Mongolian Journal of Engineering and Applied Sciences, 2022, 4–1

Компьютерын ухаан

Гүн сургалтын арга ашигласан Монгол дохионы хэлний хөрвүүлэгч

З.Цолмон*, Г.Саруул

МУИС, ХШУИС, Мэдээлэл, Компьютерын ухааны тэнхим, Машин оюуны лаборатори

Received on 2022.04.04; Revised on 2022.10.12; Accepted on 2022.11.03

*Холбоо барих зохиогч: tsolmonz@num.edu.mn

Хураангуй

Энэхүү ажлаар хүний биеийн төрх байдлыг илэрхийлэх гурван хэмжээст координатын цэгүүдийг ашиглан гүн сургалтын аргаар Монгол дохионы хэлний анхны хөрвүүлэгчийг бүтээхийг зорив. Дэлхий дээр 300 гаруй дохионы хэл байдаг бөгөөд улс бүр өөрсдийн дохионы хэлтэй байдаг. Дохионы хэлээр харилцахдаа үсэг үсгээр үг бүтээж бус харин үг үгээр өгүүлбэр бүтээж харилцдаг. Бидний боловсруулсан Монгол хэлний дохионы хэлний хөрвүүлэгч нь нийтийн хоолны газар ашиглагдах 11 өгүүлбэрийг сонгон тэдгээр өгүүлбэрийг илэрхийлэх үйл хөдлөлийн 869 видеог бэлтгэн машин сургалтад ашиглав. Бичлэг бүрийн кадр бүрээс хүний нүүр, цээж, баруун, зүүн гараас нийтдээ 1662 ширхэг хүний биеийн төрх байдлыг илэрхийлэх цэгүүдийг тооцоолсон гаргаж авсан бөгөөд тэрхүү өгөгдлөө ашиглан Монгол дохионы хэл хөрвүүлэгч машин сургалтын загварыг сургасан болно. Бидний сургасан загвар 96% оновчтой дохионы хэлийг текст болгон хөрвүүлж байна.

Түлхүүр үг: Монгол дохионы хэл, Sequence-To-Sequence, Long Short Term Memory, гүн сургалт

1 Удиртгал

Монгол улсад 2021 оны байдлаар хэл ярианы бэрхшээлтэй 8300, сонсголын бэрхшээлтэй 12671 иргэн буюу сонсгол, хэл ярианы бэрхшээлтэй нийт 20,000 орчим иргэд байдаг. Сонгол, хэл ярианы бэрхшээлтэй иргэд дохионы хэлээр харилцдаг бол энгийн иргэд дохионы хэлийг дийлэнх нь мэддэггүй. Тухайлбал 2022 онд Монгол улсад албан ёсны дохионы хэлний хэлмэрч 20 илүүгүй байдаг гэсэн байна. Сонсгол, хэл ярианы бэрхшээлтэй иргэд дохионы хэлний хэлмэрчээр дамжуулж нийгмийн харилцаанд оролцох шаардлага байнга үүсдэг бөгөөд ингэж нийгмийн харилцаанд оролцоход хүндрэлтэй асуудал олон үүсдэг. Тухайлбал хувийн эмзэг мэдээллээ хэлмэрчээр дамжуулах шаардлагатай болох мөн ганцараа байх үед яаралтай тусламж авах шаардлага гарахад бусдад өөрийгөө ойлгуулж тусламж авч чадахгүй байх зэрэг олон бэрхшээл байдаг байна. Тиймээс сонсгол хэл ярианы бэрхшээлтэй иргэд нийгмийн харилцаанд эрх тэгш оролцох боломж хомс байдаг учир нийгмээс өөрсдийгөө тусгаарлагдах шалтгаанд хүргэдэг. Бид нийгмийн цөөнх бүлэг гэж эдгээр иргэддээ туслах сонирхлоор Монгол дохионы хэлний хөрвүүлэгчийг гүн сургалтын аргыг ашиглан хөгжүүлэхээр зорьсон болно.

Дохионы хэл нь үсэг илэрхийлэх гарын дохионуудаар үг бүтээж бус үг, үйлдэл илэрхийлэх гарын хөдөлгөөний болон нүүрний хувирал хэрэглэж өгүүлбэр үүсгэж харилцдаг [1]. Өөрөөр хэлбэл дохионы хэлээр ярилцахдаа үсэг эвлүүлж үг бүтээж ярилцдаггүй үг эвлүүлж өгүүлбэр бүтээж ярилцдаг байна. Тийм учир ганц зураг буюу бичлэгний ганц фрэймийг онцлох биш үйл хөдлөлийг илэрхийлсэн цуваа олон фрэймүүдийг боловсруулах хэрэгтэй. Жишээ нь байшинд гал гарч байна гэсэн өгүүлбэрийг хэлэхдээ байшин болон гал гэсэн үг илэрхийлэх үйл хөдлөлийг гаргадаг. Бид хүний дохионы хэлний үйл хөдлөлийн дүрс бичлэгийг сургалтын өгөгдөл болгон боловсруулахдаа машин орчуулга болон зурагт гарчиг оноох даалгавар дээр маш сайн ажилладаг гүн сургалтын Recurrent Neural Network (RNN) [2] ашигласан бол энэхүү архитектурын эсээр нь long-short term memory (LSTM) [3] сонгож хийв. Гүн сургалтын сайн загвар үүсгэхэд сургалтын их хэмжээний өгөгдөл шаардлагатай болдог. Өгөгдөл бага байх тусам сургасан загвар нь илүү их алдаатай болж заримдаа бид түүнийг *overfit* боллоо гэж үздэг.

Энэ ажлыг хийхэд тулгарсан хамгийн гол асуудал нь гүн сургалт ашиглан дохионы хэлний хөрвүүлэгчийг сургахад хангалттай өгөгдөл цуглуулах байв. Учир нь Монгол хэлний дохионы хэлний бичлэгний хөмрөг сан Монголд одоогоор хэн ч үүсгээгүй байна. Дохионы хэл нь улс орон бүрд ондоо бөгөөд одоогоор дэлхий даяар 300 гаруй дохионы хэл хэрэглэгдэж байна. Бид энэ ажлаар сонсгол, хэл ярианы бэрхшээлтэй иргэдийг нийгмийн харилцаанд бие даан, тэгш оролцох боломжийг олгохын тулд эхний удаа олон нийтийн үйлчилгээний байгууллага болох

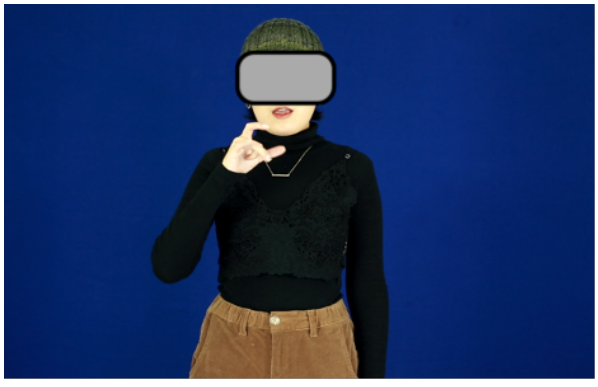
түргэн хоолны газар сонсгол, хэл ярианы бэрхшээлтэй хүн үйлчлүүлэхэд зориулсан энгийн 12-н өгүүлбэр сонгож тэдгээр өгүүлбэрийг илэрхийлэх дохионы хэл хөрвүүлэх системийг туршилтаар хөгжүүлэв.

1.1 Өмнөх ажлууд

Дохионы хэлний цагаан толгойн үсэгнүүдийн гарын дохиог таних маш олон судалгаанууд байдаг бөгөөд хамгийн түгээмэл нь [4] Майкрософтийн кинектийн камераар гарын цэгүүдийг тэмдэглэн Англи хэлний цагаан толгой үсэг таних санамсаргүй ойн (Random forest) арга ашигласан хийсэн хүний гарын дохион танигч нь 92% оновчтой таньдаг байна.

Далд марковын загвар буюу Hidden Markov model (HMM) ашигласан өөр нэг ажил болох [5] нь англи хэлний өгүүлбэрийн түвшинд дохионы хэлнийг хөрвүүлдэг бодит хугацааны системийг хөгжүүлсэн байна. Ингэхдээ хоёр янзын туршилт хийсэн байна. 1) танилтыг илүү оновчтой болгохын тулд цул өнгөөр будсан бээлийтэй болон 2) бээлийгүй хийсэн байна. Бээлий ашигласан загвартай систем нь 99.2% оновчтой таниж байсан бол бээлийгүй системийн оновчтой танилт 84.7% байв.

Зураг 1: Дохионы хэл: Салфетка авъя



Зураг 2: Дохионы хэл: Гэр бүлийн багц байгаа юу?

Гүн сургалтын sequence-to-sequence (seq2seq) [6] архитектур ашиглан Солонгос дохионы хэлний хөрвүүлэгч [7] яаралтай тусламжийн үед ашиглагдах 100 өгүүлбэрийн хүрээнд хийсэн ажил 93.78% оновчлолтой хөрвүүлдэг байна.

Seq2seq машин сургалтын загвар RNN гүн сургалтын архитектурыг ашигладаг. Энэхүү арга нь цуваа оролтыг боловсруулан өөр цуваа оролтод хувирган гаргах үндсэн үйлдэлтэй. Анх Google компани машин орчуулгад ашигласан бөгөөд машин орчуулгын хөгжлийг шинэ шатанд гаргасан байдаг. Seq2seq загвар нь цахим хэл боловсруулалтын маш олон хэрэглээнд одоо ашиглагдаж байна. Тухайлбал, зурагт гарчиг оноох, хугацааны цуваат өгөгдөл дээрээс ирээдүйг таамаглах хийхэд зэрэг загваруудад ашиглагддаг. Seq2seq машин сургалтын загвар нь RNN-ийн үндсэн архитектурыг ашигладаг бөгөөд LSTM болон Gated Recurrent units (GRU) [8] гэсэн хоёр эсийн архитектурыг нейроны эс байдлаар ашигладаг. LSTM нь RNN архитектурын гол дутагдал болох оролтын цуваа хэтэрхий урт бол градиент замхрах асуудлыг тодорхой хэмжээгээр шийдэж өгсөнөөрөө давуу талтай болсон байдаг боловч хэтэрхий урт цувааны хувьд бүрэн шийдэж чадаагүй.

Seq2seq архитектурын үндсэн загвар нь энкодер, декодер гэсэн хоёр хэсгээс бүрддэг ба оролтын цуваа энкодероор боловсруулагдан декодероор зорилтот цуваанд хувиран гаргадаг. Оролтын цувааны урт гаралтын цувааны урт хоёр ялгаатай байж болдог нь энэ архитектурын давуу тал юм.

2 Туршилтын хэсэг

2.1 Дохионы хэлний өгөгдөл

Дэлхийн улс болгон өөрийн дохионы хэлтэй бөгөөд Монгол дохионы хэлний хувьд ийм төрлийн хөрвүүлэгч хийх судалгааны ажил болон сургалтад ашиглах өгөгдөл бүрдүүлэх ажил огт хийгдээгүй байна.

Бид энэ ажлаар нийтийн хоолны газар болох KFC-д түгээмэл хэрэглэгдэх 11 өгүүлбэр болон өгүүлбэр хооронд хийгдэх хоосон үйлдлийг илэрхийлсэн үйл хөдлөл гэсэн нийт 12 өгүүлбэр сонгон Монгол дохионы хэлэнд хөрвүүлж нийт 869 ширхэг өндөр чанартай (HD) бичлэгийг 45 FPS-тэй камераар урдаас харсан байдлаар бэлдсэн.

Сонгосон 12 өгүүлбэрийг дохионы хэлний мэргэжлийн 5 хэлмэрчээр цэнхэр суурь дээр өгүүлбэр тус бүрийг илэрхийлэх үйл хөдлөлийг 10н удаа давтан хийлгүүлж тус бүрд нь бичлэг хийсэн. Нэмэлтээр бичлэгийг сайн дурын оюутнуудаар тухайн хэлмэрчдийн үйл хөдлөлийг харуулж сурган дахин давтуулсан бичлэг хийж өгөгдлийг нэмсэн. Зураг 1-д Салфетка авья гэсэн өгүүлбэрийг илэрхийлж байгаа хэлмэрчийг харуулж байгаа бол зураг 2-т Гэр бүлийн багц байгаа юу? гэсэн өгүүлбэрийг илэрхийлж байгаа хэлмэрчийн зургийг үзүүлэв. Энэ мэт нийт 5 мэргэжлийн хэлмэрч болон оюутнууд ашигласан сургалтын өгөгдлийг бэлдэхэд тусгайлсан өрөө болон ихээхэн цаг хугацаа шаардсан болно.

11 өгүүлбэрийг босоо, бага зэрэг хажуу талаас

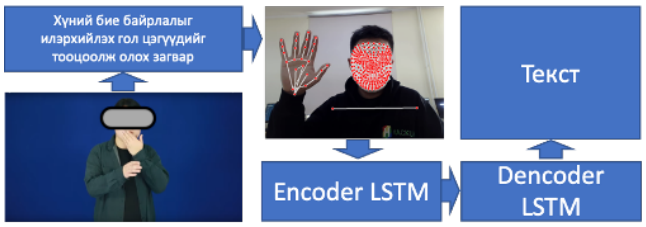
бичлэг хийж бэлдсэн ба нэмээд юу ч илэрхийлээгүй үеийн биеийн үйл хөдлөлийг илэрхийлэх бие төрх байдлыг бичиж тэрхүү төлвийг таньдаг гаралтын хоосон утгатай төлөвийг мөн нэмж өгсөн.

Хүснэгт 1: Монгол дохионы хэлний сургалтад ашиглагдсан 12 өгүүлбэр

Дугаар	Монгол өгүүлбэр
1	Сайн байна уу?
2	Гэр бүлийн багц байгаа юу?
3	Ариун цэврийн өрөө хаана байгаа вэ?
4	Уучлаарай
5	Сальфетка авъя
6	Баяртай
7	Баярлалаа
8	Найзуудын багц байгаа юу?
9	Ямар ямар багц байгаа вэ?
10	Тооцоогоо хийе
11	Төмс авъя
12	(Юу ч хийхгүй зогсох)

2.2 Дохионы хэлний загвар

Монгол дохионы хэлний хөрвүүлэгч загварыг зураг 3-т үзүүлсэн ерөнхий архитектураар угсрав. Бид хүний биеийн байрлалыг илэрхийлэх гол цэгүүдийг тооцоолж олох загвар



Зураг 3: Дохионы хэлийг хөрвүүлэх загварын ерөнхий архитектур

дийг тооцоолохдоо хүний биеийн байрлал тооцоологч өмнө сургасан гүн сургалтын загвар openPose [9], mediaPipe [10] ашиглаж нүүр, цээжин бие, баруун болон зүүн гарын нийт 1662 цэгийг бичлэгний фрэйм болгон дээр хадгалж авсан [6]. нүүр хэсгийн 1404, цээж хэсгийн 132, баруун болон зүүн гар тус бүр 63 цэгээс бүрдэнэ.

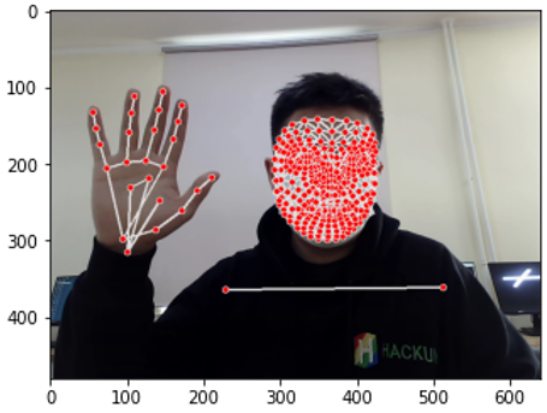
Хүснэгт 2: Зургалтын өгөгдлийн мэдээлэл

Үзүүлэлт	Нийт	Сургалт	Тест
Нийт бичлэг	869	696	173
Хугацаа (минут)	144.83	116.03	28.8
Фрэймийн тоо	112,050	89,640	22,410
Хэлмэрчийн тоо	Хэлмэрч 5 + Оюутнууд		
Бичлэгийн өнцөг	2		

Convolutional Neural Network (CNN) хэрэглээд гарын дохиог ялган хийж болох байсан боловч CNN архитектураар гаргасан загварын оновчлол нь орчноос ихээр хамаардаг тул /гадаа, дотор, гэрэл гэх мэт./ [11] бид ашиглаагүй болно. Монгол хэлний дохионы хэлний цагаан толгойн бүх үсэг зөвхөн гарын хуруууны байрлалаар илэрхийлэгддэггүй тухайлбал Ө үсэг нь гарын хуруууны байрлал нь О үсэгтэйгээ адил боловч сэгсрэх хөдөлгөөнөөр үзүүлдэгээрээ ялгагддаг. Тиймээс оролтын өгөгдөл нь зөвхөн ганц зураг бус үйл хөдлөл илэрхийлсэн дараалал бүхий цуварсан зурагнуудын оролт болох бөгөөд энэ төрлийн бодлогод тохирсон хамгийн сүүлийн үеийн дэвшилтэт архитектур болох seq2seq загварыг ашиглахаар шийдсэн. Бид seq2seq загварт LSTM эсийг сонгосон. GRU-гийн хувьд hyperparameter-ийн тоо LSTM-ээс бага учир зардал багатай сургаж болох боловч бид LSTM оролтын цуваанг хангалттай урт таар оруулж болдог тул сонгосон болно. LSTM архитектур хэдийн оролтын цуваа хангалттай урт авдаг боловч оролтын цувааны урт их болох тусам цувааны бүх элементийн мэдээллээс гаралтын цуваагаа тооцоолох тогтмол урттай вектор бодоход хүндрэлтэй болдог. Бидэнд тийм урт оролтын цуваа одоогоор байхгүй тул илүү сайжруулсан төвөгтэй архитектур ашиглах шаардлагагүй гэж үзсэн болно. Тэрхүү илүү сайжруулсан төвөгтэй архитектурын шийдэл нь Attention эсвэл Transformer [12] архитектурууд боловч энэ удаагийн туршилтад хэрэглээгүй.

2.3 MediaPipe ашиглан хүний биеийн цэгийг таних

Бидний сургах загвар ямар ч орчинд хүний биеийг төрх байдлыг илрүүлж таньж чаддаг байх ёстой. Тиймд бид бэлэн сургасан сайн шалгагдсан загвар болох mediaPipe-ийг сонгон ашиглаж, хүний биеийн төрх байрлалын гол цэгүүдийн координатыг тооцоолуулж түүний тоон утгыг ашиглав.



Зураг 4: Хүний биеийн байрлал тодорхойлох гол цэгийн байрлал

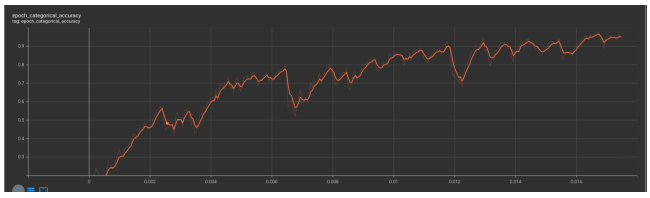
Дохионы хэлийн өгөгдлийн 1 бичлэг ойролцоогоор 10 орчим секунд урттай бичигдсэн бол 1 секундийн бичлэгийг 45 фрэймд хувааж, фрэйм бүр

дээр гол цэгүүдийг тооцуулан, дохионы хэлний хөрвүүлэх загварын оролтод өгөх өгөгдөл болгон авав.

Бид туршилтаа гүн сургалтын нээлттэй эхийн tensorflow санг ашиглан пайтон программчлалын хэл дээр хөгжүүлэн загварыг хэрэгжүүлэв. Загвар сургахад хэд хэдэн параметрийн тохируулга хийсэн бөгөөд бидний туршилтаар хамгийн тохиромжтой утгуудыг параметр тус бүрээр олоход learning rate нь 0.001, epoch-н тоо 300, batch хэмжээ 128 байв. Бидний загварын LSTM-д оролтын өгөгдлийн хэмжээ нь 45x1662 тензор юм. Фрэйм бүрийн хувьд 1662 ширхэг хүний биеийн цэгийг тооцоолуулж нийт нэг секундын бичлэгний 45 фрэйм тус бүрээр тооцсон болно.

3 Үр дүн

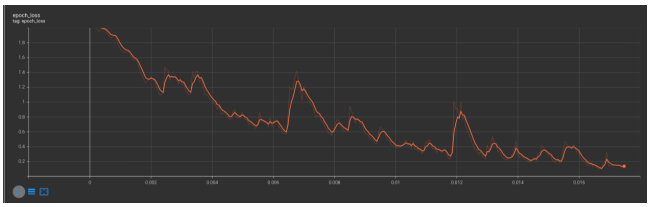
Машин сургалтын ангилалын хийдэг загварыг үнэлэхэд оновчлолыг ашигладаг. Оновчлол буюу accuracy-г тооцоолохдоо шалгах өгөгдлийг оруулж ямар утга илэрхийлж байгаа таамаглалыг зөв утга буюу өгүүлбэртэй харьцуулан буюу нийт шалгах өгөгдлийн тоотой оновчтой таамагласан тооны харьцааг хувиар илэрхийлсэн болно. Зураг 5-д боловсруулсан загвар сургалтын өгөгдөл дээр сургалцаж байх явцад оновчлол хэрхэн өөрлөгдөж байгааг харуулав. Сургалт хийсний дараа бидний загвар Монгол дохионы хэлний бидний сонгосон 12 өгүүлбэрийг ойролцоогоор 96%-ийн оновчтой таньж байна.



Зураг 5: Туршилтын хөрвүүлэх загварын accuracy график

Зураг 6-д загварын сургалт хийж байх үеийн алдааны өөрчлөлтийн графикийг харуулав. Бид загварын алдааны функцээр олон ангилалын cross-entropy сонгосон бөгөөд энэ функц нь алдааг тооцохдоо таамаглаж байгаа ангилал нь жинхэнэ ангилал байх магадлалын зөрүүг тооцож нийт шалгасан өгөгдлүүдийн алдааны дундажыг тооцож байгаа болно.

Загварыг сургах явцын оновчлол болон алдааны график дээрээс харахад огцом өөрчлөлтийн цэгүүд байгаа нь бидний бэлдсэн сургалтын өгөгдөл тийм ч хангалттай биш байгааг илтгэж буй хэдий ч загвар нь суралцах чадвартай болсон гэдгийг хангалттай харж болно.



Зураг 6: Туршилтын хөрвүүлэх загварын loss график

4 Дүгнэлт

Судалгааны үр дүнгээс харвал гүн сургалтын LSTM архитектур хэрэглэж бага өгөгдлөөр overfit, underfit болгохгүй сургах боломжтойгоос гадна загварын хөрвүүлэх нарийвчлал орчны гэрэл, сүүдрээс хамаардаг дутагдлыг mediaPipe сангийн тусламжтай өгөгдлөө бэлдэж ашигласанаар танилтын хувь сайжруулсан илүү үр дүнтэй болж чадлаа. Бидний дохионы хэлний хөрвүүлэгчийн гаралт нь текст бөгөөд гаралтын текстээс бусад хэлэнд орчуулах, дуу хөрвүүлэгч ашиглах бүрэн боломжтой юм. Тухайлбал орчуулга хийснээр дохионы олон хэлний хооронд хөрвүүлэх боломжтой болж байна.

Загварын оновчтой таамаглах нарийвчлалыг илүү тогтвортой болгоход сургалтын өгөгдлийг илүү их бэлдэх шаардлагатай байгаа бөгөөд өгөгдөл бэлдэхэд хүн хүч, цаг хугацаа их шаардсан ажил болох нь харагдаж байна. Монгол улсад энэ төрлийн суурь өгөгдөл үүсгэх нь ирээдүйд олон судлаачдад хэрэгцээтэй чухал ажил болох нь гарцаагүй билээ.

Манай дохионы хэл хөрвүүлэх системийг ашиглан эхний ээлжинд нийтийн түргэн хоолны үйлчилгээний газрууд сонсгол, хэл ярианы бэрхшээлтэй иргэд үйлчлэх боломжтой болж байгаа юм. Иймд цаашид бид Монгол дохионы хэлний өгөгдлийн хөмрөг сан бүтээх бие даасан судалгааны ажил болгохоор төлөвлөж байна. Мөн дохионы хэлний хөрвүүлэгч загвараа улам сайжруулж илүү урт өгүүлбэрүүд дээр ажиллах чадвартай Attention болон transformer архитектураар сургахаар төлөвлөж байгаа болно.

Талархал

Сургалтын өгөгдөл бэлтгэж өгсөн D-COMP 2022 тэмцээнд оролцсон ELPIS багийн нийт гишүүддээ гүн талархал илэрхийлье.

Зохиогчийн оролцоо

Г.Саруул нь уг судалгааны ажлын санаа гаргасан бол ерөнхий загварын архитектур, гүн сургалтын загвар боловсруулалт, туршилт болон дүн шинжилгээг бүх зохиогч нар ижил оролцоотой гүйцэтгэсэн болно. Г.Саруул нь өгүүллийн эхний хувилбарыг бичсэн бол З.Цолмон өгүүллийн утга, агуулгын

алдааг засаж сүүлийн хувилбарыг бичсэн.

Ашиг сонирхлын зөрчилгүйн баталгаа

Бүх зохиогчид ашиг сонирхлын зөрчилгүй болохыг баталж байна.

Ашигласан ном

[1] Cooper H, Holt B, Bowden R. Sign language recognition. In: Visual analysis of humans. Springer; 2011. p. 539–562.

[2] Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:14061078. 2014.

[3] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997;9(8):1735–1780.

[4] Dong C, Leu MC, Yin Z. American sign language alphabet recognition using microsoft kinect. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops; 2015. p. 44–52.

[5] Gattupalli S, Ghaderi A, Athitsos V. Evaluation of deep learning based pose estimation for sign language recognition. In: Proceedings of the 9th ACM international conference on PErvasive technologies related to assistive environments; 2016. p. 1–7.

[6] Von Agris U, Knorr M, Kraiss KF. The significance of facial features for automatic sign language recognition. In: 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition. IEEE; 2008. p. 1–6.

[7] Ko SK, Kim CJ, Jung H, Cho C. Neural sign language translation based on human keypoint estimation. Applied Sciences. 2019;9(13):2683.

[8] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. Advances in neural information processing systems. 2014;27.

[9] Cao Z, Simon T, Wei SE, Sheikh Y. Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 7291–7299.

[10] Zhang F, Bazarevsky V, Vakunov A, Tkachenka A, Sung G, Chang CL, et al. Mediapipe hands: On-device real-time hand tracking. arXiv preprint arXiv:200610214. 2020.

[11] Oberweger M, Wohlhart P, Lepetit V. Hands deep in deep learning for hand pose estimation. arXiv preprint arXiv:150206807. 2015.

[12] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.