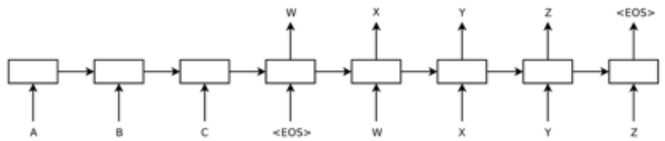
Darba autors: Eduards Mukans
Stud. apl. nr.: em18044
Darba vadītājs:
profesors Dr. Sc. Comp. Guntis Bārzdiņš

Neironu mašīntulkošanas arhitektūras

LATVIJAS UNIVERSITĀTE
ANNO 1919

1. Teorētiskā daļa

1.1. Sequence to Sequence

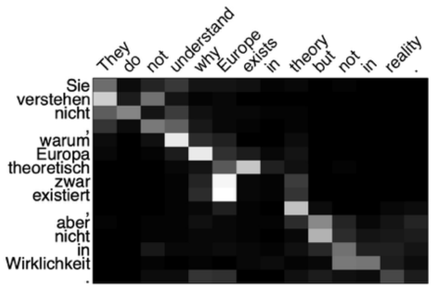


Modelis saņem secību «ABC» un izvada secību «WXYZ». Pielietojot šo modeli nav obligāti kā secības garums sakritīs ar izejas garumu. Modeļa galvenie elementi ir enkodētāji un dekodētāji, kas attēlā ir apzīmēti ar baltiem taisnstūriem.

Enkodēšana ir datu konvertēšanas process no vienas formas uz citu. Dekodēšana ir enkodēšanas pretējais process.

Mašīntulkošanas gadījumā, enkodētājs pārveido teikumu uz matemātisko vektoru un dekodētājs saprot šo vektoru un mēģina atrast atbilsošu teikumu citā valodā. Pēc būtības enkodētājs un dekodētājs ir rekurentie neironu tīkli.

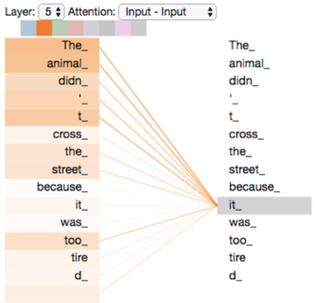
1.2. Attention



"Attention" modelis atšķiras no klasiskā "sequence-to-sequence" modeļa divos veidos. Pirmkārt, enkodētājam jānosūta visi slēpti stāvokļi dekoderam, lai pārtulkotu teikumu. Otrkārt, pirms izveidot pārtulkotu rezultātu "attention" dekodētājs veic papildus darbību. Lai pievērstu uzmanību tikai šīm teikuma daļām, kas attiecas uz šo dekodēšanas laika posmu, dekodētājs rīkojas šādi:

1. Apskata saņemtu enkodēto slēpto stāvokļu kopu. Katra enkodētāja slēptais stāvoklis visvairāk ir saistīts ar konkrētu vārdu tulkojamā teikumā;
2. Piešķir katram slēptajam stāvoklim punktu skaitu;
3. Sareizina katru slēpto stāvokli un to punktu skaitu, pielietojot softmax transformāciju.

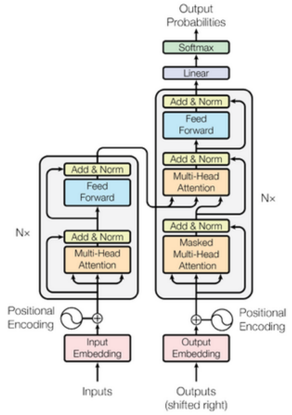
1.3. Self-Attention



"Attention" nozīme to cik liels svars ir iepriekš minētiem vārdiem, bet "self-attention" parāda attiecības starp vārdiem vienā un tajā pašā teikumā.

Ar tumšo krāsu ir parādīti vārdi, kas vairāk attiecas uz vārdu. Vārds «it» attiecas uz vārdiem «the animal». Šī informācija cilvēkiem ir viegli uztverama, bet algoritmam izskaidrot to nav viegli. Kad modelis apstrādā katru vārdu, "self-attention" metode ļauj apskatīt vārdus arī uz citām pozīcijam teikumā, lai šim vārdam iegūtu labāku atbilsošu vektoru.

1.4. Transformer



"Transformer" arhitektūra balstās uz enkodētāju un dekodētāju blokiem. Tie ir identiski enkodētāji un dekodētāji, kas atrodas viens virs otra. Enkodētāju un dekodētāju bloks strādā sekojoši:

1. Vārdu vektors no ieejas teikuma tiek padots uz enkodētāju;
2. Vektori tiek pārveidoti un tiek nodoti uz nākošo enkodētāju;
3. Rezultāts no pēdēja enkodētāja tiek nosūtīts uz visiem dekodētājiem.

2. Praktiskā daļa

Praktiskā darba ietvaros ir izstradāta mašīntulkošanas programma, kas spēj pārtulkot angļu valodas tekstu uz latviešu valodu. Ir izvelēta "transformer" arhitektūra, jo tā ir relatīvi jauna un uz tās pamata ir izveidota jaunāka pieeja - "BERT".

2.1. Tulkojumu piemēri

EN: It matters because we want to feed our people with food.

Uzbuvēts tulkotājs: Tas ir svarīgi, jo mēs vēlamies pabarot mūsu iedzīvotājus ar pārtiku.

Google Translate: Tas ir svarīgi, jo mēs vēlamies barot savus cilvēkus ar pārtiku.

Tildes Tulkotājs: Tas ir svarīgi, jo mēs gribam barot savus cilvēkus ar pārtiku.

EN: Transport is fundamental to our economy and society.

Uzbuvēts tulkotājs: Transports ir mūsu ekonomikas un sabiedrības pamatā.

Google Translate: Transports ir mūsu ekonomikas un sabiedrības pamats.

Tildes Tulkotājs: Transports ir būtisks mūsu ekonomikai un sabiedrībai.

3. Secinājumi

Iegūtos tulkojumus nav jēgas salīdzināt ar eksistējošiem tulkotājiem - "Google Translate" vai "Tildes tulkotājs", jo iegūtais modelis ir vairāk pielāgots politiski un juridiski specializētam tekstam.

Modelis spēj pietiekami precīzi tulkot tekstu no angļu uz latviešu valodu, izvēloties gramatiski pareizus un atbilstošus vārdus.

Darbā ir aprakstītas modernas arhitektūras veidi un to attīstības vēsture.