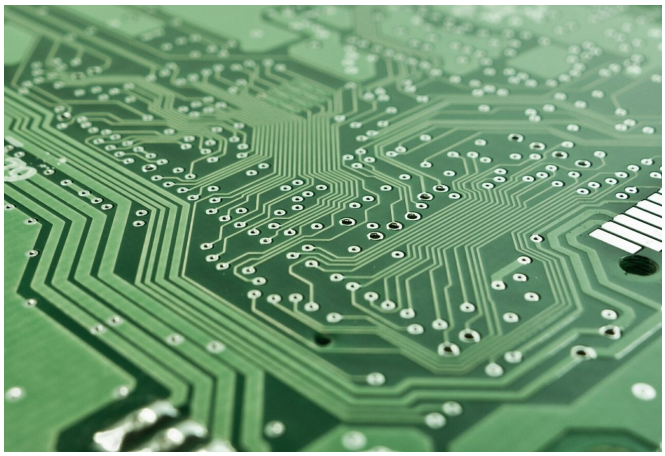
Researchers teach computer to be fluent in Finnish dialects

15 December 2021



Credit: CC0 Public Domain

Computers usually understand Finnish only as the normative standard known as kirjakieli. Finnish dialects, however, create a lot of trouble when interacting with computers, since it is impossible to speak a language without speaking in a dialect of some sort. A research group has built artificial intelligence (AI) models that can automatically detect, normalize and generate Finnish dialects. The results were published in *The 2021 Conference on Empirical Methods in Natural Language Processing*.

Collecting data for making an AI understand dialectal Finnish and Swedish has been on the news recently. The methods devised by the research group of Mika Hämäläinen, Niko Partanen, Khalid Alnajjar and Jack Rueter from the University of Helsinki take this further and enable an AI to be fluent in the Finnish dialects.

Within the paradigm of computational creativity, they have developed a method for converting standard Finnish into one of the 23 Finnish subdialects. Computers should not only be able to understand dialectal Finnish, but they should also be able to express themselves in a dialect.

"With our method, an intelligent system such as a robot can say akku on lopussa (battery is low), for example in Etelä-Karjala dialect akku o lopussa, Etelä-Satakunta dialect akku ol lopus or Länsi-Uusimaa dialect akku o lopus," Hämäläinen says.

For example, the commonly used algorithm of Google Translate fails to translate a dialectal Finnish sentence Oisko sulla jotai esimerkkei siit (Do you happen to have some examples of that) producing a completely incorrect "English" translation Oisko sulla something like that just because Google Translate has been built to work exclusively on standard Finnish. This same phenomenon can be observed with any AI tools that support Finnish like Apple Siri or dictation in macOS.

Dialects are detected from both spoken audio and text

The research shows that detecting dialects is a difficult task when relying on plain text. Dialect identification is easier when the model has access to audio as well because many dialects are marked with distinctive phonetic properties. Thus the latest research published by the researchers deals with detecting dialects from both spoken audio and text.

"The process of normalizing dialects to standard text has many benefits. It allows analyzing dialectal materials using tools for the Standard Finnish, and we can also use the normalized version as a search item when we want to find something from the dialectal materials", says Khalid Alnajjar.

The researchers remind that the problem of understanding dialects is complex and no model can understand natural language like humans do. But the created models open many more interesting directions for research, such as the

degree to which a dialect deviates from the norm and what are the syntactic differences between different language varieties.

"With this we can improve the current state of Finnish natural language processing solutions and build AI models tailored for individuals. For example, we have already reached impressive results in speech recognition of one person's speech, even in endangered languages", Niko Partanen says

The research group has also developed a similar normalization methodology for the dialects of Swedish spoken in Finland (Hämäläinen et al., 2020b) and historical Finnish (Hämäläinen et al., 2021b).

The dialect generator can be tested online and the dialect normalizer and generator code have been released openly on Github. The dialect identification code can be found on Github as well.

More information: Use online: uralicnlp.com/murre

Github links:
github.com/mikahama/murre
github.com/Rootroo-ltd/Finnish … ialectIdentification

Provided by University of Helsinki

APA citation: Researchers teach computer to be fluent in Finnish dialects (2021, December 15) retrieved 19 August 2022 from https://techxplore.com/news/2021-12-fluent-finnish-dialects.html