WATCH · SANSKRIT COMPUTING / STUDY NOTES

Dharmamitra: Sebastian Nehrdich (UC Berkeley)

Study notes by Sanskrit AI Universe, based on this video

Video by Sebastian Nehrdich (UC Berkeley) · Heidelberg - KRS · 24 Nov 2024

Not the creator's own slides

Sanskrit AI Universe

About these notes

— Official title: Dharmamitra: Sebastian Nehrdich (UC Berkeley)

— Speaker and channel: Sebastian Nehrdich (UC Berkeley) · Heidelberg - KRS

— Published 24 Nov 2024 · Running time 1 h 07 min · https://www.youtube.com/watch?v=SMD6RGr-w9Q

— Official description: Sebastian Nehrdich will talk about Dharmamitra's Sanskrit-specific capabilities, particularly our new language model that achieves state-of-the-art accuracy in fundamental Sanskrit processing tasks such as word segmentation, lemmatization, and morphological analysis. He will demonstrate how these technical advances translate into practical tools for Sanskrit scholars from assisting in basic text analysis to enabling sophisticated corpus-wide semantic search and machine translation. The talk will

— These pages are study notes drawn from the video's own title, description and captions, plus the journal's already published Watch note. They are not a transcript and not a substitute for watching the film.

— The journal treats videos as leads, not as sources you should cite on their own. Check the papers and editions behind a claim.

— Caption source: English (auto-generated). Auto-captions can mis-hear words; obvious proper-name spellings were given in IAST where the journal already uses them. Distinctive phrases follow the captions closely.

From the official description

— Sebastian Nehrdich will talk about Dharmamitra's Sanskrit-specific capabilities, particularly our new language model that achieves state-of-the-art accuracy in fundamental Sanskrit processing tasks such as word segmentation, lemmatization, and morphological analysis.

— He will demonstrate how these technical advances translate into practical tools for Sanskrit scholars from assisting in basic text analysis to enabling sophisticated corpus-wide semantic search and machine translation.

— The talk will showcase examples of how our system can provide detailed grammatical explanations, annotated transtations, and tacilitate textual research via semantic search even across language poundaries.

— These tools are designed to serve poth beginning Sanskrit students and advanced scholars conducting specialized research.

— He will also demonstrate how Dharmamitra's capabilities can be used as building blocks for Sanskrit digitization and annotation projects.

From the talk · 0:09–7:13

— This already is a very good hint at the fact that this is really a topic that is of great relevance to all of us I mean especially with.

— In Sanskrit at Oxford University actually I think in the in the early 80s and he later switched into AI research at a time.

— When AI research was not yet as fancy as it is nowadays so he has done it in the late 80s in the '90s and throughout the 2000s so.

— The models that we have and so on so we are very happy for that and then J sandan of course some of you might know him he has.

From the talk · 7:16–14:03

— And it will be released in the next week or two where we on the one hand provide the machine translation but then also a grammatical analysis.

— The other languages it's a bit you might try that but be aware that maybe the performance will not always be as good we offer grammatical explanations for Sanskrit.

— But it at least gives us a rough direction of where the model is heading and the score here is chrf which is a character based comparison score.

— Also confirmed by user feedback so far that those people that I have talked with prefer to use claw over Google translate or gbt as well.

From the talk · 14:04–20:56

— For Sanskrit grammatical analysis by a considerable margin so we have done a segmentation paper in 2018 and I think.

— The error rate by 50% and this leads us to an errow rate that is roughly as high as a single human annotator so.

— If you ask a human to do a grammatical annotation of a Sanskrit text somebody who is decently educated in Sanskrit what you get and what.

— This model is producing is somewhat on a similar level of course if you do revision if you have multiple annotators that is a different story but.

From the talk · 20:59–27:20

— These kind of systems so we were very fortunate that the Kumar Jaa project which is working on translating the Tibetan Buddhist scriptures into Chinese but they.

— How other people feel about this I'm sure not everybody is really benefiting equally from this and that's totally fine and yeah we are very happy to just hear.

— What people are making out of this system coming to the semantic search part which is not yet officially released and I think the timeline for.

— So that's that's how I actually got into contact with Oliver back then in order to see what we can do to design algorithms that can help in.

From the talk · 27:22–34:04

— That go into or that drive the development of this search engine and our solution that we are proposing for this is the Metra search it has an interface.

— Here to enter a text you can do some additional filtering by Language by category and so on so even down to the level of individual text.

— The moment operates so you can enter a query in English famous first verse of the of the bhavat Gita is the example that I choose here.

— What Chad gbt for example does because CLA and C gbt don't have that ability to retrieve specific passages and to actually lead the user to the source text.

From the talk · 34:07–40:32

— If the input query is in any other language then we first need to make sure that we have the same language to search in and then.

— We combine a lot of different search algorithms as well as different databases here so we do a traditional bm25 Sur directly on the Sanskrit data we.

— That kind of search on machine translated versions of the text and in that way we really have different Nets so to speak like a fisherman.

— The standards of editions does vary a lot so some editions follow Western punctuations in other cases we only have the Dand as it is done in.

From the talk · 40:34–47:29

— Operations like coine distance in order to determine how similar two paragraphs are which is of course very handy because then we can do standard mathematical methods.

— So in a sense it really serves as the base as the basis of this project it is in size 5.7 gigabytes that's roughly 7 billion tokens.

— For Tibetan in this case and the ca collection for Chinese and then we have done our own efforts on creating a large Sanskrit data set so.

— We have combined different web resources like gret digital Sanskrit Buddhist Cannon as well as ocrs that we have done ourselves also in collaboration with AI for B.

From the talk · 47:33–54:06

— This on a cluster of 8 a100 gpus the training took two weeks for one Epoch so for two passes of the data set takes one month if.

— This is something that really worked for our purposes so we have our recipe is similar to that it's not identical but it's similar to that we have.

— The training data into English that we are using in order to teach the model how to translate Sanskrit into English and other languages into English so that's.

— Also have paril data for other Target languages and I think this is for philological purposes the very interesting data set so based on our algorithmic Solutions.

From the talk · 54:08–1:00:44

— This is where we have this Metra instruction data set here which is used in the last stage so after the pre-training for four weeks.

— If you want to translate a whole page of a book or more than that then we also have instructions of that in our data set then we.

— Those examples that we provide and then suddenly the model will show those improvements in Behavior so that's a very use ful capability and then dama Mitra.

— We create a draft machine translation with our model and then what we do next we also take this srit input sentence and use the search engine.

From the talk · 1:00:46–1:07:06

— Play around with this llm and see how it performs you can actually download the model on hugging phase we have our organization there which is called Buddhist dnlp.

— The user can look into the database and see what kind of similar passages we have this is something that actually people requested quite a lot so far and.

— And grammar checking of the input I do know from my own experience that it's not easy to type Sanskrit without making mistakes in the diacritics especially.

— What AI is and what it isn't and whether it's sentient or not it's all very philosophical I might not be the right person to answer that.

Sources and credit

— Study notes by Sanskrit AI Universe, based on the video “Dharmamitra: Sebastian Nehrdich (UC Berkeley)”. These are not the creator's own slides.

— Video by Sebastian Nehrdich (UC Berkeley), posted on the YouTube channel Heidelberg - KRS.

— Watch the video: https://www.youtube.com/watch?v=SMD6RGr-w9Q

— Captions used: English (auto-generated) (en), retrieved on 24 November 2026 via a public Invidious instance. Auto-captions can mis-hear words; obvious proper-name spellings were given in IAST where the journal already uses them.

— Also used: the official video title and description, and the Watch catalogue note already published in this journal (presenter and why the video is listed).

— If the Founding Editor later supplies his own deck for this video, replace the PPTX and PDF in src_assets/slides/ (see the journal README, “How to swap in RK's own slides”).