.131

Figure 12: ROUGE SCORE

dataset is not quite suitablbe for precise text summarization tasks or maybe the quality of raw datsset is not good enough for summarization tasks.

2. Even most generated summaries quality is not acceptable, but we could say that the quality of seq2seq with attention and pre-trained embeddings is better than seq2seq with attention, which is also better than seq2seq without attention.

3. Even the ROUGE Score is not good, but we could find that the ROUGE score of seq2seq with attention and pre-trained embeddings is higher than seq2seq with attention, which is also higher than seq2seq without attention. Therefore this is consistent with the previous conclusion.

4. Therefore we could say that the performance of seq2seq model with attention and pre-trained embeddings is the best and certainly better than seq2seq with attention, while the very basic seq2seq model without attention mechanism performs worst in our experiment.

References

Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation

Sequence to Sequence Learning with Neural Networks

Bahdanau et al., 2014 and Luong et al., 2015 papers