Automating news summarization with Deep Learning

Maurício Steinert
Orientador: Felipe Rech Meneguzzi
Computer Science
✉ mauricio.steinert@acad.pucrs.br

Motivation

- Summaries allow users to navigate through a large volume of data quickly and search for specific information.
- Manually generating summaries is a time-consuming and error prone task due to writer background on the matter and the influence of personal opinion reflected into the summary.
- This work aims to develop a technique to automate text summarization given as input a news corpus extracted from the Internet.

Automatic Text Summarization

- Automatic text summarization produces a summary, i.e., a short length version of the source text that contains their most relevant content.
- Text summarization is a challenging task that must deal with issues such as redundancy and the temporal dimension [1].
- Text summarization techniques are broadly classified as:
  - extractive summarization, where sentences that better capture the meaning of the whole text is selected as summary; and
  - abstractive summarization, where the summary is generated from scratch, without reusing existing sentences in the source text.
- The first automatic text summarization attempt dates from 1958 [2] motivated by the growing number of academic articles available at that time.
- Researchers developed different techniques over the years, from statistical models to complex neural network architectures.

Text Summarization Techniques

We developed three text summarization techniques:

- Vector Offset Summarization technique
  - An unsupervised learning technique that identify which sentence in a text best represent the overall content.
  - Computes the word embedding offset between the whole text and each sentence, choosing as summary the sentence with lower offset (Figure 1).

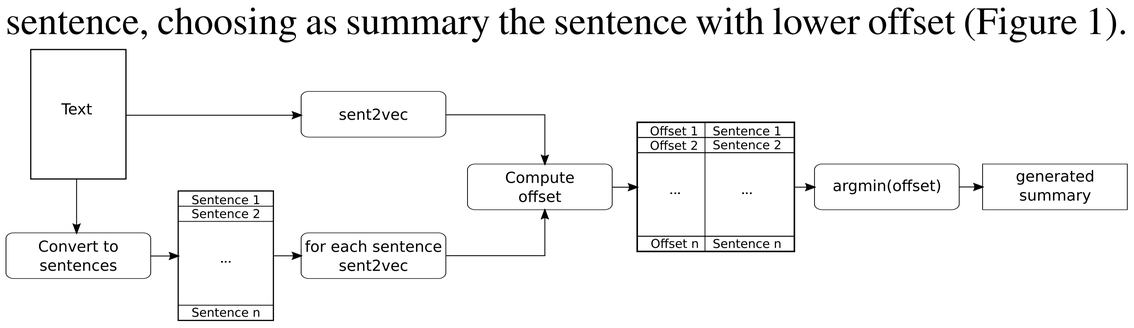

Figure 1: Vector Offset Summarization workflow.

- Feed-Forward Neural Network Summarization Technique
  - A supervised learning technique that use a feed-forward Neural Network to identify which sentence must be selected as summary.
  - Receives as input all sentences in text converted into word embedding representation and an one-hot vector that represent the sentence selected as summary (Figure 3).
  - Return the index of sentence with highest probability of being summary (classification problem).

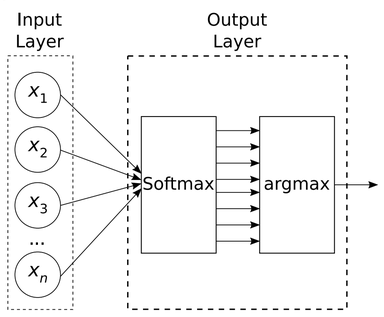

Figure 2: Feed-Forward Neural Network Summarization architecture.

- Recurrent Neural Network Summarization Technique
  - A bidirectional Recurrent Neural Network with Long-Short Term Memory (LSTM) units capable of retaining previous information and looking ahead information.
  - Receive as input each sentence converted to word embedding representation and respective ROUGE score when compared with ground-truth summary.
  - Return a vector with ROUGE score for each sentence (regression problem).

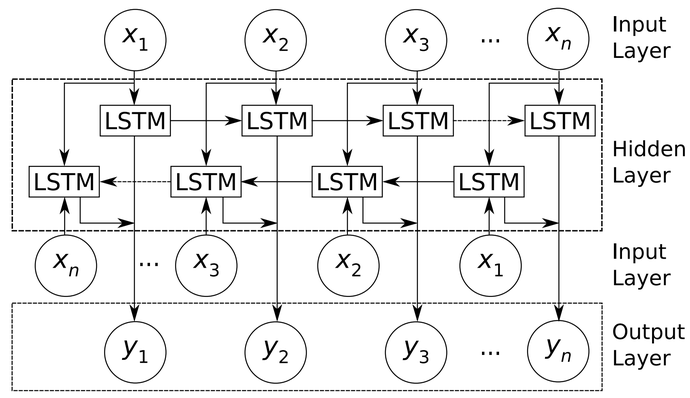

Figure 3: Bi-directional Recurrent Neural Network architecture with LSTM units.

Experiments and Results

- We tested Vector Offset Summarization running over 298,017 examples, acquiring results close to best ROUGE scores that can be obtained over CNN/Dailymail dataset (Figure 4). Our technique chooses the same sentence as best ROUGE score in 35 % of examples.

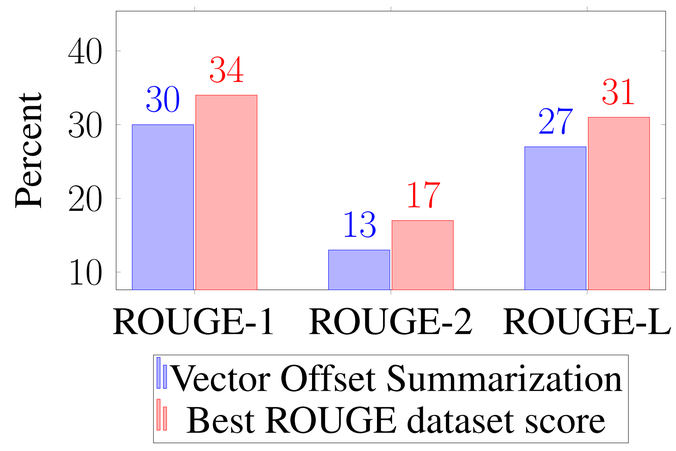

Figure 4: Performance evaluation of vector offset technique over CNN/DailyMail dataset using ROUGE scores.

- Feed-forward Neural Network running over 10,000 training examples and 10,000 test examples result in an overfitting condition. This model gets a good accuracy over training set (around 75 % with 100 epochs), but fails to generalize over unseen examples. Figure 5 show results as follow:
  - Feed-Forward NN 1 with 16 hidden units, no dropout.
  - Feed-Forward NN 2 with 32 hidden units, no dropout.
  - Feed-Forward NN 3 with 32 hidden units, dropout setted up to 0.1.
- Recurrent Neural Network with LSTM units trained over 1,000 examples got an overall accuracy of 13 % over training set and 12 % over test set as shown in Figure 5:
  - Recurrent NN 1 with 8 LSTM units.
  - Recurrent NN 1 with 16 LSTM units.

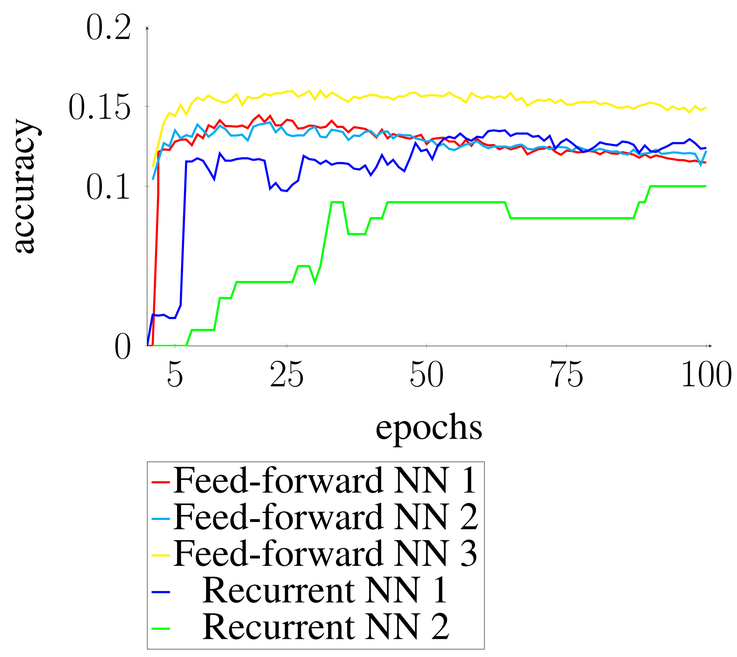

Figure 5: Neural Network test accuracy in function of number of epochs.

References

[1] Mahak Gambhir and Vishal Gupta. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1–66, Jan 2017.

[2] Hans Peter Luhn. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165, 1958.