Sentence embeddings as a path to semantic text similarity and document summarisation

Aman Berry

1 Introduction

1.1 Problem statement

One area which recent advances in deep learning has particularly propagated forward is natural language processing. This field can be abstractly split into two subcategories - natural language understanding and natural language generation. The motivation behind understanding is for computers to develop a human-level understandings of the syntax and semantics of language, whilst generation can be viewed as a continuation of understanding in order to generate human-level language.

This paper seeks to broach both subjects: how can we build a system which is able to understand language, and then use this understanding to generate summaries? Thus far, the generation of summaries, particularly abstractive (over extractive), has proved extremely difficult, even for the most advanced models.

Generally, earlier methods consisted of classical machine learning algorithms, but herein lay a distinct inability to capture the semantics within text. For example, naive Bayes methods were considered relevant in text classification tasks. ...

Recently, methods for encoding different text structures (words, sentences, documents) has become more sophisticated. Generally, ...

While the ideas of semantic text similarity and document summarisation may seem initially distant, they certainly go hand in hand. The motivation behind modern semantic text similarity is the development of sentence embeddings which allow similar sentences to be close (with respect to a similarity measure) in some high dimensional space. These embeddings are inherently relevant to document summarisation. In the case of extractive summarisation, we can use similar sentences to develop a notion of the most important sentences in a document, which can then be presented as the summary. Sentence embeddings are also used to input to a decoder to create an abstractive summary. For these reasons, developing good sentence embeddings is a pathway to developing useful text similarity and summarisation systems.

Furthermore, during work on this paper, work was conducted in the form of an internship on a semantic text similarity system for a medical company. This system is presented in this paper, with the system then applied to an out-of-domain case (which the rest of the methods are also tested upon).

1.2 Research questions

In this work,

1.3 Research methods

2 Related work & context

In this section we look to the past for methods of text similarity and document summarisation.

2.1 Text similarity

The most basic method of matching text comes down to matching words in the sense of word overlap.

2.2 Word embeddings

One hot encoding ... Word2vec ...

2.3 Sentence embeddings

Doc2vec Sentence Transformers

2.4 Document embeddings

3 Methods

We explore a variety of methods related to

3.1 Semantic text similarity

Semantic text similarity

3.2 Lead-N sentences as summary

Experiments done. (baseline)

3.3 Ranking sentences as summary

Lex-rank or text-rank (experiments tbd) This section allows us to join together the similarity ideas and the summarisation.

3.4 LSTM

3.5 Bert Sum Abs?

Can bring in and make improvements to PreSumm paper. This won't include the sentence embeddings from earlier, it will use different ones finetuned specifically for summarisation. Good comparison for how significant task specific embeddings can be.