WEBVTT

1

00:00:13.160 --> 00:00:14.630

Jisun An: Alright

2

00:00:14.820 --> 00:00:37.609

Jisun An: welcome to the course of fundamentals and applications of large language models. Today I will go through a little bit of overview of the course and introduction. And actually, the next lecture is called World Representations, and we will go through 1st middle part of that, so that that lecture will continue for the next 2 lectures.

3

00:00:38.380 --> 00:00:48.110

Jisun An: So everyone knows that the larger language model is something very, very popular. It has about 200 million users in this

4

00:00:48.110 --> 00:01:13.110

Jisun An: end of August, and then it added 50, I mean, and the Chat Gpt is probably the most popular large language models, and I assume that everyone have been using it for different ways, right? It has increasing number of the users recently and last month it reported that they have 300 million weekly users, weekly active users. And now it's not only the Chat Gpt, but, as you know, that there are many other

5

00:01:13.110 --> 00:01:28.569

Jisun An: different types of the large language models, including our cloud perplexity, or co-pilot, or gemini, and etc. Etc. And across all the language models that are increasing users, which means that it gets really popular and really comes to our daily life.

6

00:01:29.340 --> 00:01:46.810

Jisun An: And the one thing that was really interesting to me was that what happened in the final week of the last term. I don't know whether you were experiencing that, but I saw that I heard that I wasn't using Gpt at that time, but I heard that Gpt. Got failed during the final week.

7

00:01:46.860 --> 00:02:13.970

Jisun An: and for about it was on December 11th or 12, th from 3 Pm. To 7 Pm. So for about 4 h, so over the undergrads, and the graduate who were working for their final exam was shocked, and there were many, many memes and sharings in the Tiktok that they were really complaining about, what should we do without Gpt when preparing for the final exam? I mean, this was the point that I realized that. Okay, now, this is really something

8

00:02:13.970 --> 00:02:19.439

Jisun An: getting into really deep into our lives, and especially for the young generations.

9

00:02:20.507 --> 00:02:39.170

Jisun An: So what do we do with the edit? So in my own experiences. I guess I'm also the international immigrants. And I and my mother thing is Korean, because I'm from Korea. And so I get a lot of help with writings and editing definitely. I do most like grammar checks. I asked to write an emails.

10

00:02:39.300 --> 00:03:02.469

Jisun An: Many, many help with the writing editings and also coding so well. Now, I don't do coding much. All my students are doing coding but while I'm preparing this course I was doing a bit of hands on exercise with working with the edit, and I was really surprised at how well they can support for the coding, and I assume that you all have experienced that. But I was shocked, shocked.

11

00:03:02.470 --> 00:03:20.120

Jisun An: And then I also started to do more and more knowledge searching. So I found that for not the knowledge, but no more searching. I still prefer just to use the Google. But if I'm looking for a particular knowledge, then I found that I mean, Gpt summarizes really well. So I really liked it in using that way.

12

00:03:20.330 --> 00:03:36.209

Jisun An: And also, I'm also using translations, etc. Etc. So I mean, these are some common kind of practice or application that I'm doing. But is there any other thing, that other interesting way that you are using this add items apart this, apart from these 4,

13

00:03:37.230 --> 00:03:43.020

Jisun An: anything, most of it making plan.

14

00:03:43.590 --> 00:03:50.899

Jisun An: Oh, travel plan. Okay, that sounds interesting. Any any other in experience.

15

00:03:51.590 --> 00:04:15.339

Jisun An: So I don't know. I didn't include it here. But now, I think different models are supporting multimodals, so probably using something images or like Pdfs. Oh, the one thing is, I did it. Probably I was going to let you know. Tell you more about it later. But I created a image for this course. I don't know whether you've seen it. If you go to canvas the image that you saw it

16

00:04:15.380 --> 00:04:33.580

Jisun An: for the course so that I created help with the chat gpt so so now I mean, as you know, that we can also do it with with joint, with the I think they were using Dali to as on as on like external Api. But I was still doing be able to do it through the chat. Gpt.

17

00:04:34.320 --> 00:04:48.778

Jisun An: yeah. So they were. Actually, this interesting paper. I think it was coming from the people who are working with the vicuna, the another another model of the Rrm. And they actually opened a chat arena, and

18

00:04:49.140 --> 00:05:13.920

Jisun An: chat arena, which was this was interface you can actually go to. Oh, I didn't put the URL here, but if you can search for Chat arena. There is a place where you can compare different models of the outputs outputs of the different models, so you can ask the same question, and then they will, using the Api. They will get the results from 2 different models. And then you can actually compare the results from the 2 different models. And and so they just leave this place

19

00:05:13.920 --> 00:05:21.810

Jisun An: out there and then they just collect all this. You know different how actually, people are using these different models, and it was more than

20

00:05:21.830 --> 00:05:32.490

Jisun An: more than 1 million few 1 million kind of instances to collect the collective answers from all different users. And they did a simple clustering topic, kind of modeling.

21

00:05:32.520 --> 00:05:48.133

Jisun An: and to see what people were using doing with the Nnms. And I mean, you should know that these are very particular platform. Probably those very techy who are interested in large language model who also wanted to compare different large language model were coming and joining to this

22

00:05:48.620 --> 00:06:17.080

Jisun An: of experiment, not experiment, but creating this data set. But, as you see, there are the most popular one was discussing software errors and dissolutions, which is about like programming and debugging. But then there are something about like asking geography, travel, and global culture inquiries. So I think that was something very interesting and also they were also asking, creating and improving business strategies and products. So maybe trying to gather some ideas about different things

23

00:06:17.080 --> 00:06:32.499

Jisun An: and role playing various characters in conversation. So that was, I mean, it's another big thing, AI agents and the character AI. Probably you heard the movie. But that was something people were doing. And there are also, like many other types of the role playing here and there.

24

00:06:32.650 --> 00:06:50.210

Jisun An: and the last cluster is inquiries about specific plant growth conditions. I know that there are people who are really into growing plant, but apparently these were some of the most asked questions, or most had conversation with these models.

25

00:06:50.760 --> 00:07:15.630

Jisun An: So people are all doing this interesting thing. So one thing, another thing that I find it interesting, which Gpt was not able to do was retriever or working with the most recent information. So one of the weakness the larger engine model had because it is trained on certain data set that's already existing is usually harder to catch up with, like what's actually happening at the moment, or real time kind of

26

00:07:15.820 --> 00:07:40.669

Jisun An: actions. But then I just do it. I tried to retrieve how? I asked. So what's the top news today? Yesterday? I mean on on January 13th to the Chatgpt, and they actually using now the Apis to carry over to some of the few news media like Ap. Cnn. New York Times, and etc. I think they have some partnership. And then they actually got the news article, summarize it, and then

27

00:07:40.850 --> 00:08:05.230

Jisun An: show me here, and also, not only for the U.S.A. But like how? What's the top news in Korea, and they were also able to retrieving and authorizing. So I mean, these were not possible, I think, like a month ago, and then, as you see they are integrating different functionalities and enabling many, many things that I thought it never be able to do it. But I think the changes is far faster than I expected.

28

00:08:06.120 --> 00:08:16.650

Jisun An: but at the same time is not the I mean. They are not able to do everything, but they sometimes do fail in different ways. I think there's a 1 share here.

29

00:08:17.616 --> 00:08:37.730

Jisun An: so this is actually a cases collection of the failure cases. But I think this was tested on Gpt. 3.5, which was on one year old model. So I would just say, it's already outdated model. But back then, at least they couldn't do these things. So write a sentence that ends with the letter S.

30

00:08:38.110 --> 00:09:04.559

Jisun An: Write a sentence that contains the letter F exactly 3 times, and then some kind of like keys. Question like Mike's Mom's had 4 keys. The 3 of them are Louise Drake's and Matilda. What is the name of the 4th key? So this kind of key is like a requires a little bit of reasoning, and it's a bit of a keys and like, basically, I'm sorry, but I don't have information on the 4th key. It was the answer from the Gpt. 3.5. And then.

31

00:09:05.166 --> 00:09:10.413

Jisun An: what is the taller Mount Everest of the very Khalifa

32

00:09:11.280 --> 00:09:22.640

Jisun An: like that. So so these kind of questions they couldn't properly answer. But then I tried the exact the same questions to the Gpt. 4 or and they were now able to answer

33

00:09:22.850 --> 00:09:33.574

Jisun An: perfectly well. I think these were all correct. information. So now they were able to tell. I mean serving some of like minimum, I mean, at least like the easy

34

00:09:34.140 --> 00:10:01.989

Jisun An: huddles or reasoning classrooms. And they even know that the heathrow and the same pancreas are basically are in the station in the same city. So basically, there is not possible to have the airport, nonstop flight between the 2, etc. Etc. The only thing they couldn't do it was this second one. They still failed, but I think I just asked this questions, but if I were instructing them better. I probably assume that they probably do it very well.

35

00:10:02.320 --> 00:10:06.380

Jisun An: So once again, they are just improving. Improving. Over time.

36

00:10:06.947 --> 00:10:16.312

Jisun An: But still there are some other domains where the Rrm still may still pay and can give a more significant impact.

37

00:10:16.840 --> 00:10:38.390

Jisun An: basically. So one of such problem is the hallucination. So probably everyone heard about this. So because it eventually the language models are model that predicts the next tokens. So whatever that they say may not be really the truth. So, even though it sounded reasonably and perfectly sound. But it may not be the picture information, or may it may still mislead.

38

00:10:38.390 --> 00:10:47.620

Jisun An: so there are like the technique, like the retriever augmented generations that are coming up to accompany so we do with this hallucination

39

00:10:48.020 --> 00:11:14.420

Jisun An: issue. But still, and and one of the particular domain that has been most important. The factory information is the most important was that the legal area and I think they have been integrating this new system with the rags, but still with recent study, found that these tools with rags are still misled, like 70% of the times. So still there are some rooms for the improvements.

40

00:11:15.080 --> 00:11:23.056

Jisun An: Another important issue was that so? There was this one tragic cases where the 14 year old.

41

00:11:23.660 --> 00:11:41.519

Jisun An: was a chat, had a chat with the AI, and somehow made a suicide. And then now the parents were kind of blaming, blaming that the AI was one of the reason that they led. Of course these are just news articles, and it is just there should be something more taking, deeper looking in. But then.

42

00:11:41.520 --> 00:12:06.140

Jisun An: I think for the like younger generation, everything's a really dangerous, and you really cannot tell where how that will affect them so safeguard of this largely model will be very, very important. And one of the areas like, so the AI alignment which is trying to make these models to align well with our values, and our intention is one of the big research area which we will also cover later. In this

43

00:12:06.140 --> 00:12:06.860

Jisun An: course

44

00:12:07.600 --> 00:12:24.329

Jisun An: another significant issue is jailbreaking. So I think the very initial version of the initial version of the Gpts or any other models, they were bluntly let you know that how to build a bomb at home. So and now I assume that

45

00:12:24.570 --> 00:12:52.330

Jisun An: I mean so it's almost like a fight between, like the virus and the vaccine kind of fight. So the models are trying to be safeguard all these homes. But then there are hackers that for trying to jailbreak. So I think this fight is actually still going on. And this was like a month ago, and another hacker tricked Gpt. To giving the details of the how to make the homemade bombs. So this is another security issues. And and lastly.

46

00:12:54.100 --> 00:13:14.498

Jisun An: we already suffer from like phishing, and the skimming and then now Gpt makes that even even much easier and in a larger scale. So they actually reported that, like the amount of these phishing emails that you are receiving was like 300 increase in the last few months or so. So these are some of the

47

00:13:15.190 --> 00:13:28.630

Jisun An: issues that you may also think while you're working with the large eggs model. This is the model, and this is not, I mean, even the human is not perfect, but models are not perfect, and you should be always aware there's a like potential risk in using these models.

48

00:13:29.350 --> 00:13:30.770

Jisun An: But anyhow,

49

00:13:32.170 --> 00:13:55.439

Jisun An: but anyhow, I think all of you here are actually, we have very diverse pool of the students. Some of you are coming from the informatics. Some of you are from computer science. Some of you are coming from cognitive science, psychology, public health, very interdisciplinary, but I assume that in one way or the other you will use this large language, model in your future career or in your daily life.

50

00:13:55.440 --> 00:14:07.089

Jisun An: So some of you will actually build the system using the larger model. Some of you, some of you may do the research with the large language model. Some of you just using this large language model in your daily task.

51

00:14:07.090 --> 00:14:29.619

Jisun An: So we were going to all use it. But what I think the most important when you're using this larger language. Models are because it has the risks and the issues and the challenges. You at least need to understand how it actually works and what is the most how it has been developed so that you can understand or avoid potential issues or the challenges and then confront the challenges.

52

00:14:30.690 --> 00:14:38.039

Jisun An: So. And but the problem is that the development is just so quick and so fast, and you probably

53

00:14:38.200 --> 00:14:59.419

Jisun An: maybe haven't heard of some of them. But just. There are so many, many models around us. So how will you be able to catch up all of them? And obviously this course will not be able to cover all of these models. But what we are trying to achieve here is that there are some fundamental components in the blocks that you need to. You need to know to understand how things has been developed.

54

00:14:59.420 --> 00:15:25.369

Jisun An: So the course will mainly focus on those building blocks and the major components so that you can understand how it started and how it ended up now and then. In the future you will see even growing trees of these models, and hopefully, when the new model comes in, at least you will understand why this model came and what it came for, and potentially, maybe, how to improve or use it for your own task

55

00:15:25.850 --> 00:15:29.409

Jisun An: and different application of this, add them

56

00:15:29.690 --> 00:15:33.089

Jisun An: is something what this course will aim to deliver.

57

00:15:33.940 --> 00:15:49.319

Jisun An: Right? So what is a large language model. So I think everyone knows it. But I just want to define it clearly before moving forward. So what is a large link? Model? To know to

58

00:15:49.610 --> 00:16:04.449

Jisun An: how good to guide. What is the what is Nlm, we need to actually start from the Nlp. So the Nlp is the natural language processing and Nlp is a form of AI. I think we have 2 chairs here. Can you come in? Yeah.

59

00:16:04.750 --> 00:16:07.329

Jisun An: I don't know why we are more than

60

00:16:07.550 --> 00:16:10.033

Jisun An: it's okay. Anyhow. So

61

00:16:10.650 --> 00:16:24.370

Jisun An: and then Nfp is the form of the AI that aims to enable computers to understand, interpret, and interact with the human languages, whether it be written, spoken, or even scribed.

62

00:16:24.370 --> 00:16:43.779

Jisun An: So I technically spoken languages, or sound, or like the images, and like extracting the text from them, those all be part of the Nfp. But like we, we were mostly talking in the like the written text. But basically, the Nfp is something that trying to computer to

63

00:16:43.910 --> 00:16:46.930

Jisun An: do the processing of the human languages.

64

00:16:48.800 --> 00:16:52.360

Jisun An: And then there are multiple different applications of the Nrp. Probably

65

00:16:52.460 --> 00:17:13.790

Jisun An: you no problem in most of them. But you can do text analysis. You can do language, generation, code, generation, machine translation, speak recognition and processing information retriever named entity, recognition and question and answering, and these are the various applications of the Nfp. And you, do you? Do you? Is there anything that you don't know?

66

00:17:14.490 --> 00:17:16.109

Jisun An: It's so obvious? Right?

67

00:17:17.000 --> 00:17:24.049

Jisun An: Oh, and then the text classification. And there are different techniques used in this Nlp, so probably like

68

00:17:24.839 --> 00:17:54.610

Jisun An: more than 50 years ago, it started with the rule-based approaches right to tackle those applications they use like predefined linguistic rules to process the languages. And and then there was like a statistical method. So they were using probabilistic model to analyze, like the language patterns like the engrams. And then now there was like the machine learning and the deep learning approaches that trying to tackle the Nfp applications. So initially, there were very traditional machine learning models like support vector machines or

69

00:17:55.214 --> 00:18:09.440

Jisun An: Like logistic regressions. And then, like the the simple kind of neural networks. And now there are deep learning techniques like Rns or Stm's. And they transform. So here are so basically

70

00:18:09.620 --> 00:18:21.609

Jisun An: transformer. Probably many of you have heard of it. This is the the central building block that was enabled the current large scale language model exist.

71

00:18:23.100 --> 00:18:29.739

Jisun An: So within this context, the Rrm is an advanced type of Nlp models.

72

00:18:30.116 --> 00:18:52.489

Jisun An: Designed to understand, generate and manipulate human language by leveraging this deep learning technology. So the Rnm is the part of the Nlp. And it helps to tackle different applications. So you can say that AI nlp, and there's a large language models. But then the application of the large language model, I think it really expand to many different domains.

73

00:18:52.490 --> 00:19:04.869

Jisun An: So usually these Edms are pre-trained on, like the massive data set from diverse text sources. And also it's built on the transformer architecture. So that would be something very important that we will talk later.

74

00:19:05.450 --> 00:19:07.929

Jisun An: And then the applications of the Rrms.

75

00:19:08.406 --> 00:19:35.220

Jisun An: Rrm, can handle most of the applications of the Nlp that you've just seen it in the previous slide. But importantly, they do not require specialized algorithms or specialized data set for each of the applications. So if you think about those applications before the Rrm era, you probably needed to have a special models, algorithm or data set to just to tackle that particular task.

76

00:19:35.260 --> 00:19:52.260

Jisun An: So do you have model for translation separately. You have a model for named entity recognition separately. You need a model for different task. But then now the large language model and the beauty of it. It was so large and trained in a way that they can tackle all these tasks

77

00:19:52.690 --> 00:19:54.000

Jisun An: at once.

78

00:19:54.820 --> 00:20:01.660

Jisun An: So so I think that's the like. The slight differences between Np. And Nfp. Models and the large lengthy models.

79

00:20:02.280 --> 00:20:16.419

Jisun An: And so, and yeah, so this was the largest instead of creating individual model for specific and epitask. Now you can have one single M, and that can help perform for the right range of C task effectively.

80

00:20:17.270 --> 00:20:45.779

Jisun An: So so that's the Rrm as we define. And in this class we will try to learn the principles and the architecture and the training method of these large language models. So that will be the fundamental part. And then we will also do some hands-on experience on how to fine tune these models, or how to apply these models and various applications of the Rrms. Like retriever augmented generations, and the AI agents, etc. And also we also

81

00:20:46.480 --> 00:20:52.369

Jisun An: highlight how you can critically assess the ethical, social, and practical implication of these models.

82

00:20:54.526 --> 00:21:19.080

Jisun An: So these, the course we have about 4 formed with the 4 different topics. So the 1st one will be the language modeling fundamentals. So we will start from how we can represent words and how we can model the language, how we can encode the sequences. And what is this transformer? Which is, I assume, that everyone have heard of it, how the what is the transformer architecture, and how it works.

83

00:21:19.260 --> 00:21:35.250

Jisun An: and then how these models are trained, and how you can inference from these models. So the the steps of the pre-training, fine tuning instruction, prompting reinforcement, learning. I assume, that probably you heard of these words here and there, but I hope to have

84

00:21:35.320 --> 00:21:49.689

Jisun An: hope you to have a more clear idea of each of these terms throughout this course, and finally we will tackle. We will look at a few edited applications, such as, like the reasonings and the retrieval, augmented generations and the language agents.

85

00:21:49.970 --> 00:21:56.472

Jisun An: and finally, how the edit M has been applied to some various domains in different

86

00:21:58.580 --> 00:22:11.770

Jisun An: domains like education, finance, business, law, healthcare, and the social science, and this last part will be achieved not only by me, but also with you. So the projects, basically, I hope that

87

00:22:11.870 --> 00:22:36.289

Jisun An: the project scope will be wide enough so that you can basically do anything you'd like in different domain. And I hope that can help all of us to understand what's actually happening in different other domains. So any particular domain that you are interested in, and maybe you can come up with some ideas at how you can use or apply these, add items in different domains, so hope as collectively, we can learn more about different domains for the edit.

88

00:22:37.210 --> 00:23:01.924

Jisun An: So I will go through briefly about just a quick class format and structure, and all the information you can find it on the on the canvas, and then I will put. I will try to put everything on the 1st page, so that would be the only page that you will need to know. So teaching team. So myself, I'm I'm an assistance assistant professor at iu informatics. And my research area is

89

00:23:02.560 --> 00:23:30.507

Jisun An: in general, computational social science and social computing and the applications of the Nap and the Llms and the social impacts of the Llms. My office hour is from Tuesday, from 9, 30 to 1030, and my zoom links and the office hour will be through the zoom. So please use that zoom room, and our ta is Fan Huang. He's he's there. He's now 3rd year Phd. Students here working with me at Iu Ladi.

90

00:23:30.930 --> 00:23:41.799

Jisun An: he's been working with large language models, especially AI alignment with the humans. He has a great experience with working with this models. So

91

00:23:41.800 --> 00:24:01.270

Jisun An: basically, if you have any questions about the programming, and that's what you can go with it. And then his office hour will be Friday 3 to 4, and his zoom link is over there. So you can always also send an email to us if you have any particular questions, and I think that'd be better to also answer that if you cannot come to this office hours

92

00:24:02.240 --> 00:24:09.240

Jisun An: and so pre liquid. So I think this will not be very, very technical, but at least

93

00:24:09.400 --> 00:24:25.710

Jisun An: you should have some basic understanding of the linear algebra probability and the statistics. And also you need to be proficient in Python. I assume that you already know how to programming, and especially because you will have individual programming assignments and also the projects work.

94

00:24:25.976 --> 00:24:41.420

Jisun An: And also even though I will try to give you more basic concepts. But like really basic thing, like how to split, how to use this, how to build the like basic machine learning models. I think that will not be covered in this course. So I hope that you kind of already have those basic concepts.

95

00:24:43.710 --> 00:25:09.319

Jisun An: And so so how the evaluation of the course will be so the big portion will be the group projects which will be the 40%. The once again the scope of the projects will be quite broad and diverse. But we will have 3 components in it. Portal presentation, final presentation, final submission. So which will be each of 10% and 10% and 20% respectively.

96

00:25:09.440 --> 00:25:18.249

Jisun An: And we will have a theoretical assignment, 15% practical assignment, 15% exam, 20% participant and 10%

97

00:25:19.060 --> 00:25:23.609

Jisun An: so group projects. So I think it's still too early. But once again,

98

00:25:24.210 --> 00:25:36.329

Jisun An: I wrote a scope in this way, because these are some of the things that I am interested in. But we are pretty open, I think, whichever topics or domain that you're interested in. And so you can.

99

00:25:36.560 --> 00:26:02.310

Jisun An: You can think it as a more research oriented. So you can come up as a more you have particular question or hypothesis, and you want to test Vdr editance, or you can go with more practical driven projects where you have particular problem that you want to solve. And you want to just provide a solution with the edit them for that particular scenario. I think either should go okay. But we will have about 5 students per group

100

00:26:02.310 --> 00:26:08.720

Jisun An: and like presentations and the final reports. But more details will come out later.

101

00:26:09.862 --> 00:26:34.147

Jisun An: If you have. No, I mean not a great idea. Then maybe you can also check like different and what things people have been working on. Maybe you can also gather some ideas from them. They were trying to build some interface that are connects to the Wiki search. So I think that the 1st one they kind of tried to given a keyword that users are providing they were

102

00:26:34.784 --> 00:26:57.769

Jisun An: using the so retriever. Sorry, retrieve some information from the Wikipedia, and then they summarize and then present it. So, even though this is a simple idea. But if you think that for certain domain, if that kind of functionality is useful, then maybe building, that would be one of the projects. But once again, this is more practical side of the projects. But you're also welcome to take, like the research oriented projects as well

103

00:26:58.930 --> 00:27:04.490

Jisun An: the theoretical assignment. So we will have 2 individual theoretical assignment.

104

00:27:04.490 --> 00:27:29.459

Jisun An: And basically, these will be multiple choice questions of which you basically like the keys. And these are initially helping you to understand better about the lecture contents. So for some of you, maybe something just goes too fast, so it will be a good time to reflect whether you actually understand or not. I hope you will not going to use Gpt to solve those keys. So I mean eventually you will have exam, which will be

105

00:27:29.460 --> 00:27:49.219

Jisun An: very similar to this piece. So at least go through once and try to solve by yourself. I assume, I assume, that nowadays these levels of questions can be easily solved by, and I know that. But these are for your own sake, and so I hope you you manage themself by yourself well.

106

00:27:49.340 --> 00:28:08.180

Jisun An: and then we will have one individual programming assignment using Edm. And this one of details will be announced soon. But we will. We will ask you to fully use Gpt to do the programming and then and then see what how we can do. So that will be something interesting. I will give you more details later

107

00:28:09.030 --> 00:28:16.279

Jisun An: and and in doing so, I just want to ask, how familiar are you with the python.

108

00:28:17.500 --> 00:28:26.339

Jisun An: python, python? Everyone right? Right? I assume everyone. Yeah, if you're not familiar with python, I'm sorry this course may not be for you. Are.

109

00:28:26.760 --> 00:28:34.590

Jisun An: How many of you are familiar with the R, okay, cool. Javascript.

110

00:28:35.560 --> 00:28:40.360

Jisun An: Okay, okay, this is popular. One. Okay, thank you. And go no. One.

111

00:28:40.910 --> 00:28:50.340

Jisun An: Okay. Rust roast 2, 3, okay. And then finally, Julia, how do you know him?

112

00:28:50.720 --> 00:29:14.470

Jisun An: Okay, cool, I see. So you are on your spectrum. Okay, I see. Thank you. Thanks a lot. I will. Yeah. So the details will be announced later. The lastly, the exam will be actually 20%. So it's actually quite a lot. And it will be very similar to what you will see in the theoretical exam theoretical assignment, so that it can guide you in a way that how you will prepare this thing, but it will be in person. Close the book.

113

00:29:14.580 --> 00:29:16.621

Jisun An: so be prepared for that.

114

00:29:17.330 --> 00:29:41.629

Jisun An: Oh, and then the attendance, so we will take you on that attendance. But we will drop the 4 lowest grade for attendance, meaning that you are allowed to miss 4 classes and maximum without penalty. So you don't need to send me any email, let me know, like whether you're missing or not. Up to 4 you will be fine, but if you are now going over 4, then maybe you can send me an email, and that let me know that you you were gonna miss the

115

00:29:41.650 --> 00:30:05.200

Jisun An: lecture. But anyhow, to to check here to mark your attendance. Please go to canvas and assignment attendance, and you will see the one with the thursday, January 16, th and you will basically need to write. Put this password in it, and then mark that you are present in the lecture, so that will mark your attendance. But please do that while you are here.

116

00:30:24.580 --> 00:30:30.850

Jisun An: Oh, and the password will will be added, and let me know if it doesn't work. It should be fine.

117

00:30:33.470 --> 00:30:39.437

Jisun An: right just a quick, I mean schedule. You can find it from the campus. But

118

00:30:40.400 --> 00:30:55.050

Jisun An: so you you see that the 1st 3, 4 weeks will be very intense lecture based. I'm sorry just this is something that we just had to. We go through. So the 1st 4 weeks we will start from the word representation till the prompting.

119

00:30:55.050 --> 00:31:22.420

Jisun An: There will be a lot of information that I will crash into and squeeze into you. So just just warning. So it will be very lecture intensive. But then, after that we will have a few lab sessions which will be led by our Ta, and also some interesting topics of like AI alignment, and the safety or rags or agent. So later, in this course, it will be a bit more relaxed.

120

00:31:22.920 --> 00:31:30.129

Jisun An: And also we'll be focusing more on doing projects later on once again from up to week 5.

121

00:31:30.460 --> 00:31:53.439

Jisun An: I think it will be quite intense, and hopefully it will be more relaxed later. Over the courses. So we will have the project proposal presentation right before the spring break, and then after that we will have some application related sessions, and the final last 3 weeks will be all about project. So the our final project presentation will be on week 15.th

122

00:31:53.440 --> 00:32:12.879

Jisun An: So the week before that, we will have a project hectic. So basically, this will be project consultation. So we will still be here. But we will just help you to go through your projects. And the last week we also will be the final report week. But this will be online and the option. So we're not going to take the attendance for the last week's course, but

123

00:32:12.880 --> 00:32:27.909

Jisun An: what you will need to do here. So after the presentation, we will give you feedback on your projects, and for your final report, you should reflect all the comments for your final report, and this will be the time that that hope you to kind of work on those final reports.

124

00:32:28.770 --> 00:32:48.390

Jisun An: So so that's the thing, and these will be the key dates for this course. So mark your calendar in advance. What are the requirements? And and also there may be modifications, but I will let you know, but this will be something. and probably the exam date will not change. So just mark that date. That will be the most important thing.

125

00:32:49.450 --> 00:33:13.200

Jisun An: All right. I think. Last thing is AI policy. So so usually, AI tools are permitted this work process. But for the projects, the idea of the project should be your own. But you can concert with AI. But but hopefully the idea is coming from yourself. And also you can do you can use AI, but just just you need to be transparent about it. So

126

00:33:13.200 --> 00:33:25.390

Jisun An: for your final report you will be required to write a short paragraph about how you actually use this AI in in your final projects. So as long as you declare about them. Everything should be fine.

127

00:33:26.489 --> 00:33:34.210

Jisun An: I think this is all about the class. Do you have any questions about formats and the structure?

128

00:33:35.300 --> 00:33:42.500

Jisun An: Any prerequisites? Okay, all good. Oh, yes, oh, yes, I'm

129

00:33:44.783 --> 00:33:54.056

Jisun An: I will not gonna randomly assign, that's for sure. So I will. I will probably ask you to form the group, and then

130

00:33:57.380 --> 00:34:06.709

Jisun An: What? What are we going to do? I will leave open groups in the in the canvas, and then everyone can sign up on the Excel file later.

131

00:34:07.390 --> 00:34:23.617

Jisun An: so the reason that I'm hesitating here is so we have like mixture of the master students and the Phd students. So I I haven't decided whether we need to split them or doing just them all together. So I will. I will give just a little moment for thinking about it.

132

00:34:25.440 --> 00:34:26.280

Jisun An: yeah.

133

00:34:27.780 --> 00:34:33.229

Jisun An: So if you already know anyone then maybe it would be good to.

134

00:34:33.790 --> 00:34:37.659

Jisun An: So the Guru formation will start in week about 3.

135

00:34:39.710 --> 00:34:56.630

Jisun An: sorry. There are many things that actually, that's a really good question. I have many things in my hand. So normally, I randomly assigned and I guess I guess that may or may or may not happen. I will. I will let you know next week. Yeah, let me think about it a little bit more. Yeah, thank you. Yeah.

136

00:34:56.840 --> 00:34:57.620

Jisun An: yeah.

137

00:34:58.650 --> 00:35:03.280

Jisun An: Yeah. Okay. All right.

138

00:35:11.950 --> 00:35:12.855

Jisun An: Alright.

139

00:35:14.480 --> 00:35:41.909

Jisun An: So let me start with the the 1st lectures of this course we will talk about the word representations and the text classifiers and declarations. Many slides are from Graham, New, who are from Cmu and decent Stanford. They have really great and ab courses, and I got a lot of inspiration from their courses. So mind that these slides are coming from those lectures. Some of them are coming from those lectures.

140

00:35:43.320 --> 00:35:50.140

Jisun An: So, as I told you, our aim, probably until Week 4 is

141

00:35:50.340 --> 00:36:02.209

Jisun An: for you to understand the transformer. And there are many different components that you you need to know, to understand the transformer. And they are how they actually trained.

142

00:36:02.910 --> 00:36:13.369

Jisun An: And these are some of the key components, which is the so word models and word representations, neural networks, and the sequential model which we were going to do it this week. And the next week.

143

00:36:14.480 --> 00:36:33.899

Jisun An: So let me start with a Nlp system building overview. So even though this is not entirely Nlp course. But I mean, I think, though, if you understand how the Nlp system can be built, then I think it's make it easier how everything has been developed together to toward to the transformer.

144

00:36:34.210 --> 00:36:47.490

Jisun An: So the general framework for the Nfp system is basically you create a function that maps on input X to the output Y where X and or Y involve the languages.

145

00:36:47.700 --> 00:37:12.479

Jisun An: So let's say that we have these different systems where inputs are the text, or it can be also the image and the outputs can be continuing text or text in other language or labels or linguistic structure, or the text. So what kind of system do we build here? So if the input is text. And if the output is the continuing text, what the system do we build?

146

00:37:15.840 --> 00:37:20.810

Jisun An: It's like the next token prediction. So therapy languages.

147

00:37:22.060 --> 00:37:32.309

Jisun An: Sorry autoregressive model. Yes, exactly. So. That would be the autoregressive model which is the language modeling. And if we have the input text, and the output is the text in other language

148

00:37:33.170 --> 00:37:37.349

Jisun An: translation. And and if the output is the label

149

00:37:38.560 --> 00:37:42.090

Jisun An: classification, and if it's the linguistic structure.

150

00:37:42.860 --> 00:37:50.449

Jisun An: so language analysis, there's no no big name for that. And if it's the input, X is the image, and the output is the text

151

00:37:51.020 --> 00:37:54.750

Jisun An: caption, right? So these are some of the Nlp task.

152

00:37:55.490 --> 00:38:16.619

Jisun An: And then the method for creating an Ep system. Once again, there are rule based. So still, maybe for certain system rule based can be easily applicable, and there are prompting, which is basically, you are now using this larger language models and given a particular prompt, you are kind of getting an answer from the input.

153

00:38:16.850 --> 00:38:42.390

Jisun An: And there are also the fine tuning where you are now using the data to tune the model themselves to work better for a particular task. So here, one distinction is that the 1st 2 is basically you don't require any training data to do your task right. You can just having set a rules, or you can just using the to get the output for the fine tuning. You will probably need a train data to train the model.

154

00:38:43.420 --> 00:38:56.840

Jisun An: So let's let's assume that we are making a very simple rule based Nlp system. So and the example task and everyone's probably familiar with is the let's assume that we are building a sentiment analysis system.

155

00:38:56.960 --> 00:39:11.280

Jisun An: So the task is given a review on a reviewing website. So the review is now our input X, and whether it's label Y is positive or negative, or the neutral. So the output is the 3 labels.

156

00:39:11.760 --> 00:39:28.950

Jisun An: So here the input is a review text, and I'll put this sentiment label. So in an example, if I hate this music movie, I love this movie. I saw this movie then the 1st 1. 0, 0, I got messed it up so the 1st one should be positive. Second, one should be the negative.

157

00:39:29.060 --> 00:39:30.026

Jisun An: and then

158

00:39:31.280 --> 00:39:51.069

Jisun An: so the label should be different. But this will be our sentiment analysis. Then what would be the 3 steps of text classification? So what do you usually do is, firstly, you are extracting some features from the input. Right? So you assume that you identify what are the most important parts from this input text?

159

00:39:51.403 --> 00:40:12.390

Jisun An: And then, once you extract these features, you now kind of define a score. Some function that compute the score. Given these features right? And then, based on this scores, you make a decision. So you need also the decision function. So given the score, then you need to now select the most appropriate option, which is one of the 3 outputs

160

00:40:12.710 --> 00:40:14.669

Jisun An: so more formally here.

161

00:40:15.380 --> 00:40:41.580

Jisun An: they just be given X, which is our input. You will have some function that extract the features, and then the score calculation will be a very simple metrics multiplication. So, assuming that the W is your weight metrics and the H is now your feature, and then the multiplication of them will be your score, and then given your score, you will have some decision function where it will end up to the label, either minus 1 0 or one.

162

00:40:42.680 --> 00:40:58.119

Jisun An: And let me give you example. So assuming once again, we are building a very simple, rule-based, sentimental classifier, right? So, assuming that we have list of the lexicons that have we define as a good words or the bad words and then then.

163

00:40:58.160 --> 00:41:16.360

Jisun An: for the feature we can define. 2 features. One is the number of good words. So in the input text, how many words are in part of those good words lexicons and second feature could be number of bad words. So how many of them are actually listed in those bad words?

164

00:41:17.210 --> 00:41:19.809

Jisun An: What was the second step after getting the featuring

165

00:41:20.880 --> 00:41:30.150

Jisun An: scores right? So the scoring function will be super simple, just a linear function where each of the feature has the weights. So these weight

166

00:41:30.270 --> 00:41:47.040

Jisun An: will determine how important each of this feature is right. So they. So we define the score function. So usually we have wait for different feature, and we will have some bias which will be the default kind of score when neither of the features exist, and this will be our score.

167

00:41:47.460 --> 00:41:52.409

Jisun An: and initially, maybe we can set like one for the w 1

168

00:41:54.180 --> 00:42:08.469

Jisun An: minus one for W. 2, and the bias will be like 0 point 5. So you can now see that as the value is more positive, the score will be more positive, and then, as the value is more negative.

169

00:42:10.350 --> 00:42:33.060

Jisun An: I'm sorry, as the f. 1 is greater than the value will be more positive, as the F 2 is getting higher than the value will be more negative, right? So we designed the score function in that way and then. Now, once you have this score, maybe you can have a very simple decision rule where, if the S is greater than 0, then it's the positive. If it's smaller than 0, then

170

00:42:33.440 --> 00:42:38.329

Jisun An: it's a negative, and if it's a 0, then it's a neutral. So this would be our decision function.

171

00:42:38.440 --> 00:42:50.870

Jisun An: And once you decide, then, given what our aim was given X, we get the label right. So now, after these 3 steps, we got the labels so we can evaluate. Maybe the accuracy on the test set.

172

00:42:53.020 --> 00:43:04.289

Jisun An: So these are very simple rule, based sentiment classifier. But obviously, as you know, that for this will not probably give us good results. Then how can we improve this classifier?

173

00:43:06.440 --> 00:43:16.980

Jisun An: And assuming that we have a list of the movie reviews, and we are trying to classify their sentiment. So how can we improve this particular rule based classifier?

174

00:43:17.900 --> 00:43:23.740

Jisun An: Any idea, any idea?

175

00:43:31.650 --> 00:43:55.599

Jisun An: We can add more words in the in the lexicons. Right? So if you think about it, we only have now 6 words in each class. And obviously this is not really only the way that you express the good movie or the bad movie, right? So if it's the rule base that you probably are just increasing the like, the lexicons. Is there any other way that you can improve this classifier?

176

00:43:57.360 --> 00:44:06.120

Jisun An: So, assuming that now you are, you are having some training data, and you train your model and then you kind of get the evaluations

177

00:44:07.890 --> 00:44:10.180

Jisun An: any other way that you can. Yes.

178

00:44:11.380 --> 00:44:21.290

Jisun An: yes, directly. So maybe we can modify the score function, maybe within. In my particular training data. There are far more

179

00:44:21.890 --> 00:44:32.179

Jisun An: positive reviews. So maybe we need to give less weight to the positive values. I don't know what we're going to happen. But basically, you can update the score function themselves to improve this classifier. What else?

180

00:44:37.660 --> 00:44:40.430

Jisun An: Yes, oh, so just yeah.

181

00:44:41.090 --> 00:44:45.710

Jisun An: More than one word out of time. Like, if it says.

182

00:44:45.980 --> 00:44:56.700

Jisun An: Oh, I see. So you're going for like the diagram consideration, that's actually a good idea. Yeah, yeah, yeah, that's

183

00:44:56.970 --> 00:45:13.039

Jisun An: but they will need to change. Oh, but actually, they're actually working. Yeah. So basically, you will see whether that diagram included in in a text, yeah, that's actually a good idea. So that will be part of increasing these lexicons, but in a very interesting way, considering the diagrams as well. Great.

184

00:45:13.340 --> 00:45:28.119

Jisun An: Oh, is it the same. Oh, great, great, yeah. So that's another idea. And maybe the last thing is that you can also update the decision function. Right? So you have now the rule that only split the database on the score 0. But maybe that is actually not a good point.

185

00:45:28.570 --> 00:45:34.339

Jisun An: According to your training data, maybe it's actually less or slightly more. So you can also update decision.

186

00:45:34.940 --> 00:45:45.773

Jisun An: but but the rule based classifier definitely, not the best one. And there are very some difficult cases. 1st one is the low frequency word. So

187

00:45:46.440 --> 00:46:13.650

Jisun An: tedious, mocking glitch. I even don't know what mocking actually means. So there are very, very low frequency words, and if you think about, if you are just adding more and more lexicons, and there will be these words that are very, very low, frequent, but still existing in your training data, and then they may actually lower your performance of the system. So maybe solution for this is that you can keep adding more and more lexicons

188

00:46:13.650 --> 00:46:20.300

Jisun An: to improve the performance of the rule base. But basically your feature will be now just too many.

189

00:46:20.820 --> 00:46:28.899

Jisun An: Maybe you can also incorporate some external sources. Maybe there's already like a sentiment dictionary, and you can just use it as your lexicans therapy solution.

190

00:46:29.050 --> 00:46:54.090

Jisun An: Another problem would be conjugations. So now entertain and entertain me, I mean because these lexicons are exact match. So, even though you have entertain in your lexicons, still they will not be able to capture entertainingly or magnificently, etc. So once again, those conjugated words, compounding words will not be captured in the rule, base or harder to captured in the rule based systems.

191

00:46:54.400 --> 00:47:01.600

Jisun An: the the solution maybe we can use like a root from the pos of post tech,

192

00:47:02.649 --> 00:47:08.040

Jisun An: part of the speech of the world. But this would also require like morphological analysis.

193

00:47:08.970 --> 00:47:20.083

Jisun An: and some other difficulty would be definitely a negative negation. And actually, you guys had this idea of using the diagrams and adding it as a lexicon. So this could be actually solved.

194

00:47:20.780 --> 00:47:48.509

Jisun An: and and yeah, so maybe there could be some other solution, that if just if litigation modifies a word, then maybe you just disregard it. So that would be also a possibility. And also there are some other tricky cases like using metaphors or the analogy. So all these will be something very difficult to be solved by the rule-based classifier. So basically, solution is not very obvious using the rule-based system.

195

00:47:48.700 --> 00:48:08.440

Jisun An: And finally, if it was like the other languages, yeah. Now, the Rube's method, I don't know. Maybe learn Korean to to build this. Yeah, maybe so, Ruby's is not really a perfect system. So people now started to work with the machine learning based. Nlp.

196

00:48:09.280 --> 00:48:33.319

Jisun An: so I assume that you probably familiar with the machine learning, but basically what they are trying to do. We talked about features and the scoring functions right? And so what machine learning is trying to do is basically they learn these features and the score functions from the data. So rather than so you, ruby is basically, you are creating these features, and you are also changing the scoring functions. But what machine learning is trying to do is

197

00:48:33.360 --> 00:48:42.179

Jisun An: determining those 2 from the data, and they have the algorithms to be able to learn the features and these scoring functions

198

00:48:42.340 --> 00:48:45.209

Jisun An: and helps to influence using this model.

199

00:48:46.280 --> 00:49:14.830

Jisun An: So one of the 1st attempt of the machine learning for the building for sentiment analysis system was using the back of words. So now, instead of I mean, it's a kind of similar in a way that in the rule base we were also using the word and trying to match. If the those words are existing in the input or not. So here we are also using word as a kind of basic units of considerations. And now here we consider each of the word as a feature

200

00:49:15.443 --> 00:49:35.469

Jisun An: and then and then, and so that's the reason that they are coded as a back of the words. So we consider all of the words that appeared in the input as a feature. And then we are trying to find the weight for each of these words. And then and then, like. We are kind of having getting the score by multiplying this too.

201

00:49:36.260 --> 00:49:52.820

Jisun An: And so here we already defined the feature themselves, because we are using the word as a feature. But then these weight vectors, so determining how important each of these words are in doing sentiment analysis will be learned from training and from our training data.

202

00:49:53.690 --> 00:50:15.750

Jisun An: And these words, and to be used to be considered each of the word as a feature, we need to represent these words in a vector and the easiest way of representing word is the one hot encoding. So now the each word so assuming that you, your total number, total set of words are n words.

203

00:50:15.750 --> 00:50:35.729

Jisun An: and then so you can assume that you have N columns of the table. And then one hot encoding basically does giving one for the particular word, there you are yourself, and then give all the Zeros to the other self. So now a word, it will be represented by the identity themselves. Right?

204

00:50:36.370 --> 00:50:41.669

Jisun An: So this is cool. That's like the one hot encoding, which is like the very, very simple. So

205

00:50:42.030 --> 00:51:03.272

Jisun An: if you go back to here and now assuming that this is each of them are one hot encoding, and the weight will be. Now, then, if you doing like the multiplication of this row of one vector and the the column of these weights. Then it will basically tell how important this one word is, and that will give the score.

206

00:51:03.980 --> 00:51:26.429

Jisun An: Oh, oh, so so what? So if you are now giving a input and assuming this input has, I hate this movie. Then they will look up from this table the 4 that 4 particular word, and extracting those 4 lows rows, and then multiplying by the weight. And then we give the score, and that will be the whether they will be like the positive neutral or the negative.

207

00:51:28.330 --> 00:51:31.199

Jisun An: Does this make sense any questions here?

208

00:51:34.540 --> 00:51:53.625

Jisun An: So now, if we are doing this bag of wars? But then I I guess maybe for some of you, if we haven't done really the training, then this could be a little confusing. But assuming that you have a training data, and then, after the training, what this model will learn will be

209

00:51:54.790 --> 00:52:13.629

Jisun An: the list of the learns that list of the words that was in our training set. And the weight factor will tell how important each of these words are. So that's like the basic, very basic thing that you will learn after the training. And we will talk more about this training next week. But for now, just to have it, just keep it in that way.

210

00:52:14.240 --> 00:52:19.910

Jisun An: So assuming that we are doing that, what problem mentioned before would this solve?

211

00:52:20.040 --> 00:52:43.240

Jisun An: So we talked about the conjugations, low frequency word, different languages, analogy negation. So out of these 5, if we, instead of the rule-based classifier, if we apply this backwards approach, and after the score. It will be the same right. We will have the decision function, and that will give us label, What will this

212

00:52:44.700 --> 00:52:47.219

Jisun An: solve any of those 5.

213

00:52:58.200 --> 00:53:04.619

Jisun An: So let's think about that one by one, would it? Would it? So? The low frequency were the problem.

214

00:53:07.400 --> 00:53:36.349

Jisun An: maybe partially, maybe not entirely. But but at least you don't need to add low frequency word one by one to your lexicon. Right? So this model will learn. If that low frequency word is important enough, then they will add it. I mean they will already include it in the feature, or and also they will be already in the feature set, because we will consider all the words in the training as a feature, but just the weight will determine how how important that word is. Right. So it's already in it.

215

00:53:36.350 --> 00:53:39.379

Jisun An: So it's kind of soft, but not entirely.

216

00:53:39.430 --> 00:53:41.980

Jisun An: And what about the second one? Conjugation

217

00:53:48.170 --> 00:54:16.110

Jisun An: may or may not it also? Not entirely so. Because, because in the test, that there may be like other conjugation, that they never appear. And also these conjugation also can be very low frequency. But once again those compound words that are frequently appear and important, that will be part of the feature, and then we also have the high weight, so that part also partially can be solved. Not entirely. Once again. What about the negation?

218

00:54:19.950 --> 00:54:36.696

Jisun An: Probably not. And because this one bag of words? It doesn't consider the order of the word or the combination of the words, they will use just a single word as a feature. So this this will not. There's no way that they can distinguish between. I don't. I do not hate, and I do not love

219

00:54:37.330 --> 00:54:39.389

Jisun An: And then what's the analogy?

220

00:54:40.430 --> 00:54:46.520

Jisun An: No way right. There's no way that analogy can be solved with this model, and the last one was the

221

00:54:46.630 --> 00:55:07.109

Jisun An: other language, I mean, yeah, they cannot be solved here. Right so so. But but at least can handle that those painful procedure where you need to pick the words and then adding to the lexicons. At least this was help reserving that, and a little bit of the conjugation problem.

222

00:55:08.150 --> 00:55:35.009

Jisun An: And then these weight vectors, what eventually this weight vector represents. So if it was binary classification, assuming there was only the positive and the negative. Then this will be each word will have a singular scholar, where the positive, indicating the yes and the negative indicating no. So the weight vector directly will tell you how important this value is to

223

00:55:35.230 --> 00:55:57.895

Jisun An: being a positive label or the negative label. And if it was like the multi class. Then you will now each word have different value for different class, so the love can be, have positive value for the verified positive and the positive, and then have a negative value for the negative and very negative, and the neutral, etc. Etc. So this will be

224

00:55:58.510 --> 00:56:16.290

Jisun An: Now, what the weight factor. After that we train this bag of word model, the weight vector will have. And this can actually give you some idea of how each of this word is relating to the sentiment analysis. So this is something that what you can expect from this vector representation

225

00:56:18.540 --> 00:56:21.469

Jisun An: than what's missing in this model.

226

00:56:22.810 --> 00:56:47.760

Jisun An: So, firstly, the conjugated or the compound word, I love this movie versus, I love this movie. Once again it was handling partially, but it was not perfectly handled, and also the handling of the word similarity. So I love this movie. Or I adore this movie. They are basically the same thing and should be positive. But this model basically cannot really handle those cases. If the adore is not in the training data.

227

00:56:47.760 --> 00:57:07.680

Jisun An: then this beggar words will not learn that adore is relating to the positive review, but we know that adore is similar to the love, and then those it should be also relating to the positive review. So if a word doesn't occur in the training data. Then beggar words will not ever learn that these words are relating to positive reviews

228

00:57:08.220 --> 00:57:17.869

Jisun An: and then handling of the combination words. So like, I love this movie versus, I don't love this movie. There's no way that this bag of word model will handle that. Those cases.

229

00:57:18.060 --> 00:57:30.250

Jisun An: And lastly, the sentence structure, which is, it has an interesting story, but is a boring overview. So basically, these are the longer structure, and this will not be able to hinder.

230

00:57:31.520 --> 00:57:54.349

Jisun An: So this conjugation, so the compound word can be solved with the subword models, which is another word like tokenization. Probably you've heard of it. This word similarity can be solved with the word embeddings, and then the combination of the features is now, if you in instead of the back of word. If you are now using neural networks, then these can be

231

00:57:54.350 --> 00:58:16.769

Jisun An: so neural. Networks basically can help to find which features are together, works for the classifying something. So now that will be helping you to understand what features are related and those combinations of feature is possible to be considered. And the lastly, the sentence structure, will be possible to get solved with the sequencer model.

232

00:58:17.220 --> 00:58:22.540

Jisun An: So these are the 4 components that we will talk today and next week.

233

00:58:23.500 --> 00:58:25.440

Jisun An: Okay, any any questions?

234

00:58:26.250 --> 00:58:27.139

Jisun An: Oh, yes.

235

00:58:28.700 --> 00:58:41.939

Jisun An: Tf, idf, tf, idf is a way to in. So in the back of world model. So for now we are just using one to indicate that

236

00:58:42.050 --> 00:58:53.249

Jisun An: the word is existing. So so basically, we are giving the same weight to every word right? But then maybe you can use tfidf value to instead of just a better word.

237

00:58:53.630 --> 00:59:14.824

Jisun An: So they assume, you know, representing words way. I think it is a similar. It's a still give like a very sparse, but it's a similar to bow, and even not going to handle most of these problems. But tfi def is the way to measure the importance of the word in the corpus. So

238

00:59:15.490 --> 00:59:26.910

Jisun An: eventually training data. After the training, you will know which word is more or less important, Tfidf is basically another way to find the importance of the word. Yeah.

239

00:59:27.300 --> 00:59:37.829

Jisun An: But thanks for the question, that's all that should go back. Yeah. I actually thinking whether to have it in the slide or not, and I didn't edit it. But that was good. Yeah, anyhow. So another question.

240

00:59:40.160 --> 00:59:48.649

Jisun An: right? I hope I can do this software model, because it's not very I I know I told you this will be very lecture

241

00:59:48.670 --> 01:00:16.502

Jisun An: like, yeah, it'll be very lecture oriented for the 1st few lectures. So let's talk about the subword model. Also. In other words, it's also called as a tokenization. Once again, this is, we trying to aim the problem of the conjugation that these bow backwards model just cannot handle them, because they, looking at, I mean, they handle to certain extents, but not perfectly. And then how can we? And also there's another problem.

242

01:00:17.530 --> 01:00:21.969

Jisun An: so if we create like a feature vector using at word level.

243

01:00:22.250 --> 01:00:24.109

Jisun An: How many more do we have?

244

01:00:24.530 --> 01:00:48.609

Jisun An: Right? It's I don't know, actually, how many words do we have? But it's a lot, right? So your feature vector will increasing a lot, a lot. So basically you will have. I told you that when we creating this bag of words word embeddings, you need all the columns, that number of the words that you have. Right? So if you're training Corpus becoming larger and including more and more corpus, then your vector will be. Now.

245

01:00:48.610 --> 01:00:57.430

Jisun An: columns will be increasing like a lot. So let's assuming that you have like a million columns, and you only have one for your own word and the 0 for all the other

246

01:00:57.754 --> 01:01:11.709

Jisun An: features. Then basically, this will be very sparse metrics. And when you have these sparse metrics, firstly, it is very income inefficiency for your memory, for your computer. And also it will be very taking long time to compute them as well.

247

01:01:12.010 --> 01:01:19.789

Jisun An: So basically having every words as a feature, basically, is not a good strategy to have it as a model.

248

01:01:20.140 --> 01:01:23.099

Jisun An: And the forward model trying to handle that issue.

249

01:01:24.084 --> 01:01:43.639

Jisun An: So the subword model. Basic idea is that we split less common words into multiple subword tokens. So, for example, if the companies are expanding, then we are splitting this this word into the tokens of company and ies, or expanding an ing.

250

01:01:46.870 --> 01:01:48.420

Jisun An: And more importantly.

251

01:01:48.770 --> 01:01:58.639

Jisun An: there are these method that we're going to introduce. Basically, they use the data to tell us how to tokenize. So what would be the best way to tokenize these words

252

01:01:59.756 --> 01:02:26.013

Jisun An: the benefits. So we will be able to share the parameters between words variance and then also compounds words, and also we reduce the parameter size and save the compute and the memories, as I told you. So, so basically those so basically, we will only need to keep just the expense. And then every other word. If all the conjugation words can be shared like ing or ies, so those more

253

01:02:26.440 --> 01:02:30.070

Jisun An: morphemes will be handled well with these models

254

01:02:31.210 --> 01:02:43.330

Jisun An: so subword model. There are 3 common methods. The 1st one is the bi pair encoding and word piece, and the unigram language, modeling tokenizations, and these are the most common ones.

255

01:02:43.330 --> 01:03:01.550

Jisun An: and these submodels usually have 2 parts. One is the the learner, part, token learner, part where we take the training course, and from them we decide which vocabulary should be included, meaning that so which features we will have in our model.

256

01:03:02.240 --> 01:03:26.629

Jisun An: And then the segmentary send. Now, giving that our model came up with the how to tokenize the word. Then basically given a any role test sentences, they can split the sentence into the tokens. So firstly, you need to learn how to tokenize the word, and then once we have the model, then we can given any, input. You can tokenize them based on these models.

257

01:03:27.150 --> 01:03:38.410

Jisun An: So let me begin with the byte pair encoding, and this is used a lot in various transformer models, including Gpt. Gpt. 2 Roberta Bart and the D. Berta as well.

258

01:03:38.540 --> 01:03:39.650

Jisun An: And then

259

01:03:39.930 --> 01:03:52.868

Jisun An: then we basically start from vocabulary be the set of all individual characters. So Abcd or Abcd, so basically, they are, start with the characters. And then they repeat,

260

01:03:53.690 --> 01:04:02.419

Jisun An: you basically look at your training data and then choose the 2 symbols that are most frequently urgent audit in the training corpus.

261

01:04:03.220 --> 01:04:09.040

Jisun An: and then they assume that the merged version, A B is your new vocabulary.

262

01:04:09.950 --> 01:04:17.679

Jisun An: and then you replace in the training corpus. You replace all the A BA, and B to Ab.

263

01:04:18.350 --> 01:04:25.819

Jisun An: and then you just repeat this until until until certain. And this rule can be. Maybe K merges.

264

01:04:26.450 --> 01:04:50.510

Jisun An: I will show you more examples. No worries. So I mean, this is the official algorithm. You can have a look in their paper. Some of the few is that so so usually the support algorithms are run inside a space separated tokens and then we also commonly add the the underbar at the end of the word, so that we know that that's the end of the word.

265

01:04:50.810 --> 01:04:58.940

Jisun An: and also and then letters, and then we separate the letters. So that, like the how this Ppe algorithm starts with

266

01:04:59.540 --> 01:05:08.000

Jisun An: so so this was the the example that showed in the original paper. So, assuming that our original Corpus is very small, we have

267

01:05:08.040 --> 01:05:37.070

Jisun An: about 20 words, so low, low lowest voice, newer, newer, etc. So the 1st step that we were going to do. We assume that once again we will have the underbar at the end of the each of the words. So that's the one thing. So we will start with the character level. So from these we assume that we have vocabulary of Underbar DEIL NORS. TW. So these are separate characters, and these are our current vocabulary

268

01:05:37.810 --> 01:05:44.659

Jisun An: and our corpus. So instead of having this list of the words, we simply count it. So we have 5 low.

269

01:05:44.770 --> 01:05:51.509

Jisun An: 2 lowest, 6 newer, 3 wider, and 2 new, and once again at the end of each of the work. We have this underbar.

270

01:05:51.830 --> 01:05:58.580

Jisun An: So the next step, what we do is we look at the the 2 characters that are most frequent.

271

01:05:59.320 --> 01:06:03.860

Jisun An: Anyone can catch a lot of be that Oof.

272

01:06:06.210 --> 01:06:08.470

Jisun An: W. No. 2 characters.

273

01:06:10.620 --> 01:06:18.270

Jisun An: R rw, oh, it should be consequent, consecutive, consecutive, sorry. Yeah. Consecutive.

274

01:06:21.380 --> 01:06:22.429

Jisun An: ER.

275

01:06:23.190 --> 01:06:27.440

Jisun An: So we have 6 newer, 3 wider. So we have 9 er in total

276

01:06:27.550 --> 01:06:39.790

Jisun An: right? So basically from the corpus. Once again, you look for the 2 consecutive characters that are most frequent, which in this case will be er so we will merge enr into the er

277

01:06:40.666 --> 01:06:45.030

Jisun An: and then our vocabulary will now have new vocabulary er.

278

01:06:45.800 --> 01:06:53.819

Jisun An: and after that we will update our corpus by merging this. So all the newer, wider er became er

279

01:06:54.500 --> 01:07:04.190

Jisun An: so this will be the what Ppe will do with their training data to learn which vocabularies are most commonly happen, and we will repeat this

280

01:07:04.240 --> 01:07:33.440

Jisun An: in the next run. So now the 1st one is the one that we just did with the er, and the next one is the er and underbar is the 2 consecutive vocabulary. Sorry not the character anymore. 2 consecutive vocabulary that are most frequent, because er was 9 times so er underbar should be the next one. So what they do is adding this er underbar to the vocabulary, and then they also updated the corpus. The following one will be Ne. So they do the Ne.

281

01:07:33.440 --> 01:07:47.060

Jisun An: So now you get the idea right, and they repeat to end merges until they basically found. they decide like as a last merge and low underbar. And this will be now our current vocabulary.

282

01:07:47.530 --> 01:07:59.570

Jisun An: So this is the learner part. So based on this. Now we learn that, given our training data, these vocabulary will be the set of the most frequent subwords.

283

01:08:00.760 --> 01:08:09.359

Jisun An: Then, now, there will be segmental algorithm. So given. Now, we have this vocabulary. How should we tokenize? Given any any given input

284

01:08:10.440 --> 01:08:38.659

Jisun An: on the test data. Now, we run each merge and learn from the training data greedily, meaning that we will apply the rule from the start. So there's an order of this vocabulary that we added, and we just merge as we see the 1st the 1st rule that we see we will apply to them. And once again, the test frequency will not play any role the the because run from the training and the test run will not be really considered here.

285

01:08:39.270 --> 01:08:40.950

Jisun An: So one example. So

286

01:08:41.649 --> 01:09:06.735

Jisun An: as as long as as soon as the if the text input had e and r, then they will merge er first, st and then they will see the er underbar to er underbar, and etc. Etc. So here, if we have, like the newer in our test set. Then that would be tokenized as a forward, because we actually had newer in one of our

287

01:09:08.093 --> 01:09:14.333

Jisun An: vocabulary. But it will start with the ERER, the underbar, and then Ne and then

288

01:09:15.660 --> 01:09:21.040

Jisun An: and then they will meet the newer, and then they will. So the newer underbar will be merging together.

289

01:09:21.210 --> 01:09:39.399

Jisun An: But then, on the other hand, if they had lower then, because they will merge low and erw. But then, if our vocabulary didn't, because our vocabulary didn't include the lower underbar. It will just stop there. So this will be our output of the tokenization

290

01:09:40.439 --> 01:09:41.180

Jisun An: clear

291

01:09:44.071 --> 01:09:59.178

Jisun An: the properties of these ppe tokens so they usually include the frequent words and frequent subwords, and many of the morphism morphemes like East or er will be will have those subwords.

292

01:10:00.676 --> 01:10:01.869

Jisun An: so like

293

01:10:02.100 --> 01:10:09.270

Jisun An: the morphine, is like the smallest meaning bearing unit of the language. Unlike list has the 3 morphine unlikely East.

294

01:10:09.690 --> 01:10:34.580

Jisun An: So, but Bp kind of will capture them. But at least, if you think about in compared with the the number of the words in your training data, set these, a software model will have far less number of vocabies. Which means that your feature vector will be far more smaller. So if you have, input and if you now change that into the embeddings, then it will be far smaller.

295

01:10:34.650 --> 01:11:02.450

Jisun An: So the word piece is also very similar to Bpe. The only differences is that instead of selecting the most frequent pair, they will compute a score for each pair using this formula. So if you see that it's the frequency of the pair, divided by the 1st frequency of the 1st element, multiplied by the frequency of the second element, so that will now incorporate how important each of these 2 words, and then if we are

296

01:11:03.210 --> 01:11:07.819

Jisun An: if and then basically so, if the if so, maybe

297

01:11:08.380 --> 01:11:25.440

Jisun An: the merging may not be very worthy, then keeping them separate. So they just want to incorporate that. So instead of just using the free frequency themselves, they just normalize by the importance of each of these award. And so they just instead of using that, they will use this formula

298

01:11:26.310 --> 01:11:33.179

Jisun An: and then select those longest sub award that that that have the

299

01:11:34.212 --> 01:11:49.519

Jisun An: highest score of this function, and then they will do the merge. So the the way that they will select the vocab vocabulary will be similar to what Ppe do, but it will be based on these scores.

300

01:11:50.090 --> 01:11:54.900

Jisun An: and the unigram model is something slightly different. And

301

01:11:55.080 --> 01:12:24.679

Jisun An: now this will consider the entire vocabulary. So they will basically build a unigram model themselves, which you may not be very familiar with at the moment. But unigram model is basically, you are counting the probability that you're observing that word in a training corpus. So you'll be simply like the proportion of the word across the entire training corpus, and then you are basically picking a vocabulary, then maximize the local likelihood of the corpus, meaning that

302

01:12:28.750 --> 01:12:40.049

Jisun An: So if if you are taking those vocabulary, and if the entire corpus has higher kind of

303

01:12:40.870 --> 01:12:46.590

Jisun An: probability, then that will select them, and then adding it. So

304

01:12:46.790 --> 01:12:54.581

Jisun An: we have too many left, and I think that's the reason that I'm starts kind of mumble. So I will. Actually, we only have like a few slides

305

01:12:54.990 --> 01:13:03.530

Jisun An: left. But I will probably just continuing from here because I don't want to rush it. Things in 2 year so up to here. Oh, do you have any questions

306

01:13:11.170 --> 01:13:12.380

Jisun An: you mean here?

307

01:13:24.600 --> 01:13:28.430

Jisun An: Oh, because wow!

308

01:13:31.100 --> 01:13:35.260

Jisun An: Because the er now considered to be one character.

309

01:13:36.020 --> 01:13:38.920

Jisun An: so er underbar is the another character.

310

01:13:42.240 --> 01:13:43.110

Jisun An: Oh.

311

01:13:54.116 --> 01:14:02.840

Jisun An: and then, once our one again. Erc is one character and spaces one character, and that's the most common 2.

312

01:14:03.120 --> 01:14:08.080

Jisun An: Next, that is vocabulary, because everyone

313

01:14:08.920 --> 01:14:15.550

Jisun An: so it gets added on every month it gets added on. And then.

314

01:14:15.870 --> 01:14:22.729

Jisun An: when we start the training, every word will basically assume that there will be underbar attached to every word.

315

01:14:23.810 --> 01:14:26.419

Jisun An: So that's our starting training purpose.

316

01:14:29.850 --> 01:14:31.359

Jisun An: any other quick question.

317

01:14:33.870 --> 01:14:38.180

Jisun An: Okay, I will resume from the unigram model, which?

318

01:14:38.290 --> 01:14:50.729

Jisun An: Yeah, we have a lot of content. So let's see. But I, I, yeah. Yeah, thanks a lot for joining today, and I will see you next Tuesday. Have a great rest of the week, and I will see you next week. Thank you.