УДК 004.588

GOOGLE TRANSLATE НЕЙРОТҮЙҮНДҮК МАШИНАЛЫК КОТОРУУ СИСТЕМАСЫНЫН ИШТӨӨ ПРИНЦИПТЕРИН АНАЛИЗДӨӨ

Көчкөнбаева Бүажар Осмоналиевна, к.т.н., ОшМУ, Кыргызстан, 723500, Ош ш., Ленинк.331, e-mail: buajar@mail.ru

Эгембердиева Жылдыз Сраждиновна, окутуучу, КУУ, Кыргызстан, 723503, Ош ш., Н. Исанов к. 79, e-mail: Egemberdieva8787@mail.ru

Аннотация. Бүгүнкү күндө машиналык которуу системалары жүздөп саналат, бирок которулган текст эч качан чыныгы тексттин абсолюттук көчүрмөсү боло албайт, анткени ар бир тилдин өзүнө тиешелүү артыкчылык жана кемчиликтери бар. Ошондуктан которуу жолу менен биз оюубузду толук бере албайбыз, башкача айтканда табигый тилдеги текстти бир тилден экинчи тилге маанисин жоготпой которуу оор маселе.
Бул макалада жасалма интеллект теориясына негизделип иштелип чыккан нейротүйүндүк Google Translate машиналык которуу системасынын мисалында азыркы күндөгү которуу системаларынын абалы жөнүндөгү изилдөөлөр каралды.

Ачкыч сөздөр: машиналык которуу, нейротуйундук машиналык которуу, алгоритм, сөз айкаштары, статистикалык которуу

ANALYSIS OF THE PRINCIPLES OF OPERATION OF THE GOOGLE TRANSLATE NEURAL NETWORK MACHINE TRANSLATION SYSTEM

***Kochkonbaeva Buazhar Osmonalievna**, Ph.D., OshSU, Kyrgyzstan, 723500, Osh ,Lenin street* 331, e-mail: *buajar*@mail.ru
***Egemberdieva Zhulduz Srazhdinovna,** Lecturer, OshKUU, Kyrgyzstan, 723503, Osh, N. Isanov street 79, e-mail: Egemberdieva8787@mail.ru*

Annotation. Today there are hundreds of machine translation systems, but the translated text can never be an absolute copy of the original text, because each language has its own advantages and disadvantages. Therefore, we cannot fully express our thoughts through translation, in other words, it is difficult to translate a text from one language into another without losing its meaning.
This article discusses the current state of translation systems using the example of the neural network machine translation system Google Translate, which is based on the theory of artificial intelligence.

Keywords: machine translation, neural network machine translation, algorithm, phrases, statistical translation

Маселенин коюулушу

Машиналык которуу системаларынын тарыхы “Джорджтаун экспериментине” барып такалат. 1954-жылы IBM компаниясы жана Джорджтаун университети менен биргеликте орус тилинен англис тилине которуучу машиналык которуу системасын жалпыга жарыялаган. Ал учурдагы которуу системасы 250 сөздөн жана грамматикасы 6 эрежеден түзүлгөн. Бул ачылыш машиналык которуу багытындагы изилдөөлөргө багыт берген.

Бирок 1966-жылы Америкалык ALPAC (Automatic Language Processing Advisory Committee) комиссиясы машиналык которуу системаларын иштеп чыгуу пайдасыз экендигин айтып чыккан. Бул докладдан кийин машиналык которуу багытындагы изилдөөлөр бир кыйла токтоп калган. Бирок эсептөө техникасынын тынымсыз өсүүсү менен өткөн кылымдын 70-80 жылдары машиналык которуу багытында статистикалык системалардын пайда болуусун шарттаган.

2006-жылдын 28-апрелинен баштап Google Translate машиналык которуу системасы иштей баштаган. Баарыбыз билгендей Google Translate онлайн машиналык которуу системасына кыргыз тилин кийирүү маселеси 2011 жылы кыргыз өкмөтүнүн колдоосу менен атайын лингвисттердин тобу тарабынан башталып, кошуу үчүн 1 миллион 300 миңден кем эмес сөздүктү чогултуу талап кылынган. Ошентип 2016 жылдан бери улуттук тилибиз дүйнө жүзүндөгү биринчи орунда турган онлайн которуу системасында колдонулуп келет. Бүгүнкү күндө системада 108 тилде онлайн машиналык которуу жүргүзүлөт жана ар күнү 500 миллионго жакын суроо талаптар иштелип чыгат.

Машиналык которуу жасалма интеллект системасы болгондуктан, ал атайын алгоритмдин негизинде иштейт. Алгоритм дүйнө жүзүндөгү бардык тилдер үчүн туура иштей алабы? Бул макаланы жазууда ушул коюулган суроого жооп берүү менен дүйнөлүк машиналык которуу системасынын иштөө принциптерин карап чыгууну максат кылып алдык.

Нейрондук машиналык которуу системасынын архитектурасы

2016-жылы Google компаниясы онлайн которуунун сапатын жакшыртуу максатында жасалма нейрондук түйүндү колдонуучу нейрондук машиналык которуу системасын (Google Neural Machine Translation) сунуштайт. Машиналык которуунун нейрондук модели статистикалык усулга караганда тексттер менен иштөөнүн башка принциптерин колдонот. Google компаниясынан сырткары бул багытта Microsoft жана SYSTRAN компаниялары да иштеп келишет.

Нейрондук түйүндөр пайда болгонго чейин система айрым сөздөрдү, сөз айкаштарын жана фразаларды грамматикалык эрежелердин негизинде которуп келген. Ошондуктан кийирилген текст канчалык татаал болсо которуунун сапаты ошончолук начар болгон.

Нейрондук машиналык которуу системасы (НМКС) сүйлөмдү тексттин маанисине таянып бүтүн бойдон которот. Ошол эле учурда система кийирилген тексттин миңдеген ар кандай варианттарын сактап калбайт.

Которууда сүйлөм сөздүк сегменттерге ажырайт да, атайын декодердин жардамында тексттеги ар бир сегменттин салмагы аныкталат. Андан ары максималдуу мүмкүн болгон ыктымалдуу маани эсептелип чыгат да сегменттин котормосу алынат. Акыркы этапта грамматикалык эрежелердин негизинде сегменттер чогултулат (1-сүрөт).

Архитектураны карай турган болсок ал: кодер (encoder) тармагынан, декодер (decoder) тармагынан жана көңүл буруу (attention) тармагынан турат.

Мында (X, Y) кийирилген жана максаттуу сүйлөмдөр жубу болсун. $X = x_1, x_2, x_3, ..., x_M$ кийирилген сүйлөмдөгү M символдор удаалаштыгы болсун жана $Y = y_1, y_2, y_3, ..., y_N$ максаттуу сүйлөмдүн удаалаш N символу болсун. Анда кодер бул төмөнкү түрдөгү функция болот:

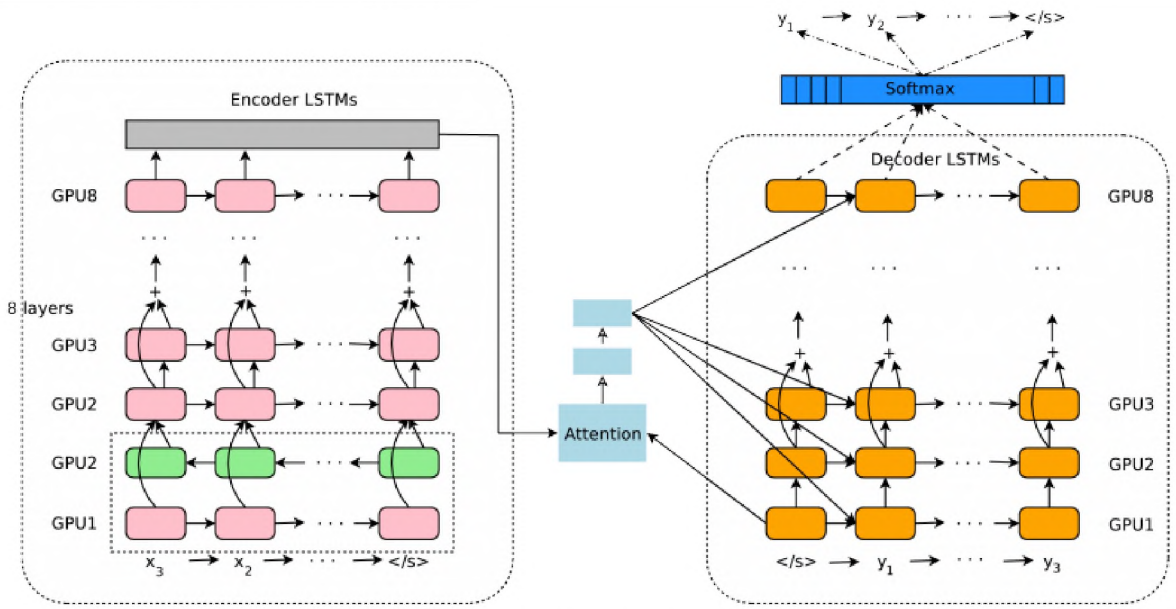
$\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_M} = EncoderRNN(x_1, x_2, x_3, ..., x_M)$

Бул теңдемеде x_1, x_2, ..., x_M белгилүү узундуктагы векторлордун тизмеси. Тизмедеги мүчөлөрдүн саны, кийирилген сүйлөмдөгү символдордун санына барабар (бул мисалда Мге барабар). Чынжырлар эрежесине таянып, P (Y | X) удаалаштыктын шарттуу ыктымалдуулугун төмөнкүдөй ажыратууга болот:

$$P(Y|X) = P(Y|\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, ..., \mathbf{x_M}) = \prod_{i=1}^{N} P(y_i|y_0, y_1, y_2, ..., y_{i-1}; \mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, ..., \mathbf{x_M})$$

мында, y_0 ар бир максаттуу сүйлөмдүн “башталашына” уланууучу атайын символ. Чыгуу убагында алар кийирилген сүйлөмдүн кодун жана учурдагы удаалаштыктын чечмелөөсүн эске алып кийинки символдун ыктымалдуулугун эсептеп чыгышат:

$P(y_i|y_0, y_1, y_2, y_3, ..., y_{i-1}; \mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, ..., \mathbf{x_M})$



1-сүрөт. Google нейротүйүндүк машиналык которуу системасынын архитектурасы[1]

Google нейрондук машиналык которуу системасынын негизинде ыктымалдуулуктарды матрицалык эсептөөчү эки багыттуу рекуренттик нейрондук түйүндөрдүн иштөө принциби жатат. Рекуренттик болгондо программа сөздөрдүн жана сөз айкаштарынын маанисин удаалаштыкта жогору жайгашкан маанилердин негизинде эсептейт. Ошонун негизинде программага бир канча варианттардын ичинен туурасын тандап алууга мүмкүнчүлүк түзүлөт. Эки багытуу болуусу нейротүйүндүн эки: анализдөөчү жана синтездөөчү агымдардын иштөөсүн туюнтат. Биринчи агымда сүйлөм мааниси боюнча бөлүктөргө бөлүнөт жана анализденет, экинчиде жалпы мааниге жараша бир кыйла жакын которуу варианты тандалып алынат. Анализдөөчү түйүн оңдон солго жана солдон оңго карай сүйлөмдөрдү окуйт жана бул тексттин толук маанисин алууга жардам берет.[8]

Нейрондук системада эң кичине бирдик катары сөз эмес, сөз фрагменти каралат да негизги көңүл сөз формасына эмес, сүйлөмдөрдүн маанисине бурулат. НМКС 32000 жакын сөз фрагментин колдонот.

Которуу сапатынын көрсөткүчү

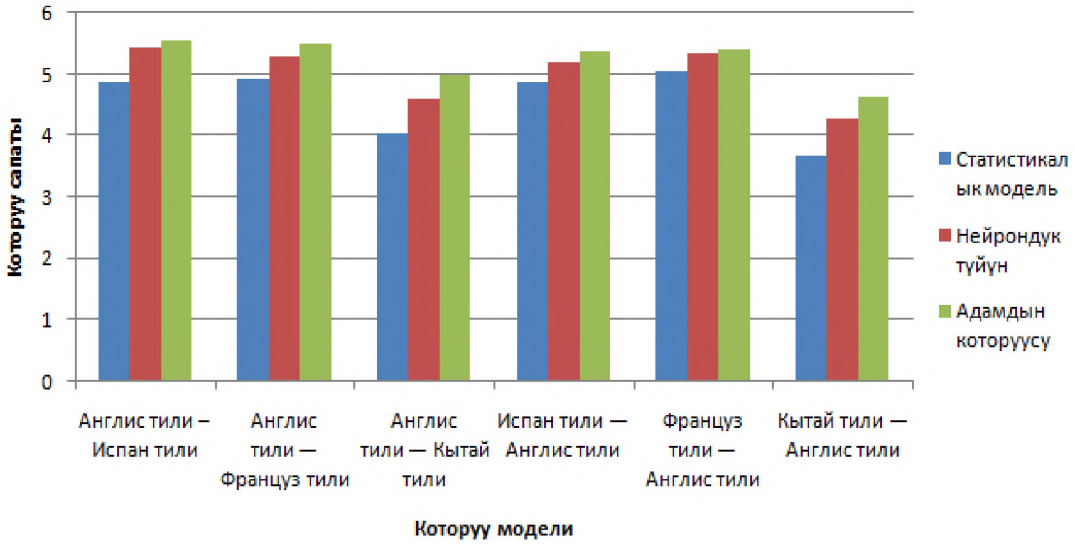
2017-жылы Google компаниясы Google Translate программасын колдонуучулардан сурамжылоо жүргүзгөн. Аларга которуунун үч жолун баалоону суранышкан: машиналык статистикалык, нейрондук жана адамдын которуусу. Жыйынтыгында кээ бир тилдер үчүн машиналык нейротүйүндүк которуу адамдын которуусу менен жакын экендиги аныкталган (1-таблица).

Таблица 1.

Которуу багыты	Статистикалык модель	Нейрондук түйүн	Адамдын которуусу
Англис тили – Испан тили	4,885	5,428	5,550
Англис тили — Француз тили	4,932	5,295	5,496
Англис тили — Кытай тили	4,035	4,594	4,987
Испан тили — Англис тили	4,872	5,187	5,372
Француз тили — Англис тили	5,046	5,343	5,404
Кытай тили — Англис тили	3,694	4,263	4,636

[1] https://arxiv.org/pdf/1609.08144.pdf

Таблицадан көрүнүп тургандай англис тилинен француз тилине, испан тилинен англис тилине которуу жыйынтыктары жогорку көрсөткүчтү көрсөтүп, адамдын которуусуна жакын экендиги графикалык диаграммада көрүнүп турат (сүрөт 2). Бул албетте, тесттик тилдер болгондуктан ушундай болуусу мүмкүн.



Сүрөт 2. Которуу жыйынтыктары

Нейротүйүндүк которуудан айырмаланып статистикалык которуу алгоритмин төмөнкүдөй берүүгө болот:

1-кадам. Файлдан же буфердик эстен берилген тексттин сүйлөмдөрүн алуу.

2-кадам. Кийирилген сүйлөмдү сөздөргө ажыратуу жана сүйлөмдүн чегин аныктоо.

3-кадам. Берилген текстти морфологиялык анализден өткөрүү – сөздүктө табылган ар бир сөз үчүн лексикалык коддорду алуу.

4-кадам. Берилген текстти синтаксистик анализден өткөрүү – сүйлөмдөгү баш жана багыныңкы сөздөрдүн дарак түрүндөгү схемасын түзүү.

5-кадам. Берилген текстке семантикалык анализ жасоо б.а. маанисине көңүл буруу.

6-кадам. Сөздөрдүн дарак түрүндөгү схемасын которуп чыгуу.

7-кадам. Которулган даракты семантикалык, синтаксистик жана морфологиялык жактан шайкеш келтирүү.

8-кадам. Которулган сүйлөмдү сактоо.

Мындай которууда эрежелер жана сөздүк базасы канчалык чоң болсо, котормо туура болот.

Google Translate программасын кыргыз тилинин мисалында карай турган болсок жогорудагы көрсөткүчтөр бир кыйла төмөн болуп калат. Анткени сөздөрдүн түзүлүү эрежеси боюнча кыргыз тили агглютинативдик тилдердин катарына кирет жана сөздөр уңгуга ар кандай мүчөлөрдүн улануусу менен ишке ашат. Ошондой эле интернет түйүнүндө кыргыз тилиндеги тексттердин аз болуусу да ушундай жыйынтыкка алып келет.

Жыйынтыктоо

Жыйынтыктап айтсак, которуу ким же эмне тарабынан болбосун, ал эч качан чыныгы тексттин абсолюттук көчүрмөсү боло албайт, анткени ар бир тилдин өзүнө тиешелүү артыкчылык жана кемчиликтери бар. Которуу адам тарабынан ишке ашса, ал котормочунун билим деңгээлине жараша кемчиликтер менен болуусу мүмкүн.

Ошондуктан которуу жолу менен биз эч качан оюубузду толук бере албайбыз, башкача айтканда табигый тилдеги текстти бир тилден экинчи тилге 100 пайыз так которуу мүмкүн эмес.

Мисалы, англис тилинде жазылган *“The student is reading a book”* деген сүйлөмдү *“Студент китеп окуп жатат”* деп которобуз. Албетте бул көпчүлүк учурда туура котормо, бирок англис тилинде жазылган сүйлөмдө, угуучулар үчүн белгилүү бир адамдын

(студенттин) белгисиз бир китепти окуп жаткандыгы айтылат. Котормодо биз аны коргон жокпуз. Ошондой эле студент деп англис тилинде мектеп окуучусун же билими бар ар кандай адамды атоосу мүмкүн. Ал эми ошол эле сүйлөмдү орус тилине которууда (Студент читает книгу) да студент кыз же бала экени так берилбей жаткандыгын көрүүгө болот. Бул болсо семантикалык анализдин изилдей турган маселеси деп ойлойбуз.

Бирок биз жогоруда карап чыккан Google Translate нейротүйүндүк машиналык которуу программасы жасалма интеллект тармагында дагы бир чоң бурулушту жасады. Программанын артыкчылыгы болуп эки багыттуу рекуренттик нейрондук түйүндөрдүн иштөө принциби эсептелет.

Кыргыз тили бул программага кошулганына аз убакыт болгонуна карабай жакшы көрсөткүчтөрдү көрсөтүүдө. Мындан ары да өз тилибизди өнүктүрүү үчүн семантикалык анализ багытына басым жасалган керектүү программалар иштелип чыгат деген ойдобуз.

Колдонулган адабияттар

1. П.В.Рыбин Теория перевода М.: 2007
2. Микел Л.Ф. Making sense of neural machine translation // Translation Spaces 6:2 (2017) 291–309, DOI 10.1075/ts.6.2.06for.
3. An open source machine learning framework for everyone. [Электронный ресурс]. Режим доступа: https://www.tensorflow.org/
4. He W., Wu Hua., Wang H. Improved neural machine translation with SMT features// Тридцатая конференция AAAI по искусственному интеллекту, 2016.
5. Luong M., Manning C. Achieving open vocabulary neural machine translation with hybrid word-character models, 2016.
6. Sennrich R., Haddow R. Improving neural machine translation models with monolingual data // 54th annual meeting of the association for computational linguistics. Berlin, 2016. P. 86-96.
7. https://arxiv.org/pdf/1609.08144.pdf (Электрондук ресурс).
8. Кудакеева Г.М. Алгоритм распознования зрительных образов / Г.М. Кудакеева, Н.Э. Табылдиева, Е.Ю. Терентьева, Т. Жамалидин уулу // Известия Кыргызского технического университета им. И. Раззакова. №4 (52). 2019. С. 42-48