Advanced NLP

Lec - 4

18/08/22

Is NLP hard? we haven't reached the hard part yet you dumbass!!!

How are Word2Vec and LM related? Traditionally, language is not viewed as temporal. Language modelling itself is employed in almost every task. The LM is only supposed to give you the likelihood. Likelihood in human setting, can stand proxy for fact checker. In real life, we need a measure for how good a model is good enough. We don't need a hard-line but more of an estimation, resulting from a series of errors from the experiments. So, using LM as a proxy has helped us computationally.

When Shannon proposed the word vectors model, it wasn't estimated that the meaning of unknown words will be captured. But weirdly enough, it does.

So, two big hypotheses in LM, likelihood maximisation and meaning encapsulation.

If a topic or a set of topics has high likelihood, it is more probable that they occur together.

Humans are great at constructing positive biases. For example, every school bus, for some or the other reason, is yellow in color. If you were to now show a school bus which is NOT yellow, any image recognition model WILL FAIL.

When supervision is READILY AVAILABLE, it's called self-supervised learning. Word2Vec brought the idea that meaning is distributed. Word2Vec is a global representation i.e, any one word will have just one vector. It is indeed a culmination of all the senses of a word crammed into one vector.

GLoVe : Global Vectors

CoVe : Context Vectors

FastText : Basically similar to Word2Vec but even the characters have meaning. You have both a character embedding and a word embedding.

The problem with all these embeddings is the fact that they were all global. A word's meaning is given by the company it keeps, sure, but that company is not general, or global. It is restricted to a local sense.

Now, RNNs have an interesting propoerty. They are basically information aggregators.

[] -> [] -> [] -> []

x1 x2 x3 x4

here, it isn't difficult to figure out that x4 will have, in some manner, a sense of x1 being carried over to it. But the order of amalgamation itself varies according to the hyperparameters fixed. RNNs are TRANSLATION-INVARIANT i.e. a mere shift in position of words occurring will not change the vector representation.

RNN is a sequential aggregation. If we add another layer on top of this layer of RNN, then I have multiple paths. Now, the second aggregator of the second layer will have a very good meaning complex. But there is another flaw in this line of thought, WHY SHOULD MEANING COMPLEXES BE DIRECTIONAL IN NATURE????

So what we do is go in both the directions i.e. we have an H_i(forward) and H_i(backward). So this arrangement itself is more complex than a simple RNN. It's an LSTM.

For effective next-word-prediction, we need to carry over the complete meaning of all the words which have appeared before.

Ask sir: * how is the averaged delta carried over to the next epoch? (for the matrix U)

- If a sense is reused because of a bidirectional pathway, is it not deterrent? What if individual senses were opposing in nature.

Search up on: Contextual embeddings