Хайлтын системд зориулсан монгол хэлний үгзүйн задлуур, үүсгүүр

Мөнхжаргалын Золжаргал*, Чагнаагийн Алтангэрэл*, Дамбасүрэнгийн Нанзадрагчаа*, Керейн Есенбек*

* Машин Оюуны Лаборатори, Монгол Улсын Их Сургууль, Улаанбаатар, Монгол

{zoljargal, altangerel, nanzadragchaa, esenbek}@num.edu.mn

***Хураангуй*—Тус өгүүлэл нь хайлтын системийн асуулга баяжуулах, их өгөгдлийг шахахад үг зүйн задлуур/үүсгүүрийг ашиглах боломж, монгол хэлний төгсгөлөг төлөвт үгзүйн задлуур, үүсгүүр болон туршилтын үр дүнгийн талаар өгүүлнэ. Бид зөв бичгийн дүрмийг үгзүйн задлалд төвлөрч хийсэн ба төгсгөлөг төлөвт програм нь үгийн нөхцөлийг залгах боломжтой. Нэр үгийн морфотактикийг бүрэн, үйл үгийн морфотактикийг нэг нөхцөлийн хувьд оруулсан. Задлуурын туршилтыг цэвэрлэсэн болон практик хэрэглээнд байдаг материалын сан дээр туршин сайжруулах зарчмаар хөгжүүлсэн.**

Түлхүүр үг— үгзүйн задлуур, төгсгөлөг төлөв, монгол хэл, зөв бичгийн дүрэм

I. Удиртгал

Үгзүйн задлуур нь хэлбэржсэн үгийн үндсийг олдог хэрэгсэл програм ба хэлийг компьютероор боловсруулахад суурь хэрэгсэл юм. Ялангуяа үгзүйн дүрэм төвөгтэй хэлнүүдэд илүү чухал байдаг [8]. Үгзүйн задлуурыг их өгөгдөл боловсруулдаг системүүд тэр дундаа мэдээлэл олборлох (Information Retrieval) системд хэлбэржсэн үгийн нөхцөлийг салгаж өгөгдлийн хэмжээг багасгахад ашигладаг. Түүнчлэн хэрэглэгчийн оруулсан асуулгын (query) алдааг засах, асуулга дахь үгийн нөхцөлийг салгаж хайлтын үр дүнг нэмэгдүүлэхэд ашиглаж байна. Орчин үеийн хайлтын системүүд (Google, Bing г.м.) нь англи хэлний боловсруулалтыг хэрэглэгчид оновчтой хайлтын үр дүн гаргахад ашигладаг ч монгол хэлний боловсруулалт хараахан хийгдэхгүй байна. Нөгөөтээгүүр манай улсад монгол агуулгаас оновчтой хайлт хийх зорилготой системүүд (moogol.mn, map.minu.mn, legalinfo.mn г.м.) хөгжүүлж байгаа ч монгол хэлний боловсруулалт ашиглахгүй мөн байна.

Өгүүлэл нь монгол хэлний хоёр түвшинт дүрэмд (two-level rule) суурилсан төгсгөлөг төлөвт үгзүйн задлуур болон хайлтын системд зориулсан өгөгдлийн хэмжээг багасгах, хэрэглэгчийн асуулгыг баяжуулах туршилтын талаар өгүүлнэ. Үгзүйн үүсгүүр нь задлуурын эсрэг процесс ба төгсгөлөг төлөвт автоматын хэрэгсэл програм үүнийг автоматаар хийдэг учир судалгааны ажлыг үгзүйн задлуурт төвлөрч дүрэм, үгийн санг тодорхойлсон.

Өмнө нь монгол хэлэнд үгзүйн задлуурын хэд хэдэн ажил хийгдсэн [1, 5]. Энэ удаагийн ажлын онцлог нь бодит хэрэглээнд байгаа, нээлттэй эхийн материалын сан дээр задлуур програмын чанарыг туршин сайжруулах замаар хөгжүүлсэн явдал юм. Ингэснээр практик ач холбогдол өндөр болно.

II. Судлагдсан Байдал

Төгсгөлөг төлөвт автоматаар монгол хэлний үгзүйн дүрмийг тодорхойлсон анхны оролдлого нь [5] байсан. Уг ажлаар юникод дэмждэггүй PC-KIMMO [8] төгсгөлөг төлөвт автоматыг хэрэгжүүлсэн хэрэгсэл програм дээр монгол хэлний дүрмийн тодорхойлолтыг бичсэн. Кирилл үсгийг таниулахын тулд 2 байт кодоор илэрхийлэгдэх кирилл үсгийг 1 байт англи үсэг рүү хөрвүүлдэг дундын програм ашиглаж байсан. Уг задлуур нь 36 үгзүйн дүрэм, 6,199 нэр үг, 18,551 үйл үг, 4,516 тэмдэг нэр бүхий үгийн сантай. Програмыг бага хэмжээний материалын сан дээр туршихад 63 хувийг задалж байжээ.

Дээрх ажлыг Ч.Алтангэрэл нар 84 дүрэм, 25 дэд хэсэгтэй болгож өргөтгөсөн [1]. Туршилтыг Ц.Дамдинсүрэн нарын монгол үсгийн дүрмийн толиноос түүвэрлэж авсан 2,000 үг дээр хийсэн. Уг ажлын дутагдалтай тал нь ижил хэлбэртэй хирнээ ялгаатай хувирдаг үгийн асуудлыг шийдэж чадаагүй.

Бадам-Осор нар [10] мэдээлэл олборлох програмын индекслэлд зориулсан монгол хэлний үгийн үндэс олох судалгааны ажлыг хийсэн. Аргазүй нь нэр болон үйлийн нөхцөлийн толь ашиглан үгийн араас хайлт хийж үндсийг олж, эгшиг жийрэглэх, гээх дүрмийг оруулсан. Мөн 1102 техникийн өгүүллийн нэр, хураангуй, түлхүүр үгийг индексэлж түлхүүр үгээр хайх туршилтыг хийсэн ба гол агуулга илэрхийлж байгаа үгийг хураангуйгаас хайхдаа үгзүйн задлуураар үндсийг олсон байна.

Дээрх ажлуудад үгзүйн задлуурт төвлөрсөн ба өгөгдөл шахах, асуулга баяжуулах талаар туршилт хийгдээгүй байдаг.

III. Төгсгөлөг Төлөвт Үгзүйн Задлуур

Үгзүйн задлуурт хоёр төрлийн төгсгөлөг төлөвт автоматыг хэрэгжүүлдэг. А) морфотактик болон Б) хоёр түвшинт дүрэм (Монгол хэлний морфотактикийн талаар IV бүлгээс үз).

Морфотактикийг төгсгөлөг төлөвт автоматад үргэлжилсэн үгийн сан хэлбэрээр илэрхийлнэ. Тухайлбал тийн ялгалын араас тийн ялгал орох дарааллыг Зураг 1-д “харуул” гэдэг үг дээр харьяалахын тийн ялгал болон хамтрахын тийн ялгал зэрэг залгах үгийн сангийн загварыг харуулав.

Үгийн сангийн аргаар хувирсан үгийн эгшиг зохицох ёс, эгшиг гээх, жийрэглэх зэргийг хоёр түвшинт дүрмээр гүйцээнэ. Зураг 1-ийн дагуу харуул гэдэг үгийг хувилгавал “харуул+ЫЫн+тЙй” болно.

Энд байгаа тусгай тэмдэглэгээнүүдийг (Ы, Й) хоёр түвшинт дүрмээр оруулж тохирох үсгээр солино (V бүлгийг үз).

```
Lexicon Noun          Lexicon Gen
   харуул n-inf ;       ЫЫн ;
Lexicon n-inf           ЫЫн+Com ;
   ;                  Lexicon Com
   Gen ;                тЙй;
```

Зураг 1. Морфотактикийг үгийн сангийн аргаар илэрхийлсэн байдал. Энд Noun: нэр үг, n-inf: нэрийн хувилал, Gen: харьяалахын тийн ялгал, Com: хамтрахын тийн ялгал.

Ц.Дамдинсүрэн, Б.Осор нарын Монгол Үсгийн Дүрмийн Толийг [9] баримтлан хоёр түвшинт дүрмийг бичлээ.

Бидний үүсгэсэн задлуур нь үгийн үндэс, нөхцөлийг илэрхийлэх 84 бүлэг тэмдэглэгээтэй ба үүнээс 16 нь нэр, үйл, оноосон нэр гэх мэт үндсэн үгийн аймгийг, 68 нь тоо, тийн ялгал, хэв, байдал зэрэг нөхцөлийг төлөөлнө. 68 тэмдэглэгээний бүлэг нь дотроо нөхцөлийн хувилбарыг илэрхийлсэн дэд тэмдэглэгээ агуулна. Жишээ нь хамтрахын тийн ялгалыг “Com” гэж ерөнхийлөн тэмдэглээд “тЙй” (тай[4]) –г “Com1”, эртний хувилбар болох “лУгАА” (лугаа[2])-г “Com2” гэж тусад нь тэмдэглэнэ. Энэ мэт дэд нөхцөлийн хувилбарыг илэрхийлэх тэмдэглэгээ 195. Нэг үгэнд орсон олон нөхцөлийг нэмэх тэмдгээр (+) тусгаарлана.

Задлуур програм нь морфотактикийг илэрхийлсэн 221 үгийн сан (Lexicon), 55 хоёр түвшинт дүрэмтэй. Хүснэгт 1 –д бидний ашигласан үндсийн тоог үгийн аймгаар бүлэглэн харуулав.

ХҮСНЭГТ 1. Үгийн сангийн хэмжээ

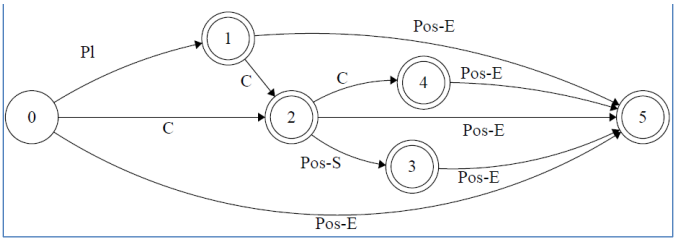
Аймаг	**Үндсийн тоо**
Нэр (тэмдэг нэр, оноосон нэр г.м.)	11,580
Үйл	12,732
Бусад (хувирдаггүй)	1,800

IV. Морфотактик

Морфотактик гэдэг нь тухайн хэлний дүрмийн дагуу дагавар, нөхцөл ар араасаа залгагдах дараалал юм. Үг нөхцөлөөр хувирсан хэлбэрийг нэг төлөв гэж үзнэ. Тухайлбал нэр үгэнд олон тооны нөхцөл залгавал “олон тоогоор хувирсан” төлөвт орно. Араас нь тийн ялгал залгавал “олон тоогоор хувирсан” төлөвөөр дамжин “тийн ялгалаар хувирсан” төлөвт шилжинэ.

Бид энэ удаагийн ажлаар зөвхөн нэр, үйлийн нөхцөлийн хувиллыг авч үзсэн. Харин дагавраар үүссэн үгийг үгийн санд багтаасан.

Нэрийн морфотактик нь тэмдэг нэр, оноосон нэр, төлөөний үг, тооны нэр зэрэг үгийн аймгийг хамтатгаж авч үзнэ. Бидний хөгжүүлсэн задлуур нь Зураг 2-ын дагуу 300 орчим нөхцөлийн дараалыг үүсгэсэн.



Зураг 2. Нэр үгийн нөхцөлийн морфотактик. Энд Pl: олон тооны нөхцөл, C: тийн ялгал, Pos-E: хамаатуулах нөхцөл, Pos-S: хамаатуулах уттат хэсэг.

Монгол хэлний үлийн морфотактикийн дагуу үүсэх нөхцөлийн боломжит дараалал нэр үгийг бодвол харьцангуй олон юм [3]. Үйл үг үүсгэх дагаврыг оролцуулалгүй зөвхөн нөхцөлийн боломжит дарааллыг тооцоход ойролцоогоор 24 мянга орчим байна [3]. Үйлийн нөхцөлийн бүх дарааллыг төгсгөлөг төлөвт үгийн санд оруулах нь цаг их авна. Тиймээс бид энэ удаагийн ажлаар зөвхөн нэг нөхцөлийн түвшинд үйлийг задлахаар хийлээ. Нөгөөтээгүүр Компьютер Хэл Шинжлэлийн Судалгааны Төвийн (одоогоор Машин Оюуны Лаборатори) үгийн аймгийн тэмдэглэгээт “МУИС” материалын сангийн [4] 2.7 сая үгийн 516,246 нь үйл үг, эдгээрээс 65,299 нь буюу нийт үйл үгийн 12.6% нь хоёроос олон нөхцөлөөр хувирсан байна. Иймд эхний ээлжид нэг нөхцөлөөр хувирсан үйлийн задлуурыг хийхэд практик ач холбогдлоо алдахгүй.

A. *Морфологи Болон Үг*

Монгол хэлний үг нь хоосон зайгаар тусгаарлагдсан байдаг ба зарим онцгой тохиолдол, оноосон нэр зэргээс бусад тохиолдолд хоосон зай хоорондын тэмдэгтийн цуваанд нэгээс илүү үг байдаггүй. Хэрэв нэгээс олон үгээс бүтсэн бол үгийг тусгаарладаггүй. Тиймээс хоёр түвшинт дүрэм нь хоосон зай хооронд үйлчилнэ. Харин чиглэхийн тийн ялгалын “руу/рүү”, асуух туслах үг “юу/юү”, “уу/үү” зэрэг нь өмнөх үгээс хоосон зайгаар тусгаарлагддаг хирнээ эгшиг зохицох ёсыг баримталдаг. Энэ мэт онцгой тохиолдлуудыг хоёр түвшинт дүрэмд тусгаж өгсөн.

B. *Салаа Утгатай Задардаг Үгс*

Уг гарвалиасаа шалтгаалж яг ижил хэлбэртэй хирнээ ялгаатай нөхцөлөөр хувирдаг үг монгол хэлэнд элбэг байдаг. Тухайлбал “дарга”, “чарга” гэдэг хоёр үг хэлбэрийн хувьд ижил ч харьяалахын тийн ялгалаар хувирахдаа харгалзан “-ын”, “-ны” гэсэн хоёр өөр хувилбар авч “даргын”, “чарганы” болдог. Иймээс үгсийг хоёр түвшинт дүрмийн аргаар ялгах боломжгүй юм.

Мөн хэлбэр нь ижил утга нь ялгаатай үгс нэг нөхцөлөөр хувирахдаа тухайн нөхцөлийн өөр өөр хувилбарыг авна. Жишээлбэл “гэрийн өрх” гэдгийн “өрх”-ийг харьяалахын тийн ялгалаар хувилгавал “өрхний” болно. Харин “өрх гэр” гэдгийн “өрх”-ийг хувилгавал “өрхийн” болно.

Дээрх дүрмээр ерөнхийлөн тодорхойлох боломжгүй тохиолдлуудыг үгийн сангийн аргаар шийдсэн.

V. Хоёр Түвшинт Дүрэм

Зөв бичгийн дүрмийг төгсгөлөг төлөвт автоматад хэрэгжүүлэхдээ үсгүүдийг бүлэг болгож тусгай

тэмдэглэгээгээр (archiphoneme) тэмдэглэсэн. Хүснэгт 2-т тэмдэглэгээг харуулав.

ХҮСНЭГТ 2. Үсгийн тусгай тэмдэглэгээ

Тэмд.	**Үсэг**	**Тэмд.**	**Үсэг**	**Тэмд.**	**Үсэг**
Б	б, в	А	а, э	Й	а, э, о
Д	д, т	Э	а, о	Ө	а, и, о
Ж	ж, ч	И	а, и	Ү	а, э, о, ө
Ы	й, и, ы	У	у, ү	Я	я, ё, е
Ъ	ъ, ь				

Үсгийн тусгай тэмдэглэгээг жишээгээр тайлбарлавал: “чарга+х.тя” хувиллыг “чарга+нЫЫ” гэж тэмдэглэнэ. Энд байгаа “ЫЫ” нь эгшиг зохицох дүрмийн дагуу эр үгэнд “ы”, эм үгэнд “ий” болно.

A. Эгшиг Зохицох

Монгол хэлний эгшиг зохицох ёсыг хоёр түвшинт дүрмээр илэрхийлсэн. Эгшиг зохицох ёсонд орж байгаа тусгай тэмдэглэгээнүүд {Ү}, {Й}, {У}, {Ы}, {А}, {Я} болно. Жишээ нь “оруул+нҮ” гэж илэрхийлээд “Ү” тэмдэглэгээг эгшиг зохицох ёсоор “а, э, о, ө” эгшгүүдийн аль нэгнээр орлуулна. Хүснэгт 3., 4.-т эгшиг зохицлын жишээг харуулав.

ХҮСНЭГТ 3. {Ү} тэмдэглэгээний эгшиг зохицол

Өмнөх эгшиг	**Хувилал**	**Өмнөх эгшиг**	**Хувилал**
а, аа, ай, я, яа	а	и, ий	э
у, уу, уй, юу	а	ө, өө, еө	ө
э, ээ, эй	э	о, оо, ой, ё, ёо	о
ү, үү, үй, юү	э		

ХҮСНЭГТ 4. {Я} тэмдэглэгээний эгшиг зохицол

Өмнөх эгшиг	**Хувилал**	**Өмнөх эгшиг**	**Хувилал**
а, аа, ай, я, яа	я	и, ий	е
у, уу, уй, юу	я	ө, өө, еө	е
э, ээ, эй	е	о, оо, ой, ё, ёо	ё
ү, үү, үй, юү	е		

B. Авиа Ижилсэх

Монгол хэлний авиа ижилсэх ёсоор {Д}, {Б}, {Ж} тусгай тэмдэглэгээнүүдийг тодорхойлсон. Эдгээр нь мөн хоёр түвшинт дүрмээр зохицуулагдана.

C. Гээгдэх Эгшиг

Үгийн эцсийн болон эцсийн гийгүүлэгчийн өмнөх балархай эгшгийг гээх хоёр тохиолдлыг дүрэмчилсэн. Тэгэхдээ эцсийн гийгүүлэгчийн өмнөх балархай эгшгийг зарим уламжлалын зарчим баримталж бичсэн үгэнд гээдэггүй [9]. Жишээ нь “тэнхим”, “галав”, “төлөв” гэх мэт. Энэ мэт онцгой тохиолдол учирдаг тул бүх үгийг мөн ерөнхийлөн дүрэмчлэх боломжгүй болгож байгаа юм. Тиймээс гээгдэх эгшигтэй үгийн үндэс бүрийн балархай эгшгийг тусгай тэмдэглэгээгээр тэмдэглэсэн. Тухайлбал “саат{а}л” гэж үндсийг тэмдэглээд гээх тохиолдолд “{а}” тэмдэглэгээг байхгүй болгоно.

D. Зөөлний тэмдэг солигдох нь

Монгол хэлний зөв бичгийн дүрмийн дагуу зөөлний тэмдгийг “и” болгож хувиргах нь ямар нэгэн онцгой тохиолдол байхгүй учир хоёр түвшинт дүрмээр бүрэн илэрхийлж болно.

E. Тусгаарлагч Тэмдэг Үсэг

Эр, эм үгэнд туслах я, ё, е эгшгүүд өмнөх гийгүүлэгчээсээ саланги дуудагдахаар орвол өмнө нь хатуу, зөөлний тэмдгээр тусгаарладаг [9]. Энэ дүрэм нь хоёр түвшинт дүрмээр ямар нэгэн хоёрдмол утгагүйгээр илэрхийлэхэд асуудал тулгараагүй.

F. Олон Тоо

Монгол хэлний нэр үг хэд хэдэн олон тооны нөхцөлөөр хувирах тохиолдол элбэг байдаг. Тухайлбал “хөгшин” гэдэг үг “хөгшид”, “хөгшчүүл”, “хөгшчүүд” гэсэн олон тооны нөхцөл авна. Мөн үгийн үндсийн хэлбэрийг бүрэн хувилбар ч байна. Жишээ нь “өвгөн” гэдгээс “өтгөс” гэх мэт. Дээрх дүрмийн бус байдлыг харгалзан нэрийн олон тоо бүрийг үгийн үндсэнд шууд залгаж үгийн санд оруулсан.

VI. Хайлтын Асуулга Баяжуулах

Бичвэрээс хайлт хийхэд асуулга нь бичвэрт байгаа үгтэй таарахгүй байх нь элбэг. Тухайлбал Google мэт хайлтын систем нь одоогоор хэрэглэгчийн оруулсан асуулгыг бичвэр харьцуулах зарчмаар хайдаг ба ингэж хайхад хэрэглэгч өөрт хэрэгтэй үр дүнг оновчтой олж чаддаггүй.

Хайлтыг оновчтой болгохын тулд хэрэглэгчийн асуулгыг баяжуулж илүү олон хувилбараар хайлт хийж болдог. Одоогийн системүүд нь үгзүйн түвшинд а) хэрэв хэрэглэгч зөв бичгийн алдаатай үг оруулсан бол автоматаар засаад хайх, б) асуулга дахь нөхцөлөөр хувирсан үгийн үндсээр хайх гэсэн хоёр аргаар асуулгыг баяжуулдаг.

Судалгааны ажлаар дээрх хоёр асуулга баяжуулах аргыг хэрэглээнд байгаа газарзүйн хайлтын системд туршиж үзсэн.

VII. Бичвэр Өгөгдөл Шахах

Монгол хэлний үг нь өгүүлбэрийн харьцаанд ороходоо нөхцөлөөр хувирч үгийн урт нэмэгддэг. Иймд үгийн нөхцөлийг салгаж үндсээр нь индекслэлт хийвэл бичвэр өгөгдлийн хэмжээ багасна гэсэн таамаг дэвшүүлж туршилтаар баталгаажуулахыг оролдлоо.

Тэгэхдээ хэлбэржсэн үгийн нөхцөлийг орхигдуулж дан ганц үндсээр индекс хийвэл хайлтын оновчгүй үр дүн хэт ихсэх, өгүүлбэрийн утга алдагдах талтай. Тиймээс бид үг болон нөхцөлийг хамтад нь агуулдаг хирнээ эх үгээс богино тэмдэгтийн цуваа үүсгэдэг энгийн алгоритмыг боловсруулсан.

Алгоритм №-1:

1. Хувирсан үгээс нөхцөлүүдийг салгана.
2. Салгасан нөхцөлүүдийг хамтад нь нэг богино тэмдэглэгээгээр тэмдэглэнэ.
3. Үгийн үндэс дээр богино тэмдэглэгээг залгана. Хэрэв үг салаа утгатай задарсан бол хамгийн богино үндэстэйг сонгоно.
4. Хувирсан үгийн эх болон богино тэмдэглэгээт хувилбарын уртыг харьцуулж аль богиныг нь авна.

A. Нөхцөлийн Богино Тэмдэглэгээ

Туршилтын материалын сангийн бүх үгийн нөхцөлийг задлахад 279 янзын нөхцөлийн дараалал гарсан (Хүснэгт 5).

ХҮСНЭГТ 5. Нөхцөлийн дарааллын жишээ. Энд нэг нөхцөлийн ялгаатай хувилбарыг нэг тэмдэглэгээгээр тэмдэглэнэ.

Нөхцөлийн дараалал	**Жишээ**
N+Gen	*харуулын = харуул+N+Gen, чарганы=чарга+N+Gen*
N+Gen+Dat+Refl	*аавындаа = аав+N+Gen+Dat+Refl*

Нөхцөлийн боломжит дарааллуудыг давтамжаар нь эрэмбэлээд эхнээс нь дугаарласан ба уг дугаар нь нөөцлөхийн дарааллын товч тэмдэглэгээ болно. Давтамжаар нь эрэмбэлснээр их тохиолддог нөхцөлийн дараалал илүү богино буюу нэг эсвэл хоёр оронтой тоогоор илэрхийлэгдэнэ. Бидний туршилтаар хамгийн их хэрэглэгддэг нөхцөлийн дараалал харьяалахын тийн ялгал дангаараа орсон тохиолдол байсан ба 1 гэж тэмдэглэнэ. Нөхцөлийн богино тэмдэглэгээг ашиглаж үгийг бичвэл "*харуулын"* гэдгээс "*харуул1"* болно. Олон нөхцөлөөр хувирсан тохиолдолд мөн тухайн олон нөхцөлийн дараалалд харгалзах кодоор тэмдэглэнэ. Жишээ нь: "*аавындаа"* гэдгээс "*аав87"* болно.

VIII. Туршилт

Үгзүйн задлуурын чанарыг хэрэглээнд байгаа бичвэрээр шалгах нь илүү практик ач холбогдолтой. Иймд МУИС материалын сан болон 2016 оны 3 сарын 5-ны Монгол Википедиа санг ашиглан шалгасан.

Асуулга баяжуулах туршилтыг бодит хайлтын системд туршсан. Харин өгөгдөл шахах туршилтыг gogo.mn мэдээний веб сайтын 2007-2010 оны бүх мэдээн дээр хийсэн.

Бидний үүсгэсэн үгзүйн задлуур нь секундэд 60,000-80,000 үг задлах, үүсгэх хурдтай. Хурдны туршилтыг dell optiplex 340 загварын i5-2.9 GhZ процессор, 8 гига байт санах ойтой компьютероор хийсэн. Энэ хурд нь олон хэрэглэгчтэй зэрэг ажилладаг хайлтын системд үгзүйн боловсруулалт хийхэд хангалттай юм.

A. Үгзүйн Задлуур

МУИС материалын сан болон Википедиа өгөгдлийн сангийн хэдэн хувийг нь задалж байгааг тодорхойлсон (Хүснэгт 6).

Үйл үгийн хувийг тооцохдоо монгол хэлний үгийн аймгийн тэмдэглүүрийг [7] ашигласан. Википедиа нийтлэлийн тусгай тэмдэглэгээг (html таг, wiki тэмдэглэгээ г.м) Wikipedia_Extractor[1] хэрэгслээр ялгасан.

ХҮСНЭГТ 6. Туршилтын үр дүн

Материалын сан	**Үгийн тоо**	**Задалсан**	**Хувь**	**Дундаж задлал**	**Үйл %**
Википедиа	4,258,510	3,458,739	81.2	1.42	14.5
МУИС	2,638,621	2,377,606	90.1	1.60	19.0

Задлуур МУИС материалын сангийн 2,638,621 үгээс 261,015-ыг задалж чадаагүй. Нэг үгийг дунджаар 1.6 янзаар задалсан. Нэг үг дунджаар 2.37 нөхцөл авсан байна. Википедиа сангийн 4,258,510 үгээс 799,771 үгийг задлаагүй. Нэг үгийг дунджаар 1.42 янзаар задалсан. Нэг үг дунджаар 3.25 нөхцөл авсан байна.

Википедиа сангийн задлаагүй үгсийг шинжихэд а) буруу бичсэн үг, б) шинээр үүссэн үг, в) гадаад үгс элбэг тохиолдож байна. Энэ нь Википедиаг олон нийтийн хүчийг ашиглан үүсгэдэгтэй холбоотой. Мөн сангаас тоо, цэг, таслал, асуултын тэмдэг, хаалт зэрэг тусгай тэмдэглэгээнүүдийг цэвэрлээгүй учир задалж чадаагүй үгс дунд эдгээр утгат хэсгүүд багтсан гэдгийг анхаарах хэрэгтэй. Мөн Википедиа сангийн нэг үгийн нөхцөлийн дундаж тоо МУИС сангаас их байгаа нь үйл үгийн эзлэх хувь бага, нэр үгэнд хувилал их хэрэглэснээс болж байна.

B. Асуулга Баяжуулах

Туршилтыг шинээр хөгжүүлж байгаа Улаанбаатар хотын газарзүйн хайлтын системд хийсэн. Хайлтын системээс хайхдаа асуулга дахь үг бүрийн үндсийг олоод эх хувилбартай "OR" хийж хайна. Жишээ нь "банкны салбар" гэж хэрэглэгч хайсан бол задлуур "банк салбар" гэж үндсийг олоод "(банкны салбар) OR (банк салбар)" гэж хайна.

Одоогийн байдлаар хайлтын системд өдөрт 150-200 асуулга тавьдаг ба үгзүйн задлуур дунджаар 65.1 хувийг асуулгын үр дүнд оновчтой байдлаар үгийн үндсийг олж байна. Тухайлбал "ногооны зах" гэдгийг "ногоо зах" болгох бөгөөд хайлтын үр дүн ижил агуулгатай байна.

Үгзүйн задлуур хувирсан үгийн боломжит бүх задлалыг хийдэг бөгөөд хайлтын систем тэдгээрээс хамгийн эхний хувилбарыг сонгож байгаа тул хэрэглэгчийн оруулсан асуулгын жинхэнэ үндэстэй тохирохгүй байх нь элбэг байна. Тухайлбал "спорт хороо" гэсэн асуулгын "хороо" гэдэг үгийг задлуур "хор" гэдэг үндэс дээр ерөнхийлөн хамаатуулах нөхцөл "оо", "хороо" гэсэн нэрлэхийн тийн ялгалд байгаа үг гэж хоёр янзаар задлах ба асуулга баяжуулах алгоритм нь эхний хувилбарыг сонгож байна.

Хайлтын системийн асуулгын 95 хувь нь нэрийн холбоо үг байна. Энэ нь өнөөгийн хайлтын системийн боломжид тохируулан хэрэглэгч оноосон нэр, түлхүүр үгээр хайж заншсанаас болж байгаа болов уу. Мөн бидний туршсан газарзүйн хайлтын системээс хаяг, албан байгууллагын нэр, тухайн байгууллагын үйл ажиллагааны чиглэл (хүүхдийн дэлгүүр, банк г.м), шинжээр ихэвчлэн хайж байна. Монгол хэлний нэрийн холбоо үгийн цөм үг нь хамгийн ардаа байрладаг ба хэрэглэгчийн асуулгыг шинжилхэд цөм үгээс бусад үгийн нөхцөлийн хувиллыг зөв оруулж байна. Энэ ажиглалт болон үгийн үндсийг олсон асуулгын утгыг харьцуулахад үгзүйн задлуураар асуулгад байгаа бүх үгийн үндсийг олох нь эх асуулгын утга алдагдахад хүргэж байна. Тухайлбал хэрэглэгч "элчин сайдын яамны" гэж хайхад үгзүйн задлуур "элчин сайд яам" болгох ба хайлтын үр дүнд "элчин сайд"-тай холбоотой илэрцүүд гарна. Тиймээс нэрийн хэлцээр түлхүү хайдаг хайлтын системд зөвхөн цөм үгийн үндсийг олох нь тохиромжтой харагдаж байна.

C. Бичвэр Өгөгдөл Шахах

Gogo.mn сайтын түүхий материалын сан нь 48,900 нийтлэл, 23,259,822 үгтэй (token) ба нийт файлын

1 http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

хэмжээ 256.7 мига байт. Туршилтад ашиглахдаа ямар нэгэн цэвэрлэгээ хийгээгүй болно.

Үгзүйн задлуур 5,847,987 буюу нийт үгийн 25.1 хувийг задалж чадаагүй ба шинжлэхэд латин үсгээр бичигдсэн үг, зөв бичгийн алдаатай үг, тэмдэгт, тоо, оноосон нэр зэрэг орсон байна.

Үгзүйн задлууpаар задалсан үгнүүдэд Алгоритм №-1-ийг ажиллуулахад 5,670,353 ширхэг хувирсан үгийг богино тэмдэглэгээгээр орлуулж файлын хэмжээг 232.5 мига байт буюу 9.4 хувиар багасгасан байна.

Файлын хэмжээг 9.4 хувиар багасгаж байгаа нь хэлбэржсэн үгийн тоотой харьцуулахад хангалттай үр дүн биш юм. Энэ нь доорх шалтгаануудаас болсон гэж үзэж байна:

1. Практик хэрэглээнд олон нөхцөл дараалуулан хэрэглэдэггүй.
2. Задлуур нь үгийн үндсийг олохдоо монгол хэлний зөв бичгийн дүрмийн дагуу ажиллаж байгаа учир гээгдсэн эгшгийг буцаан үгийн үндсэнд оруулна. Иймд үндсийн урт нэмэгдэх талтай.
3. Практик хэрэглээнд оноосон нэр их хэрэглэж байна. Харин үгзүйн задлуур нь хэрэглээнд байгаа бүх оноосон нэрийг үгийн сандаа багтаан үндсийг нь олж чадахгүй байна.

IX. Дүгнэлт

Судалгааны ажлаар монгол хэлний үгзүйн задлуурыг хөгжүүлж хайлтын системийн асуулга баяжуулах, бичвэр өгөгдлийн хэмжээг багасгах туршилтыг хийлээ. Задлуур програмд монгол хэлний зөв бичих дүрэм болон нэр үгийн морфотактикийг бүрэн хэрэгжүүлсэн. Үйл үгийн морфотактикийг нэг нөхцөлийн хувьд хэрэгжүүлсэн нь практик ач холбогдлоо алдаагүй. Задлуурын чанарыг МУИС материалын сан болон Монгол Википедиа сан дээр туршсан ба харгалзан 90.1, 81.2 гэсэн хувьтай задалж байна.

Задлуур нь бидний туршилтын компьютер дээр секундэд 60,000-80,000 үгийг боловсруулж 96,000-128,000 үр дүн гаргаж байна.

Асуулга баяжуулах туршилтаар 65.1 хувийн үгзүйн задлал хийж байна. Өнөөгийн хайлтын системүүд нь түлхүүр үгээр индекс хийдэг учир дан ганц үгзүйн задлууpаар асуулгыг баяжуулах нь учир дутагдалтай байгаа нь туршилтаар ажиглагдсан. Бидний үүсгэсэн задлуурын санал болгож байгаа үгийн үндсүүдээс үгзүйн салаа утга таниураар (Morphological disambiguator) тухайн асуулгын утгад тохирох үндсийг санал болгох судалгааны ажлыг хийх шаардлагатай нь харагдаж байна.

Түүнчлэн хаягийг илэрхийлэх нүүдэлчин соёл нь олны танил газраас баримжаалж хэлдэг. Тухайлбал “3-р сургуулийн зүүн талын ногоон байшин”. Иймхайлтад оновчтой үр дүн гаргахын тулд монгол хэлний өгүүлбэрзүйн шинжлүүр, нэрлэсэн нэгж таниур, холбоо үг таниур, үгийн утга тодорхойлуурыг хамтад хэрэглэх нь зүйтэй.

Хэлбэржсэн үгийн нөхцөлийг богино тэмдэглэгээгээр орлуулах зарчмаар 23 сая үгтэй материалын сангийн хэмжээг багасгах туршилт хийхэд хэмжээг 9.4 хувиар багасгасан үр дүн гарсан.

Цаашид оноосон нэр, хэрэглээнд шинээр орж ирсэн үгсийг програмын үгийн санд нэмэх ажлыг хийх шаардлагатай. Үйл үгийн морфотактикийг гүйцээх нь нэн чухал байна.

References

[1] Altangerel, C., Adiyatseren, B.: Two level rules for mongolian language. In: Proceedings of the the 7th Multimedia and Information Technology Application (MITA2011). pp. 130-133 (2011) , Ulaanbaatar, Mongolia

[2] Çöltekin, c.: A freely available morphological analyzer for Turkish. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010). pp. 820-827 (2010), http://www.lrec-conf.org/proceedings/lrec2010/summaries/109.html

[3] Мөнх-Учрал, Э., Чоймаа, Ш.: Хөрвүүлэх програмд зориулсан монгол хэлний судалгаа. Докторын зэрэг горилсон бүтээл. Улаанбаатар, Монгол (2010)

[4] Purev, J., Odbayar, C.: Part of speech tagging for mongolian corpus. In: The 7th Workshop on Asian Language Resources. Singapore (2009)

[5] Purev, J., Tsolmon, Z., Altangerel, C., Cheolyoung, O.: Pc-kimmo-based description of mongolian morphology. vol. 1, pp. 41-48 (2005), http://www.jips-k.org/dlibrary/JIPS_v01_no1_paper8.pdf

[6] Washington, J., Ipasov, M., Tyers, F.: A Finite-state morphological transducer for kyrgyz. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey (may 2012)

[7] Zoljargal, M., Purev, J.: Mongolian trigram part of speech tagger. In: Proceedings of the the 7th Multimedia and Information Technology Application (MITA2011). pp. 161-163 (2011)

[8] Kimmo K., Two-level morphology: A general computational model for word-form recognition and generation. Ph.D. thesis, University of Helsinki (1983)

[9] Дамдинсүрэн Ц., Осор Б., Монгол үсгийн дүрмийн толь, БНМАУ, Ардын Боловсролын Яамны Сурах Бичиг-Сэтгүүлийн Нэгдсэн Редакцын Газар, (1983)

[10] Badam-Osor .Kh, Atsushi .F, A lemmatization Method for Modern Mongolian and its Application to indexing for Information Retrieval, Information Processing and Management: an International Journal, volume 45 issue 4, pp 438-451, (2009)