

Hvernig getum við vitað hvaða þýðingarvél er best?

Það getur verið flókið að bera saman tvær þýðingar og meta hvor er betri. Þegar þýtt er á milli tungumála er mikilvægt að rétt merking komist til skila, lesandi eða hlustandi ætti ekki að þurfa að kunna upprunamálið til að átta sig á merkingunni, en það er líka mikilvægt að þýðingin sé læsileg og á málsmáli sem er samsvarandi því sem verið er að þýða úr. Hún geti þannig þjónað sama tilgangi og upprunalegi textinn.

Flest okkar hafa notað þýðingarvélar upp að einhverju marki enda hefur gagnsemi þeirra aukist mjög síðustu ár. Það eru ekki mjög mörg ár síðan þýðingarvélar þýddu texta orð fyrir orð eða því sem næst og niðurstaðan úr svoleiðis löguðu gat verið óttaleg vitleysa. Þegar þýðingarvélar sem byggja á tauganetum komu til sögunnar fyrir tæpum áratug urðu vélarnar fljótlega betri í því að búa til læsilegan texta. Læsilegur texti þýddi þó ekki endilega að þýðingin væri rétt. Dæmið hér að neðan sýnir t.d. úttak úr tveimur tauganetsþýðingarvélum sem þýddu úr ensku yfir á íslensku árið 2021:

Enska: *He went into cardiac arrest, was later declared brain dead and taken off life support.*

Þýðing 1: *Hann fór í hjartastopp, var síðar úrskurðaður heiladauður og tekinn af lífi.*

Þýðing 2: *Hann fór í hjartastopp, var síðar lýst heila dauður og tekið burt líf stuðning.*

Hvor þýðingin er betri? *Þýðing 1* er á lýtalausri íslensku en merking þýðingarinnar vísar nokkuð frá upprunatextanum. *Þýðing 2* er á brogaðra máli en þó er líklegra að lesandinn átti sig á réttri merkingu.

Þegar við notum þýðingarvélar, viljum við að þær gagnist okkur og þýðingarnar þjóni sínu hlutverki vel. Til að velja þá bestu getum við prófað nokkrar, skoðað þýðingar á dæmum sem við látum þær þýða og treyst á innsæið til að velja þá sem okkur líst best á. En ef við höfum engin skýr viðmið mun sú leið líklega gefa okkur heldur takmarkaðar upplýsingar um hversu gott eða lélegt vélþýðingarkerfið sem við veljum er.

Fjölmargar staðlaðar aðferðir til að meta vélþýðingar hafa verið settar fram undanfarna áratugi og vel skilgreindar mæliaðferðir hafa gegnt mikilvægu hlutverki í þróun þýðingarvéla alveg frá því að fyrstu þýðingarvélarnar komu fram á sjötta áratug síðustu aldar. Þrátt fyrir langa sögu er samt enn verið að þróa nýjar aðferðir við mat á vélþýðingum og deila um hvaða aðferðir gagnast best. Stóra vandamálið er auðvitað að hægt er að þýða það sama með mismunandi hætti og allar þýðingarnar orðið góðar og þjónað sínu hlutverki vel. Þá getur verið ómögulegt að finna hlutlæga aðferð sem alltaf finnur bestu þýðinguna.

Við getum skipt aðferðum við mat á vélþýðingum í tvo flokka: sjálfvirk mat og mat með handafli. Þessar tvær leiðir hafa sína kosti og galla.

Sjálfvirk mat á vélþýðingum

Sjálfvirkar aðferðir eru fljótlegar og ódýrar og við getum borið mörg kerfi saman hratt. Þær eru þó allar einhverjum annmörkum háðar og þess vegna eru til margar sjálfvirkar aðferðir og stöðugt verið að skilgreina nýjar. Lengst af hafa verið notaðar aðferðir sem byggja á samanburði við þýðingar unnar af fólki. Þá eru textar á upprunamálinu valdir og þýðandi fenginn til að þýða þá án þess að nota þýðingarvélar sér til aðstoðar. Þær þýðingar eru svo bornar saman við úttak þýðingarvéla. Því líkari sem vélþýðingin er viðmiðunarþýðingunni, því betri er þýðingarvélin talin vera. Til að niðurstöður séu

vel marktækar er algengt að í svona mæliprófum séu a.m.k. í kringum 1000 setningar. Helsta vandamálið við að notast við viðmiðunarþýðingu er þó það að sama textann er iðulega hægt að þýða með margs konar hætti og allar þýðingar jafn réttar. Því getur góð þýðing verið metin sem léleg þýðing vegna þess að hún er of frábrugðin því sem miðað er við. Auk þess þarf stöðugt að búa til ný svona próf, við viljum t.d. vita hvort þýðingarvélar ráði við nýjungar í málinu. Ekki síður mikilvægt er að stundum notast þeir sem búa þýðingarvélar til við gömul matspróf til að fínstilla vélarnar sínar. Þau próf duga þá ekki lengur til að meta gæði vélanna með hlutlægum hætti.

Með tilkomu stórra mállíkana undanfarin ár hafa verið þróuð próf sem notast ekki við viðmiðunarþýðingar. Þá er yfirleitt um annað tveggja að ræða:

- 1) Sérhæfð líkön sem hafa verið þjálfuð sérstaklega til að gefa þýðingum einkunnir. Það er gert með því að sýna þeim ógrynni af þýðingum sem fólk hefur gefið einkunn og líkönin eiga að herma eftir því sem fólk hefur gert.
- 2) Stór mállíkon á borð við GPT-5 eða Claude eru spurð hvort þau finni villur í þýðingum. Ef þau gera það er spurt hvort villurnar séu alvarlegar eða minniháttar. Að lokum er villufjöldinn talinn saman.

Nýju aðferðirnar hafa þann kost að auðveldlega er hægt að fá samanburð á milli mismunandi kerfa án þess að þurfa að leggja í kostnað við að búa til prófunargögn. En þau hafa líka þann ókost að líkönin taka mið af þeim textum sem notaðir voru til að búa þau til, sem eru ekki endilega alltaf góðir, og eldri dómum um gæði texta, sem eru ekki endilega alltaf vandaðir heldur. Líkönin eru því mistæk og geta gefið okkur aðrar niðurstöður en fólk myndi gera.

Fólk metur þýðingarvélar

Þegar þýðingarvélar eru metnar með handafli, þ.e.a.s. þegar fólk fer kerfisbundið yfir úttak úr vélunum og metur gæðin eftir fyrirfram skilgreindri aðferðafræði, getum við fengið ítarlegri upplýsingar um gæði vélanna. Þá er hægt að skoða tiltekna þætti sérstaklega og sjá þannig hverju vélarnar eru góðar í, hverju þær klikka á, og hvaða vélar skila oftast af sér nothæfum þýðingum. Rétt eins og á við um sjálfvirku aðferðirnar eru ýmsar leiðir til að nálgast viðfangsefnið. Undanfarin ár hafa aðferðir sem byggja á því að merkja villur í þýðingum og gefa setningum eða efnisgreinum svo einkunn þótt gefa besta raun og verið næst því að skila niðurstöðum sem almenn sátt getur verið um. En mögulega getur skipt máli hver fer yfir þýðingarnar. Venjulegir notendur eru t.d. líklegir til að meta þýðingarvélar með aðeins öðrum hætti en atvinnuþýðendur. Atvinnuþýðendur eiga auðveldara með að þekkja villur sem eingöngu þýðingarvélar gera og hafa hugsanlega sumir tilhneigingu til að meta slíkar villur sem svo að þær séu alvarlegri en aðrar. Stærsti ókosturinn við mat með handafli er svo auðvitað hversu tímafrekt það er. En þegar nákvæmar niðurstöður skipta máli er enn treyst á mat sérfræðinga í þýðingum.

En hver er þá best?

Það er ljóst að það getur verið flókið að meta hvaða þýðingarvélar eru bestar. Hvernig eiga venjulegir notendur þá að snúa sér í þessu? Ekki hefur enn verið sett upp „mælaborð“ fyrir þýðingarvélar eins og hefur t.d. verið gert til að bera saman getu gervigreindarlíkana á ýmsum sviðum, a.m.k. ekki til að mæla þýðingar á eða úr íslensku. Árlega er þó haldin nokkurs konar keppni í vélþýðingum á stærstu vélþýðingaráðstefnu heims og undanfarin fimm ár hefur m.a. verið keppt í þýðingum úr ensku yfir í íslensku. Niðurstöður keppninnar 2025 voru kynntar þann 8. nóvember síðastliðinn en fjallað verður um þær í pistlinum „Áfram íslenska! - um keppni í vélþýðingum og niðurstöður hennar“ sem birtist hér á Mannamáli þann 19. nóvember!