WATCH · SANSKRIT COMPUTING / STUDY NOTES

Data Driven Approaches for Sanskrit — University Of Cambridge CST Seminar Series — Dr Amrith Krishna

Study notes by Sanskrit AI Universe, based on this video

Video by Amrith Krishna (University of Cambridge CST seminar) · Amrith Krishna · 26 Aug 2023

Not the creator's own slides

About these notes

— Official title: Data Driven Approaches for Sanskrit | University Of Cambridge CST Seminar Series | Dr Amrith Krishna

— Speaker and channel: Amrith Krishna (University of Cambridge CST seminar) · Amrith Krishna

— Published 26 Aug 2023 · Running time 47 min 10 s · https://www.youtube.com/watch?v=zTmkTghSEXo

— Official description: Seminar talk delivered by Dr Amrith Krishna at the University of Cambridge CST Seminar Series, based on his published work - A Graph-Based Framework for Structured Prediction Tasks in Sanskrit We propose a framework using Energy-Based Models for multiple structured prediction tasks in Morphologically rich free-word order languages, with a focus on Sanskrit. Ours is an arc-factored model, similar to the graph-based parsing approaches, and we consider the tasks of word-segmentation, morphological

— Catalogue caution: Posted on the speaker's own channel; few views so far.

— 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

— Typically the state of the art models for morphosyntactic tasks in morphologically rich languages still rely on hand-crafted features for their performance.

— The feature function so learnt along with the search space we construct, encode relevant linguistic information for the tasks we consider.

— This enables us to substantially reduce the training data requirements to as low as 10 \% as compared to the data requirements for the neural state of the art models.

— Our experiments in Czech and Sanskrit show the language agnostic nature of the framework, where we train highly competitive models for both the languages.

— Moreover, our framework enables to incorporate language specific constraints to prune the search space and to filter the candidates during inference.

From the talk · 0:00–4:52

— A morphological rich language or a free word or language which is the case with quite a lot of languages so let's see yeah.

— So sanskrit has been influenced by fairly advanced disciplines of prasadi phonetics and grammar which has been existent since 2500 bc and several of.

— So sanskrit is a relatively free word order language it has rich morphology rich compounding rich compounding and the words often perform santi that is.

— And because of that a lot of the texts are written in words format so this results in a completely different set of requirements for processing.

From the talk · 4:55–9:48

— The free word orderness to adhere to the meter constraints and in this process the verbal cognition takes a backseat sanskrit sentences also express santi sandhi is nothing but.

— This happens at a much larger scale in sanskrit in sanskrit yeah for example if we consider the sequence guru alambana here it is a sandhi of two possible words where the.

— First word is guru and the second is alan benner so the o in guru gets converted into va when we perform something similarly in the case of vidya.

— This is merely a phonetic level transformation the generation is unambiguous but the analysis is not as we could observe from the third and fourth case so these are the tasks that.

From the talk · 9:50–14:26

— Input representation we have extremely low requirements of task specific training data as compared to the neural counterparts so our structure prediction framework consists of consists of three separate components.

— The structure prediction model once we have that this arc factored model basically learns a function to score the edges of the graph and our inference basically searches.

— So basically our input is a graph and hence we stick to ebms for this very specific reason that they enable non-probabilistic training or structured models and this avoids.

— The need to normalize over our output space output space which is the case with crf so the easiest way to think about this model is in terms of the.

From the talk · 14:27–19:06

— So if you observe here there are characters that were not originally present in the input sequence so.

— We generate the edges and we pass it to our energy based model and for word segmentation we again use maximal click as our inference so inference so now.

— I will probably go into much detail into each of these tasks that we are addressing in this framework so word segmentation is the task where we identify.

— The seven cases result in phonetic transformations so what you see here in orange are the changes that happens due to santi so here.

From the talk · 19:08–24:05

— We passed this sentence into the like into sanskrit heritage reader it provides us with all possible analysis in terms of sandy and morphology one morphological analysis so with.

— That so if we know that a particular nominal is in ablative case then so you could see the case of a nominal in the y axis so.

— The verb so this considerably reduces the label space label space when we have to do the prediction we also perform the task of synthetic linearization so.

— We have previously mentioned that in sanskrit there are a lot of verses where the word order are often not given i mean the word order is done.

From the talk · 24:08–29:00

— So depending on the sun the operation of sandy the words the meter of certain syllables change so when we write it in.

— If they are placed successively and again i am mixing up terms from phonetics or phonetics or prosody and grammar it's because in science cut we follow phonemic orthography so.

— That is why sometimes sandy becomes essential when it comes to prosody level tasks yeah so the basically it's a two-step formation like you take a bag of words.

— There any queries so far shall i proceed yeah it seems really interesting just yeah i'm happy sure okay.

From the talk · 29:03–33:31

— The principle of dependency locality and this has been defined as sanity in sanskrit proximity so the learning and the inference methods are well known approaches since 2005.

— This is important from a point that for many morphologically rich languages people still use still use handcrafted features to get state-of-the-art results in more for synthetic parsing or even.

— We use grammatical categories such as case number tens etc and this is important here because sanskrit is a functional language so by fusion language fusion language.

— The engrams however not all co-occurrences are useful are useful that's why we use we need to identify which of those co-occurrences are useful and we basically obtain.

From the talk · 33:32–37:56

— The set of possible co-occurrences by conditioning them on morphological constraints so the constraint can be a morphological tag or it can be a grammatical category so to represent this.

— We use only one constraint in a feature but it you know it exponentially blows up and we can get more than eight hundred thousand possible features if.

— We use these feature templates to get our feature vectors so currently our model uses about 850 such features and how do we identify.

— And used forward stage path generation approach so here both the task essentially what it tries to do is.

From the talk · 37:59–42:39

— So they their model achieved a better performance in terms of perfect match like how often you get a perfect match at a sequence level so.

— We use again a beam search based inference and we obtain a blow score of 59.38 and this is this is an improvement there is a close to.

— For more than four blow score improvement in comparison to kavya guru kavi guru was the baseline we built which was published in acl 19. so this.

— For prosodic linearization we find that again our model performs much better as compared to bso or caviar guru lineal stm is again.

From the talk · 42:41–47:08

— The joint test we basically compared with different combination of our models so in the interest of time probably i will skip this is the performance.

— The biofin were coming close with 84 or 84.36 uas but our model was somewhere around 87.46 around 87.46 for the join morphosynthetic parsing task.

— Another area that we are interested in exploring is how we could formalize the morphological constraints into a logical formulation and probably use them.

— In a as a loss function or as a part of loss function so the case and the synthetic relation how can we make use of it currently.

Sources and credit

— Study notes by Sanskrit AI Universe, based on the video “Data Driven Approaches for Sanskrit | University Of Cambridge CST Seminar Series | Dr Amrith Krishna”. These are not the creator's own slides.

— Video by Amrith Krishna (University of Cambridge CST seminar), posted on the YouTube channel Amrith Krishna.

— Watch the video: https://www.youtube.com/watch?v=zTmkTghSEXo

— 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”).