Sentence Fusion

Andrés Herrera Cepero Laurine Jeannot Quentin Pouvreau
Imane Zouhri

Sentence fusion is a text-to-text generation task which takes related sentences (also called "set of simple sentences" or "simple sentences") as input and merges them into a single output sentence. It is typically used in the context of multi-document summarization or for the production of abstracts, where simple sentences need to be fused into a single more readable sentence. Hence, the aim of this supervised project was to investigate deep learning approaches to sentence fusion using the Split-and-Rephrase dataset and to evaluate them.

This work needed some notions such as knowledge about neural networks or sequence-to-sequence model. Before starting to talk about our work, let's define these notions.

Neural Networks is a biologically-inspired programming paradigm which enables a computer to learn from observational data. This approach combines small algorithms to create a network. Neural networks can be used in many different ways, for tasks such as sentence fusion but also for image recognition or Artificial Intelligence in games.

Then, a sequence-to-sequence model (also called encoder-decoder model) is a RNN (Recurrent Neural Network) architecture that consists of two RNN, one for encoding words into vectors and one to decode these vectors into words (in the target language in case of translation). The interesting part of the encoder is the final hidden state, which contains the information from all the input sequences and outputs a single vector. On the other hand, the decoder takes the vector from the encoder, considers it as the starting state and produces an output sequence, here the output being a complex sentence.

The deep learning model on which we have focused is thus a sequence-to-sequence model, that takes as input a sequence of related sentences and provides a prediction of a complex sentence as output, which has keep the meaning of the set of simple sentences.

To train neural networks, the datasets used are Split-and-Rephrase datasets (benchmark-v0.1 and benchmark-v1.0). A dataset is composed of 6 files: three files for complex sentences and three other files for simple sentences, a training data file, a testing data file and a validation data file. By combination, a complex file and its corresponding simple file form a set of complex-simple pairs.

To well-train a model, the dataset need to pass several times through the network. Each time is called an epoch. Therefore, to improve the training of the model, we have to increase the number of epochs. Here we tested different models by 3 to 23 epochs.

BLEU score were used to evaluate the predictions. This metric is basically a comparison between a human translation and the one made by a machine. To calculate this score we compare the number of n-grams (usually from 1 to 4) that a given sentence shares with the aligned sentence in the reference corpus. Here the reference sentence comes from the complex files and the "machine translated prediction" is the output of the model.

To complete this evaluation, we also defined a manual evaluation system that grades some characteristics of the predicted sentence (such as fluency, grammatical correction, etc.). As it is a manual evaluation, it was not foreseen to do it on the whole predictions but only on a significant part of it.

Globally, the BLEU score of the predictions is quite good, especially for predictions made on 13 epochs. But if we look closer on the sentences produced, there are many repetitions of words and some sentences cannot be defined as complex sentences (*e.g* they do not contain a relative clause). As this project is the reverse process of the Split-and-Rephrase, we could expect to have similar results, but since that is not the case, we may think that another approach is needed.