Supervisor's Review of Diploma Thesis

Thesis title: Empirical Models for an Indic Language Continuum
Author: Niyati Bafna
Supervisor: doc. Ing. Zdeněk Žabokrtský, Ph.D.

Thesis description

The aim of the work under review was to shed more light on into how language continua could be handled in modern NLP, with a particular focus on the continuum of languages spoken in North India and cognate induction as a selected case study application.

The thesis is structured as follows. After the introductory (unnumbered) chapter presenting the overall picture, Chapter 1 follows which summarizes the language situation in North India, and lists 26 languages (mostly Indo-Aryan ones) included into the study and tentatively divided into Band 1, Band 2, and Band 3, depending on the amount of existing data resources. Chapter 2 presents a general review of related work (however, additional passages describing more specific areas of related work appear later in the text too, e.g. in 5.2 or 9.2). Chapter 3 describes a crawler that gathers multilingual data from an existing collection of folksongs and poems for most of the languages under study; a specific crawler has been developed, as a certain reverse-engineering tuning was needed to minimize the amount of noise and to classify languages correctly; basic quantitative features of the downloaded data are described in this chapter too. Chapter 4 presents more detailed statistical analyses (probes) of the downloaded data, revealing character-level, word-level as well as subword-level overlaps within the set of languages. Chapter 5 introduces the task of cognate induction (from created multilingual collection) and outlines possible directions to deal with it, including two simple baselines. In Chapter 6, an EM-based algorithm, potentially capable of learning a more fine-grained model of orthographic distances (based on giving different EM-trained weights to different character substitutions). Chapter 7 shows a different approach to the same task, this time based on similarity of word embeddings. Chapter 8 presents how an independent evaluation data collection was assembled from an online language learning website. Chapter 9 outlines evaluation of different approaches to cognate induction, using the evaluation data. Finally, the concluding chapter summarizes contributions of the thesis.

Including references, lists of figures and tables, and a short appendix, the thesis contains 90 pages. There is no data storage medium attached to the thesis, as all experimental source codes are available at github.

Evaluation

It is striking that not only there is basically no NLP technology available for many North-Indian languages nowadays, some of which have millions or even tens of millions of speakers, but even most basic resources such as monolingual corpora are missing for some of them. The lack of any reliable data (which, honestly, seemed almost unbelievable to me at the beginning, given the speaker population sizes) was the main challenge for the presented work. More specifically, Niyati had to face quite a few highly non-trivial obstacles, from which I choose three random examples. First, not only that no reasonably sized data were available for developing machine-learning models of the language continuum, but there were basically no data even for evaluation. Second, there was no reliable language identifier that would allow to classify downloaded textual material. Third, even if the thesis focuses on languages that all use Devanagari, there were problems with different writing systems mixed in the data (e.g. with a very valuable dataset which, however, was available only in casual Roman trasliteration). Niyati always gained an in-depth insight into the problem, carefully surveyed

existing data resources and related literature, and gradually improved her solution. She is a quick programmer with excellent mathematical thinking, but at the same she always tries to interpret her findings in linguistically meaningful ways, and to support her argumentation using various visualization forms (heatmaps, dendrograms, 2D visualization of embeddings spaces). Last but not least, the thesis is clearly structured, written in flawless and elegant English, and typeset in high quality too.

There are two central contributions of the thesis: a created multilingual data collection (published in the LINDAT reposity), and a set of experiments on cognate induction using, which make use of the created collection. In my opinion, both contributions are quite novel and hopefully also valuable for the Indian NLP community. Obviously, percentage scores resulting from the evaluation stage are well below than what we are used to read in similar studies for some Indo-European languages; however, once again, I would like to emphasize the zero-resource starting point of Niyati's work. That said, being able to run any experiment across such a wide set of mostly heavily under-resourced languages, and to consistently evaluate any metric on them, is already a big success *per se*, regardless of some possibly remaining "child-diseases" (such as the surprisingly under-performing EM solution). In addition, I believe that Niyati's achievements could be useful also from the viewpoint of language evolution studies, as the language situation in India is unique in many aspects.

Finally, I would also like to mention that Niyati has invested a lot of energy into publishing her research results during her master study. Besides a paper submitted to EMNLP (under review now), which has the same topic as her thesis, she is also a co-author of an LREC paper about a multilingual morph-segmented data collection, and the main author of a NAACL paper about cross-lingual transfer of embeddings from Hindi to Marathi and Nepali.

Conclusion

Not only that the goals stated in the thesis specification have been reached, but, in my opinion, some contributions of the presented thesis are truly pioneering for a range of North-Indian languages. I can only wish that Niyati's work will find its followers in the area of Indian NLP soon.

I highly recommend to accept the thesis for the defense.

In Prague, August 19, 2022

Zdeněk Žabokrtský
Institute of Formal and Applied Linguistics
Charles University in Prague