Машин Сургалтын Аргад Суурилсан Монгол Хэлний Үгийн Алдаа Шалгах Программ

Баттөмөрийн Хажидмаа[1*], Дуламрагчаагийн Уганбаатар[1], Энхбатын Соджамц[1], Алтанхуягийн Лхагвасүрэн[1], Төмөрбаатарын Пүрэвсүрэн[2]

[1] *Шинжлэх Ухааны Академи, Математик, тоон технологийн хүрээлэн, Улаанбаатар 13330, Монгол улс*
[2] *Шинжлэх Ухааны Академи, Хэл зохиолын хүрээлэн, Улаанбаатар 13330, Монгол улс*

**Холбоо барих зохиогч: khajidmaa_b@mas.ac.mn; ORCID:0000-0003-4786-0836*

Өгүүллийн мэдээлэл: Хүлээн авсан: 2022.09.11; Зөвшөөрөгдсөн: 2022.10.01; Нийтлэгдсэн: 2022.12.26

Хураангуй: Мэдээллийн технологи хурдацтай хөгжин өөрчлөгдсөөр хүн төрөлхтний харилцааны гол хэрэглүүр болсон хэл бичиг биет байдлаас тоон хэлбэр рүү эрчимтэй шилжиж байна. Тиймдээ ч гадаадад эх хэл шинжлэлийн судалгаанууд тогтмол хийгдсээр эцсийн хэрэглэгчид шууд хэрэглэх боломжтой бөгөөд баталгаатай программуудыг зах зээлд нэвтрүүлээд байна. Анх үгийн алдаа шалгах программууд нь уламжлалт арга буюу n-gram арга дээр суурилдаг байсан бол орчин үед машин сургалтын аргуудыг түлхүү ашиглах болсон байна. Тиймээс энэхүү судалгааны ажлаараа монгол хэлний бүтцэд тохирсон машин сургалт дээр суурилсан үгийн алдаа засах загваруудыг сургахыг зорилоо. Судалгааны хүрээнд машин сургалтын BERT болон SymSpell загваруудыг сургаж, туршсан болно.

Түлхүүр үгс: BERT загвар, SymSpell загвар, Эх хэл боловсруулалт, алдаа шалгуур

1. Оршил

Эх хэл боловсруулалт (ЭХБ) [NLP-Natural language processing] нь хэл шинжлэл, компьютерын шинжлэх ухаан салбаруудыг хослуулан хүн төрөлхтний ярианы хэл соёлыг тоон хэлбэрт нормчилж, улмаар хүн компьютер хоорондын харилцааг дэмжих зорилготой судалгааны чиглэл юм.

Хэл шинжлэлийн судалгаа урьд өмнө нь хийгдэж байсан ч ЭХБ-ын судалгааны чиглэл 1950 онд Англи улсын математикч Алан Тюринг өөрийн түүхэн "Тооцоолох Машин ба Оюун ухаан" гэх бүтээлээ хэвлүүлснээс эхлэлтэй гэж үздэг. ЭХБ нь дотроо маш олон хэсгүүдэд хуваагдах бөгөөд түүний нэг чухал хэсэг нь зөв бичих дүрмийн алдаа шалгах ажил юм.

Дэлхийн ихэнх улс орнууд өөрсдийн хэлний зөв бичгийн алдаа засах программуудыг зогсолтгүй хөгжүүлсээр ирсэн ба тэдгээрээс хамгийн сайн хөгжүүлэгдээд байгаа нь харьцангуй өндөр хөгжилтэй орнуудын болон олон хүн амын хэрэглэдэг Англи, Хятад хэлний алдаа засагч программ хангамжууд байна.

Монгол хэлний хувьд үгийн алдаа шалгах программууд тодорхой хэмжээнд судлагдаж, хөгжүүлэгдэж ирсэн ч саяхнаас олон нийт, албан байгууллагуудын хэрэглээнд тархаж эхэлж байна.

Гэхдээ одоо ашиглагдаж байгаа монгол хэл бичгийн алдаа шалгах программуудад алдаа засах, сайжруулах хэрэгцээ шаардлагууд харагддаг ба тийм боломж бололцоонууд

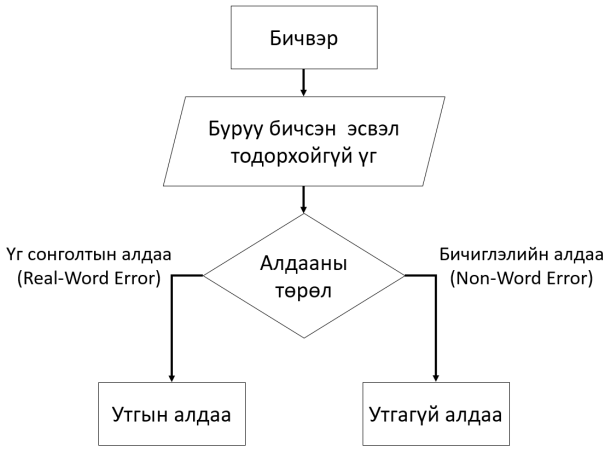
ажиглагддаг тул монгол хэл бичгийн алдаа засах программ, алгоритмуудыг цааш улам судалж, хөгжүүлэх шаардлагатай.

Тиймээс монгол хэлний үгийн алдаа шалгах ажлыг сайжруулахаар машин сургалтын аргууд болох BERT болон SymSpell загварууд дээр тулгуурлан судалгаагаа хийсэн болно.

2. Судалгаанд ашигласан өгөгдөл ба материал

2.1. Үгийн алдаа шалгах

Үгийн алдааны төрлийг үг сонголтын алдаа буюу утгын алдаа, Бичиглэлийн алдаа буюу утгагүй алдаа гэж 2 ангилдаг.



Зураг 1: Үгийн алдааны төрөл.

1. Доод түвшний ажил (Алдаа шалгагч):
   - Буруу бичилттэй үгийг олох
   - Зөв бичилттэй үгсийг санал болгож, эрэмбэлэх
   - Алдааг засах
2. Дээд түвшний ажил (Үг сонголтын алдааг засах):
   - Үгийн бичилт зөв боловч утгын алдаатай үгийг олох
   - Зөв утгатай үгсийг санал болгож, эрэмбэлэх
   - Алдааг засах

Үг биш алдааны төрлүүд:

1. Салсан бичилт (Split word): Үгийн дотор алдаатайгаар хоосон зай оруулсан бичилт
2. Нийлсэн бичилт (Run-on or Merged words): Хоёр эсвэл түүнээс дээш үгийг нийлүүлж бичсэн бичилт
3. Үсэг дутах, илүүдэх, эсвэл андуурах (insertion, deletion an substitution-IDS): Нэг ба түүнээс дээш үсэг дутуу, илүү, эсвэл андуурч оруулсан бичилт

Алдааг илрүүлэхэд өргөн ашиглагддаг арга бол үгийн толь бичгээс хайх буюу Dictionary lookup гэх арга байдаг. Хэрэв оролтын үгтэй адилхан үг жагсаалтад байхгүй бол оролтын үгийг алдаатай гэж үзнэ алдаа засах алхам руу шилждэг.

Автоматаар алдааг илрүүлэн засах 4 үндсэн алхам байдаг. Үүнд:

> Алхам 1: Алдаатай бичсэн үгийг тодорхойлох
> Үгийн алдааны төрлөөс тухайн үг ямар алдаатай гэдгийг тодорхойлно. Толь бичгээс тухайн үг олдохгүй бол алдаатай гэж үзэж болно.

Алхам 2: Алдаатай үгийг засварлах үйлдлийн тоог тодорхойлох
Засварлах зай гэдэг нь хоёр үг хоорондын ижил төстэй байдлын хэмжүүр бөгөөд бага байх тусам үгнүүд нь ижил төстэй байдаг. Тодруулбал энэ нь a үгийг b үг болгон хөрвүүлэхэд шаардагдах хамгийн бага тооны үйлдлийн тоогоор тодорхойлогддог. Энэ нь a мөрөнд тэмдэгт оруулах, устгах эсвэл солих замаар хийгддэг.

1. Нэмэх (үсэг нэмэх) : “өвл” → “өвөл”
2. Устгах (үсэг устгах): “сургалта” → “сургалт”
3. Сэлгэх (энэ нь зэргэлдээ хоёр үсгийн байруудыг сэлгэх): “ухана” → “ухаан”
4. Солих (нэг үсгийг өөр үсгээр солих): “өгөгдол” → “өгөгдөл”

Алхам 3: Үг хайх
Өгөгдлийн сангаасаа зөв бичигдсэн үгсийг хайж олно.

Алхам 4: Үгсийн магадлалыг тооцоолох
Аливаа үг сан дотор хэдэн ширхэг, сан дотор нийт хэдэн ширхэг үг байгааг ашиглан тохиромжтой үгийг магадлалын аргаар олох боломжтой.

2.2. Өгөгдлийн сан

1. Үгийн давтагдаагүй үгсийн сан
Ц. Дамдинсүрэн, Б. Осор нарын 1983 онд хэвлүүлэн гаргасан “Монгол үсгийн дүрмийн толь” бүтээлийн үгийн сангийн эх хувилбарыг хадгалан түүний онцлог сэтгэлгээ, дүрэмд тулгуурлан одоо цагт хэрэглэгдэж буй үгийн сангаар баяжуулан хөгжүүлсэн “Зөв бичгийн дүрмийн толь” дээр үндэслэсэн нийт 2 сая гаран үгтэй толь бичгийн 1048075 (нэг сая дөчин найман мянга далан таван) ширхэг давтагдаагүй үгтэй өгөгдлийн сангаар туршилтаа хийсэн.

2. Сөрөг үгийн сан
Сөрөг үгсийн жагсаалтаа гаргахдаа эерэг үг дээр эсрэг үг бүтээх угтвар болон нөхцөл залгасан. Жишээ нь: эс, бус, гүй, биш гэх мэт.

	муу
	хөнөөлтэй
	түгшүүртэй
	ууртай
	зовоох
	түгшсэн
дөлгөөн бус	хайхрамжгүй
туушtай бус	аймшигтай
тайван бус	хэрцгий
эрэлхэг бус	муухай
бүтээлч бус	бүтэхгүй
сэргэлэн бус	өөдгүй
бишрэм бус	зэвүүн
баярсал бус	заналхийлсэн
баясгалантай бус	яраглах
шалгарсан бус	дарлал
шударга бус	залхмаар
хөгжилдөх бус	сүйрсэн

Зураг 2: Сөрөг үгийн өгөгдлийн сан.

3. Арга зүй

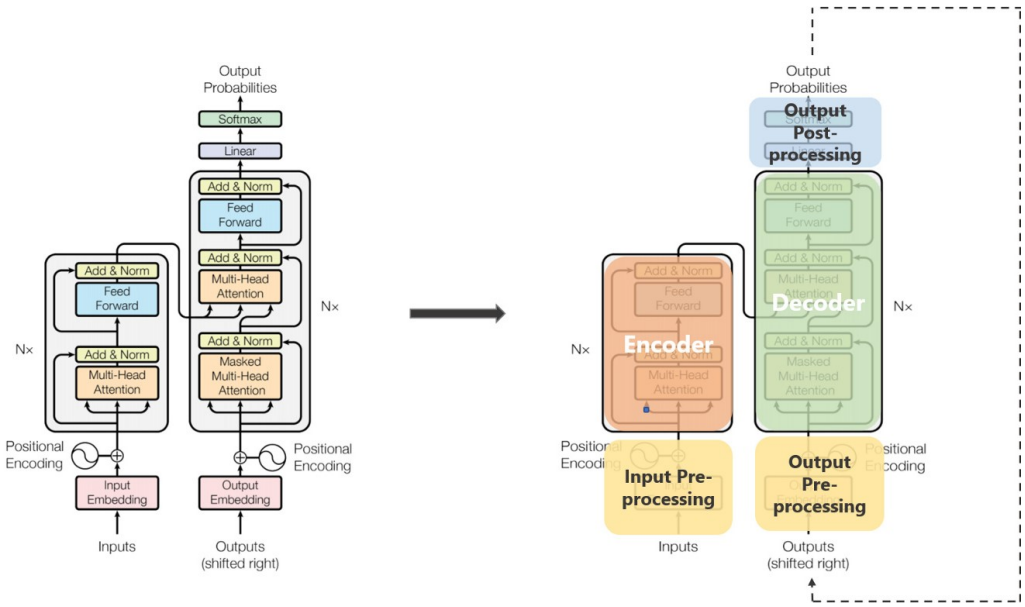
3.1. BERT загвар

BERT буюу Bidirectional Encoder Representations from Transformers загвар нь Google-ээс гаргасан трансформер төрлийн архитектур дээр суурилсан эх хэл боловсруулалтад (NLP) ашиглах боломжтой хоёр чиглэлт кодлогч бүхий машин сургалтын загвар юм. Уг загварыг Жейкоб Девлин болон түүний Google-дэх Хиймэл Оюуны баг хамт олон бүтээж, 2018 онд олон нийтэд танилцуулан, нийтлүүлсэн байдаг [1], [2]. Анх танилцуулагдаж байхдаа BERTbase болон BERTlarge гэсэн 2 хувилбартай гарч байсан төрлүүд нь BooksCorpus-аас 800 сая үгтэй, англи Wikipedia-аас 2500 сая үгийн сан бүхий өгөгдөл сургагдсан загварууд байсан юм [4].

3.2. Трансформер загварын архитектур

Трансформер нь кодлогч, код тайлагч гэсэн хоёр үндсэн хэсгүүдээс бүрдэнэ. Кодлогч хэсэг нь оролтыг кодлох давхаргуудаас бүрдэх бол код тайлагч нь кодлогчийн үүсгэсэн гаралтыг тайлах код тайлах давхаргуудаас бүрдэнэ. Ингэхдээ кодлогч нь оролтын бичвэрийг үг бүрээр нь харгалзах байрлалтай нь кодолж тусгай вектор бүхий токен гэх зүйл үүсгэн код тайлагч руу илгээнэ.

Код тайлагчийн хувьд хүсэмжит гаралтыг мөн адил үг бүрээр гэхдээ байрлалын хувьд баруун тал руу нэг алхам шилжсэн байдлаар кодлон, токен үүсгэх ба үүссэн хүсэмжит гаралтын токен болон кодлогчоос ирсэн токенуудын нэгдлээс хүсэмжит гаралтын утгын тооцоолох аргад суралцана.



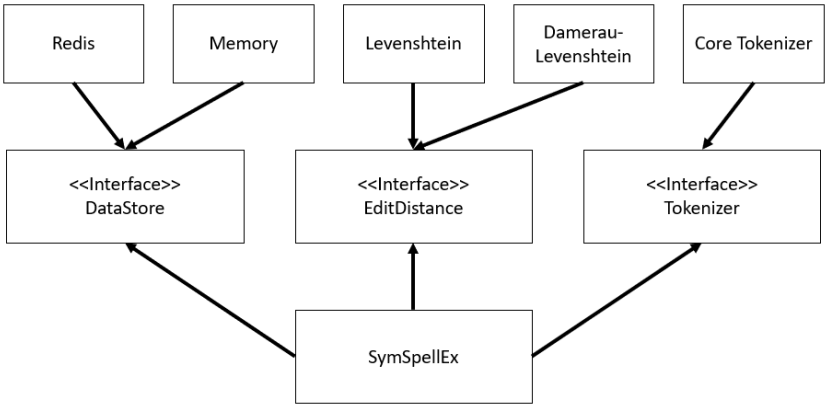
Зураг 3: Трансформер загварын архитектур [4].

3.3. SymSpell (Symmetric delete spelling correction algorithm)

SymSpell алгоритм нь олон тооны мөрүүдийн жагсаалтаас засварлах тодорхой зайд маш богино хугацаанд бүх үгийг олох алгоритм юм. SymSpell буюу Symmetric Delete зөв бичгийн алдаа засах алгоритм нь Дамерау-Левенштейн зайнаас хамаарсан засварлаж болох боломжууд болон толь бичгийн хайлтын төвөгтэй байдлыг багасгадаг. Энэ арга нь өгөгдлийн хэлнээс үл хамааран (нэмэх + устгах + сэлгэх + солих бүхий) уламжлалт аргаас зургаан дахин хурдан юм.

Бусад алгоритмуудаас сэлгэх+солих+нэмэх алхам хийлгүй зөвхөн устгах алхмууд хийдгээрээ ялгаатай. Оролтын сэлгэх+солих+нэмэх алхмууд нь толь бичгийн устгах алхам болж хувирна. Орлуулах болон оруулах нь маш их цаг хугацаа ордог. Хурд нь зөвхөн устгах боломжтой үгсийг бий болгож урьдчилсан тооцоолсноор ажилттай хэрэгждэг. Дунджаар 5 үсэгтэй үгэнд хамгийн их засварлах зай нь 3 үед 3 сая орчим зөв бичгийн алдаа гарч болзошгүй боловч SymSpell нь урьдчилан тооцоолж хайх үед бүгдийг нь хамрахын тулд ердөө 25 устгал үүсгэх шаардлагатай.

3.4. SymSpell архитектур



Зураг 4: SymSpell-ийн архитектур [6].

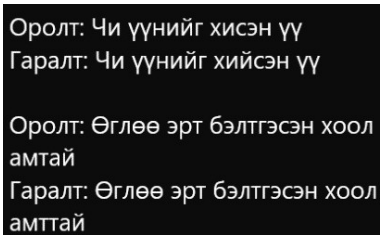
Засварлах зай:

Энэ нь хоёр үгийн хоорондох ялгааг хэмжих мөрийн хэмжүүр юм. Засварлах зайг тооцоолох нь нэг мөрийг нөгөө мөр болгон хувиргахад шаардагдах хамгийн бага үйлдлүүдийг тооцоолох замаар хоёр мөр (жишээ нь, үг) бие биетэйгээ хэр төстэй болохыг тодорхойлох арга юм.

4. Үр дүн

4.1. Bert загвар дээр хийсэн туршилт

Ц. Дамдинсүрэн, Б. Осор нарын “Зөв бичгийн дүрмийн толь” дээр үндэслэсэн нийт 2 сая гаран үгтэй сангаар сургалт хийж туршихад 83%-д нь алдаатай үгийг автоматаар засварласан.



Зураг 5: Bert загвар дээр туршсан үр дүн.

4.2. SymSpell загвар дээр хийсэн туршилт

Ц. Дамдинсүрэн, Б. Осор нарын “Зөв бичгийн дүрмийн толь” дээр үндэслэсэн нийт 2 сая үгийн сангаас сонгогдсон 1048075 үг болон 1498 сөрөг үгийн сан дээр сургалт хийсэн ба туршилт хийхдээ SymSpell загвартаа сангаас 12 сая холбоо үгийн толь бичгийг үүсгэн ашигласан.Засварлах зайг нь хамгийн ихдээ 2-оор авсан. Доорх жишээнд элемэнтийг гэдэг үгийг элементийг гэж зассан.

```
Enter sentence :программ хангамж техник хангамж өгөгдөл хүмүүс үйл ажиллагааг багтаасан мэдээллийн системийн талаарх элемэнтийг
бий болгосон
программ хангамж техник хангамж өгөгдөл хүмүүс үйл ажиллагааг багтаасан мэдээллийн системийн талаарх элемэнтийг бий болгосон
run spell checker...

original text: программ хангамж техник хангамж өгөгдөл хүмүүс үйл ажиллагааг багтаасан мэдээллийн системийн талаарх элемэнтийг
бий болгосон

corrected text: программ хангамж техник хангамж өгөгдөл хүмүүс үйл ажиллагааг багтаасан мэдээллийн системийн талаарх элементийг
бий болгосон
```

Зураг 6: SymSpell загвар дээр туршсан үр дүн.

5. Дүгнэлт

SymSpell загварыг 1 сая гаран өгөгдөл дээр сургаж, туршилт хийхэд дунджаар 80%-д нь алдааг зөв олж, засварлаж байсан. Харин 2 чиглэлтэй кодлогч трансформер болох BERT загварыг нийт 2 сая өгөгдөлтэй сан дээр сургаж, туршилт хийхэд дунджаар 83%-ийн нарийвчлалтай үр дүнд хүрсэн. Сургалтын өгөгдлийн хувьд SymSpell загварыг 2 сая үгтэй өгөгдөлтэй сан дээр сургах үед санах ойн алдаа гарч байсан тул хэмжээний хувьд BERT загвараас том хэдий ч үр дүнгийн хувьд төстэй байсан нь эдгээр загваруудыг цааш харьцуулан судлах хэрэгтэйг харуулж байгаа бөгөөд гарсан үр дүнгээс уг загваруудыг ашиглан монгол хэлний бичвэрийн алдааг илрүүлэн засварлах боломжтой гэж дүгнэж байна.

Ном зүй

[1] "Autocorrect Feature using NLP in Python," *https://www.analyticsvidhya.com/blog/2021/11/-autocorrect-feature-using-nlp-in-python/*, 2021.

[2] "Хөрвүүлэх программд зориулсан монгол хэлний цахим хөмрөгийн тухай," *Др. Э. Мөнх-Учрал (Боннын Их Сургууль, Монгол Улсын Их Сургууль)*, 2010.

[3] М. Хүрэлхүү, Д. Уганбаатар, "Машин сургалтын аргыг кирилл, монгол бичгийн алдаа засах, бичвэр хооронд хөрвүүлэхэд ашиглах нь," 2019.

[4] "Transformers in NLP: A beginner friendly explanation," *https://towardsdatascience.com-/transformers-89034557de14*.

[5] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *https://arxiv.org/pdf/1810.04805.pdf*, 2019.

[6] "SymSpellEx," *https://www.npmjs.com/package/symspell-ex/v/1.0.2*, 2020.

[7] *https://en.wikipedia.org/wiki/BERT_(language_model)#cite_note-:0-1*.

[8] *https://github.com/wolfgarbe/SymSpell*.

[9] "A quick overview of the implementation of a fast spelling correction algorithm," *//medium.com-/@agusnavce/a-quick-overview-of-the-implementation-of-a-fast-spelling-correction-algorithm-39a483a81ddc#:~:text=SymSpell,in%20check%20by%20prefix%20indexing*.

Cyrillic Word Error Program Based on Machine Learning

Khajidmaa Battumur[1*], Uuganbaatar Dulamragchaa[1], Sodjamts Enkhbat[1], Lhagvasuren Altankhuyag[1], Purevsuren Tumurbaatar[2]

[1] *Institute of Mathematics and Digital Technology, Mongolian Academy of Sciences, Ulaanbaatar 13330, Mongolia*
[2] *Institute of Language and Literature, Mongolian Academy of Sciences, Ulaanbaatar 13330, Mongolia*

**Corresponding author: khajidmaa_b@mas.ac.mn; ORCID:0000-0003-4786-0836*

Article Info: Received: 2022.09.11; Accepted: 2022.10.01; Published: 2022.12.26

Abstract: With the rapid development of information technology, the main means of human communication-language-is also shifting rapidly from physical to digital forms. That being said, natural language processing research on foreign languages have been conducted regularly with ready-to-use and reliable programs already widely available on the market. While, initially, spellchecking programs were developed using traditional methods like n-gram methods, modern approaches embrace machine learning methods. Hence, with this research, we aimed to train a natural language processing models more suitable for the Mongolian language structure of which BERT and SymSpell models were trained and tested.

Key words: BERT model, SymSpell model, Natural Language Processing, Spellchecking