Progress Report

Group: NLP
Member: Mingwei Huang (mingwei6)
Nov 14th, 2021

Completed:

- Text collection and text preprocessing
- Topic modeling for the dataset
- A word2vec model for word similarity (may not be used)
- Algorithm for segmenting paragraphs based on the topic distribution of each sentence
- Also, uncommented code is pushed to the repo

Ongoing:

- model evaluation (may need to manually label 10-20 lectures)
    - Evaluate how the model performs on segmenting paragraphs
    - (Maybe) evaluate topic modeling since I have multiple models (which needs word2vec)
- May also incorporate the data here

Future work:

- Try paragraph segmentation by measuring sentence similarities with Sci-BERT
- (Maybe) Summarize each paragraph
    - baseline - BM25
    - evaluation (may need manual labeling too)
    - still Sci-BERT on keyword extraction

Problem:

- The bag-of-word performs too much worse than an average human, and the result is barely useful.
- I suspect the usage of a transformer will not improve the performance a lot even though it can take the context into consideration. Paragraph segmentation requires an extremely fine understanding of the text which is even hard for an average human, and this level of difficulty is too hard for supervised learning, let alone the unsupervised one.

Example result of segmentation of baseline approach using topic difference of each sentence:

number segmented pieces: 15
segmented text:
["This lecture is about the syntagmatic relation discovery. An entropy. In this lecture, we're going to continue talking about word Association mining. In particular, we can talk about how to discover syntagmatic relations. And we're going to start with the introduction of entropy, which is the basis for designing some measures for discovering such relations. By definition, Syntagmatic relations hold between words that have correlated Co occurrences. That means when we see one word occurs in the context, we tend to see the occurrence of the other word. So take a more specific example, here we can ask the question whenever

eats occurs, but other words also tend to occur. Now looking at the sentence is on the left. We see some words that might occur together with eats like a cat, dog or fish is right.",
'But if I take them out and if you look at the right side where we only show eats and some other words. The question that is, can you predict what other words occur?',
'To the left or to the right.',
'Right, so this would force us to think about what other words are associated with eats.',
"If they are associated with eats, they tend to occur in the context of eats. So more specifically, our prediction problem is to take any text segment, which can be a sentence, paragraph or a document, and then I asked the question is a particular word present or absent in this segment. Right here we can ask the question about the world W is present or absent in this segment. Now, what's interesting is that some words are actually easier for it, in other words. If you take a look at the three words shown here, meet, the and Unicorn. Which one do you think it is easier to predict? Now, if you think about it for a moment, you might conclude that. The is easier to predict because it tends to occur everywhere, so I can just say with the in the semtence. Unicorn is also relatively easy. Because Unicorn is rare, is very rare. And I can bet that it doesn't occur in this sentence.",
"But meat is somewhere in between in terms of frequency, and it makes it hard to predict because it's possible that it occurs in the sentence or the segment more accurately. But it may also not occur in the segment. So now let's start this problem more formally. Alright, so the problem can be formally defined as predicting the value of a binary random variable. Here we denoted by X sub w, w denotes a word. So this random variable is associated with precisely one word. When the value of the variable is 1, it means this word is present. When it's zero, it means the word is absent, and naturally the probabilities for one and zero should sum to 1. Because a word is either present or absent in the segment. There's no other choice. So the intuition we discussed earlier can be formally stated as follows. The more random this random variable is, the more difficult the prediction would be.",
"Now the question is, how does one quantitatively measure the randomness of a random variable like X sub w, how in general, can we quantify the randomness of a variable? And that's why we need a measure called entropy. And this is a measure introduced in information theory to measure the randomness of X. There is also some connection with the information here, but that's beyond the scope of this course. So for our purpose we just treat the entropy function as a function defined on a random variable.",
"In this case it's a binary random variable, although the definition can be easily generalized for a random variable with multiple values.",
"Now the function form looks like this. There's a sum over all the possible values for this random variable inside the sum, for each value we have a product of the probability that the random variable equals this value and log of this probability. And note that there is also an negative sign there. Now, entropy in general is not negative and that can be mathematically proved. So if we expand this sum will see the equation looks like a second one I explicitly plugged in the two values zero and one. And sometimes when we have 0 log of 0, we would generally find that as zero because log of 0 is undefined. So this is the entropy function and this function will give a different value for different distributions of this random variable. And this clear it clearly depends on the probability that the random variable taking a value of one or zero. If we plotted his function against the probability that the random variable is equal to 1 and then the function looks like this. At the two ends, That means when the probability of X = 1 is very small or very large, then the entropy function has a lower value when it's .5 in the middle that it reaches the maximum. Now, if we plot the function against the probability that the X is taking a value of 0 and the function would show exactly the same curve here. And you can imagine why and so that's because the two probabilities are symmetric and completely symmetric. So an interesting question.",
'You could think about in general here is for what kind of X? Does the entropy reached maximum or minimum and we can in particular think about some special cases.',
"For example, in one case we might have a random variable that always takes the value of one, the probability is one or there is a random variable that Is equally likely taking a value of 1 or 0. In this case, the probability that X = 1 is .5. Now, which one has a higher entropy? It's easier to look at the problem by thinking of simple example. Using coin tossing, so when we think about the random experiment like a tossing a coin, it gives us a random variable that can represent the result. It can be head or tail, so we can define a random variable X sub coin so that it's one when the coin shows up as head, it's zero when the coin shows up as tail. So now we can compute the entropy of this random variable, and this entropy indicates how difficult it is to predict the outcome of a coin for coin tossing. So we can think about the two cases. One is a fair coin, it's completely fair. The coin shows up as head hotel equally likely, so the two probabilities would be, 1/2 right so both will have both equal to 1/2. Another extreme case is completely biased coin, where the coin always shows up as head, so it's a completely biased coin.",
"Now let's think about the entropies in the two cases, and if you plug in these values you can see the entropies, would be as follows for a fair coin",
"we see the entropy reaches its maximum, that's one. For the completely biased coin we see is 0 and that intuitively makes a lot of sense because a fair coin is most difficult to predict whereas a completely biased coin is very easy to predict that we can always say it's a head because it is a head all the time so they can be shown on the curve as follows. So the fair coin corresponds to the middle point, or it's very uncertain. The completely biased coin corresponds to the end point. We have a probability of 1.0 and the entropy is 0.",
"So now let's see how we can use entropy for word prediction. Now the problem, let's think about our problem right, still predicted whether W is present or absolutely in this segment. Again, think about the three words. Particularly, think about their entropies. Now we can assume high entropy words are harder to predict. And so we will now have quantitative way to tell us which word is harder to predict. Now if you look at the three words, meat, the and Unicorn again.",

"An we clearly would expect the meat to have a high entropy, then the OR Unicorn. In fact, if you look at the entropy of the, It's close to 0, because it occurs everywhere. So, it's like a completed biased coin, therefore the entropy is 0."]