Málfræðipróf fyrir tölvur

Þótt lesendur Mannamáls séu eðli málsins samkvæmt áhugasamir um íslenskt mál og almenna málfræði, að minnsta kosti upp að einhverju marki, er ekki útilokað að hrollur fari um einhverja þegar minnst er á málfræðipróf, enda vekur greining í orðflokka, setningarliði o.þ.h. ekki kátínu meðal allra grunn- og menntaskólanema. Undirritaður vinnur um þessar mundir að prófi í íslenskri málkunnáttu sem er sérstakt að því leytinu til að það er ekki ætlað fólki, heldur tölfræðilegum spálíkönum sem mötuð hafa verið á ógrynni texta – hinum alkunnu risamállíkönum.

Mállíkönin sem liggja að baki öflugum gervigreindarspjallmennum á borð við ChatGPT og félaga sæta nefnilega reglulega prófunum í „málkunnáttu“, en slík próf eru bæði hugsuð til þess að veita notendum spjallmenna upplýsingar um það hversu brúkleg þau eru til hvers konar hvunndagsverkefna, en einnig geta þau haft beint rannsóknargildi. Þannig hafa vísindamenn í hugfræðum (málvísindum, sálfræði, gervigreind o.fl.) margir hverjir áhuga á að rannsaka mállíkön með tilliti til þess hvernig þau „læra“ tungumál út frá textamagni í þjálfunargögnum og hvað það gæti sagt okkur um mennska máltöku.

En hvernig prófar maður hversu góð mállíkön eru í íslenskri málfræði? Segja má að í hinu stóra samhengi skiptist slík próf í tvo meginflokka, sem hvor um sig hefur sína kosti og galla. Í þessari grein verður stuttlega fjallað um þessa tvo flokka, með dæmum sem öll eru tekin úr íslenskri málhæfnimælistiku fyrir risamállíkön, sem unnin var á vegum Árnastofnunar.

Flokkur eitt: Líkindamælingar

Fyrir áratug síðan, þegar vindar yfirstandandi tauganetsbylgju í máltækni og gervigreind tóku að blása, kynntu Tal Linzen o.fl. til sögunnar aðferð til að mæla það hvort mállíkön hefðu tileinkað sér einhver atriði í málfræði þess tungumáls sem verið var að þjálfa á. Mállíkön þess tíma voru ekki beint hugsuð sem **spunalíkön** (eða **myndandi mállíkön**, e. *generative language model*) í sama skilningi og spjallmenni dagsins í dag, sem eru fær um að mynda svör við spurningum. Hlutverk þessara líkana var fyrst og fremst að geyma tölfræðidreifingu yfir textann í þjálfunargögnunum, sem svo væri hægt að nýta áfram við alls kyns vélræn textavinnsluverkefni, líkt og t.d. að flokka textabúta í fyrirframgefna flokka.

Nálgun Linzen og félaga byggir á dómaprófum, sem tíðkast hafa í málvísindarannsóknum um áratugabil, og felur í sér að mata líkön á **lágmarkspörum**, setningatvennum á borð við (1)a-b, þar sem setningarnar eru alveg eins fyrir utan eitt tiltekið atriði:

(1) *Dæmi um lágmarkspar í íslensku.*
(a) María er **góður** bílstjóri.
(b) María er **góð** bílstjóri.

Hugmyndin er sú að fái setning (1)a, sem lesendur hljóta að vera sammála um að sé eðlilegri frá málfræðilegu sjónarmiði, *hærri líkindi* en (1)b megi túlka það sem svo að líkanið meti setningu (1)a sem málfræðilega „réttari“ – þar sem enginn munur er á setningunum tveimur fyrir utan málfræðilegt kyn lýsingarorðsins. Ef nógu mörg sambærileg dæmi eru borin undir líkanið, og það telur réttu útgáfuna alltaf / nær alltaf líklegri að þessu leyti, mætti fullyrða að líkanið sé búið að *læra*, út frá því að hafa séð ótal dæmi af þessu tagi í þjálfunargögnunum, að í íslensku sambeygist lýsingarorð alltaf nafnorðinu sem það stendur með innan nafnliðar.

Þessari aðferð hefur talsvert verið beitt síðastliðinn áratug og til eru risavaxin prófunarsett á borð við BLiMP fyrir ensku, CLiMP og ZhoBLiMP fyrir kínversku og MultiBLiMP fyrir 101 tungumál, þar á meðal íslensku.

Kostir: Þessi prófunaraðferð nýtir beint útreikninga líkansins og ætti því að gefa eins skýra mynd og hægt er af því hvað það hefur í raun „lært“. Þá telst það kannski að vissu leyti kostur að þessi aðferð virkar betur til að mæla raunverulega getu minni líkana, sem ekki eru fær um að fylgja fyrirmælum notenda (sjá næsta kafla).

Gallar: Fyrir það fyrsta er ekki hægt að beita þessari aðferð nema þegar rannsakendur hafa beinan aðgang að líkaninu – sem er því miður ekki tilfellið með flest risamállíkön dagsins í dag (ChatGPT, Gemini, Claude o.s.frv.) sem eru lokuð og læst innan veggja stórfyrirtækjanna sem búa þau til. Þá er erfitt að túlka líkindareikninginn sem slíkan. Hversu lítil líkindi samsvara því t.d. að setning sé algjörlega ótæk að „mati“ líkansins? Einnig er takmarkað gagn að aðferð þar sem ekki stendur til boða að hafna báðum setningunum sem ótækum eða samþykkja báðar sem góðar og gildar – önnur setningin kemur alltaf betur út.

Flokkur tvö: Próf með skipunum (e. *prompts*)

Eftir tilkomu spjallmennisins ChatGPT og þeirrar kynslóðar risamállíkana sem eru fær um að svara fyrirspurnum notenda hefur það verið reynt í ýmsum rannsóknum að prófa málgetu líkana með því að skipa þeim beint fyrir að leysa einhver málfræðiverkefni. Hægt er að láta líkönin leggja mat á tækar og ótækar setningar með því að spyrja þau beint út í þær en möguleikarnir eru auðvitað miklu fleiri, sbr. eftirfarandi dæmi úr málhæfnimælistiku Árnastofnunar (þar sem fyrirmælin til líkansins eru á ensku):

(a) Here is an Icelandic sentence, followed by a question: *Hún bað mig um að hjálpa sér og ég gerði það. Hverjum hjálpaði ég?* Answer the question with only one word in Icelandic. (Rétt svar: *Henni*.)
(b) Fill in the blank in the following Icelandic sentence with the correct past tense of the verb tagged with : *Okkur langaði að krata fiskinn örlítið, þannig að við _ hann áður en hann fór í ofninn.* Answer only with one word. (Rétt svar: *krötuðum*.)
(c) Which of the slash-separated options in the following question forms part of a sentence that is grammatical in Icelandic? *Einn/Ein/Eitt húðflúranna var af stórum dreka.* Answer only with one word. (Rétt svar: *Eitt.*)

Eins og dæmin sýna er ekkert því til fyrirstöðu að leggja svipuð próf fyrir risamállíkön og fyrir manneskjur. Þannig byggjast nær öll mælaborð sem eiga að sýna getu risamállíkana (t.d. mælaborð Miðeindar á Íslandi eða hið alþjóðlega EuroEval) á prófum þar sem mállíkönin fá beinar skipanir ogeiga að fylgja þeim við að leysa hin ýmsu verkefni. Á þetta jafnt við um prófanir í málhæfni og öðrumþáttum á borð við t.d. rökleiðslu.

Kostir: Segja má að þessi aðferð krefjist ekki mikillar tæknikunnáttu af hálfu þess sem prófar. Hægt er að beita prófum af þessu tagi á lokuð líkön og prófa mun fleira en með fyrri aðferð. Og að einhverju leyti má kalla þetta skýrari samanburð við fólk, t.a.m. er hægt að hafna öllum setningum eða samþykkja allar.

Gallar: Sýnt hefur verið fram á að svör sem fást með því að spyrja líkönin *endurspegla ekki beint líkindadreifingu þeirra*, þ.e. það að líkan segist telja eitthvað atriði réttara en annað þýðir ekki að það atriði sé í raun líklegra samkvæmt útreikningum þess. Þá felur þessi aðferð í sér aukið flækjustig í prófun: Ólíkt orðalag í fyrirspurn notanda getur haft áhrif á svarið og sú krafa að líkanið „skilji“ skipunina og fylgi henni rétt bitnar frekar á minni líkönum en þeim stærri.

Lokaorð

Eins og lesendum er vonandi ljóst eftir lestur þessarar greinar er nokkuð gagn að hvorri aðferðinni fyrir sig en erfitt er að segja að önnur eigi alltaf betur við en hin. Þó er það mat undirritaðs að í fullkomnum heimi hefðu rannsakendur betra aðgengi að helstu risamállíkönum sem verða til úti í heimi og þar með væri hægt að beita fyrri aðferðinni – beinum líkindamælingum – mun oftar. Í ljósi þess að svo er ekki verður sennilega ekki hjá því komist að styðjast við seinni aðferðina í a.m.k. einhverjum tilfellum til þess að reyna að kanna betur þessi ólíkindatól og þeirra aðferðir við að tileinka sér mannlegt mál.

Heimildir

Bjarki Ármannsson, Finnur Ágúst Ingimundarson og Einar Freyr Sigurðsson. 2025. An Icelandic Linguistic Benchmark for Large Language Models. *Proceedings of the 25th Nordic Conference on Computational Linguistics (NoDaLiDa)*:37-47.
Jennifer Hu og Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*:5040–5060.
Tal Linzen, Emmanuel Dupoux og Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics* :521–535.

Mynd

Unseen Studio, Unsplash.