Монгол Хэлний Нэрлэсэн Нэгжийн Хөмрөг Байгуулах нь

Мөнхжаргалын Золжаргал
Компьютер Хэл Шинжлэлийн Судалгааны Төв
Монгол Улсын Их Сургууль
Улаанбаатар, Монгол улс
zoljargal@num.edu.mn

Нямдаваагийн Оюундарь
Компьютер Хэл Шинжлэлийн Судалгааны Төв
Монгол Улсын Их Сургууль
Улаанбаатар, Монгол улс
oyundari.n@gmail.com

Чагнаагийн Алтангэрэл
Компьютер Хэл Шинжлэлийн Судалгааны Төв
Монгол Улсын Их Сургууль
Улаанбаатар, Монгол улс
altangerel@num.edu.mn

Дамбасүрэнгийн Нанзадрагчаа
Компьютер Хэл Шинжлэлийн Судалгааны Төв
Монгол Улсын Их Сургууль
Улаанбаатар, Монгол улс
nanzadragchaa@num.edu.mn

***Хураангуй*—Нэрлэсэн нэгж таниур нь өгүүлбэрт байгаа хүн, байгууллага, орон байрын нэрийг автоматаар илрүүлэх ба мэдээлэл оновчтой хайхад чухал ач холбогдолтой хэрэгсэл юм. Монгол хэлний нэрлэсэн нэгж таниур хөгжүүлэх, бусад хэлэнд ашигладаг арга техникийг харьцуулан судлахад нэн тэргүүнд судалгааны материалын буюу нэрлэсэн нэгжийн хөмрөг шаардлагатай байна. Өгүүлэл нь монгол хэлний нэрлэсэн нэгжийн хөмрөгийг байгуулах, түүний үр дүнд тулгуурлан машин сургалтын аргыг туршиж, машины гаргаж байгаа алдааг хэлзүйн үүднээс судалсан тухай өгүүлнэ.**

Түлхүүр үг—Нэрлэсэн Нэгж (Named Entity), Нэрлэсэн Нэгж Таниур (Named Entity Recognizer), хөмрөг (corpus).

I. Удиртгал

Технологийн дэвшлийг ашиглан төрөл бүрийн мэдээлэл асар их хэмжээгээр бий болохын зэрэгцээ цаг хугацаа орон зайн хязгааргүй асар хурдтай тархаж байна. Мэдээлэл их хэмжээгээр бий болоход түлхэц өгч байгаа гол хүчин зүйл нь машины (компьютер, гар утас, таблет г.м. дээр суурилсан хүчирхэг програм хангамжууд) тусламжтай хүн бүр мэдээлэл бүтээх, түгээх боломжтой болсон явдал юм. Нөгөө талаас мэдээллийг хүлээн авч байгаа хэрэглэгч их хэмжээний мэдээлэл дотроос өөрт хэрэгтэйг ялган авах, оновчтой хайлт хийх зэрэгт мөн л машины туслалцаа шаардаж байна.

Оновчтой мэдээлэл ялган авах (Information Retrieval) ажлын суурь даалгавар нь Нэрлэсэн Нэгж Таних (ННТ) (Named Entity Recognition) юм. ННТ-ын даалгавар нь өгүүлбэрт байгаа хүн, байгууллага, газрын нэрийг ялган таних юм [3]. Тэгэхдээ системийн зорилгоос хамаарч эдгээр нэр нь янз бүр бөгөөд мэдээний салбарт дээрх гурван төрлийн нэр дээр цаг хугацаа (огноо, цаг), тоон илэрхийлэл (мөнгө, хувь хэмжээ) [4] гэсэн нэгжүүдийг авч үздэг.

Сүүлийн арван жилд дэлхийн олон судалгааны баг нэрлэсэн нэгж таних олон арга дэвшүүлсэн [6] ба ерөнхийд нь 1) толинд суурилсан арга, 2) дүрэмд суурилсан арга, 3) машин сургалтын (Machine Learning) арга гэж гурав ангилдаг. Толинд суурилсан арга нь нэрлэсэн нэгжийн жагсаалт ашиглан өгүүлбэр дотроос хайлт хийдэг ба ертөнц дээр байгаа бүх нэрийг агуулах шаардлагатай учир хөгжүүлэхэд хүндрэлтэй. Харин дүрэмд суурилсан арга нь хэлзүйн дүрэм, үгийн хэлбэржилт зэрэгт үндэслэн гараар яв цав дүрэм бичиж ашигладаг. Машин сургалтын арга нь өмнө нь тэмдэглэсэн хөмрөг дээр статистик шинжилгээ хийж магадлалын аргаар тааварлах зарчмаар тэмдэглэгээ хийдэг бөгөөд хамгийн их үр дүнтэй аргад тооцогдож байна [4,5,6,8].

Хэлзүйн үүднээс зөв бичих дүрэм (үг хувирах, том үсгээр бичих, хураах г.м.), хамт хэрэглэгддэг үг, үгийн олон хувилбар нь машинаар нэрлэсэн нэгж таних, тэдгээрийг өөр хоорондынь ялгах гол шижим болдог.

Монгол хэлэнд ННТ-ын систем хөгжүүлэх, олон улсад хэрэглэж байгаа аргуудыг турших, цаашлаад зөвхөн монгол хэлэнд байдаг онцлогт тохирсон шинэ арга хөгжүүлэх, тэдгээрийг батагтах зэрэгт ашиглах гол судалгааны материал нь нэрлэсэн нэгжийн хөмрөг юм. Түүнчлэн хэл шинжээчдэд хүн, байгууллага, орон байрын нэрийн хэрэглээ, илэрхийлэх ёс, утга, агуулга, бүтцийн онцлогийг тодорхойлох зэрэг олон сонирхолтой хэлзүйн судалгаа хийх материалын сан болно гэдэгт найдаж байна.

Өгүүллийн зорилго нь монгол хэлэнд анх удаа нэрлэсэн нэгжийн хөмрөг үүсгэх юм. Ажлын үр дүнд тулгуурлан нэрлэсэн нэгж таниур програмын эхний хувилбарыг хөгжүүлсэн бөгөөд машинаар танихад Амьд Хэл Боловсруулалтын (Natural Language Processing) үүднээс ямар асуудал үүсч байгаа, хэлзүйн үүднээс ямар онцлогтой, хүн, газар орон, байгууллагын нэр зэрэг оноосон нэрийг бодит

байдалд хэрхэн бичдэг, зөв бичгийн дүрэм хэр идэвхитэй үйлчилж байгаад дүн шинжилгээ хийхийг зорилоо.

II. Нэрлэсэн Нэгж

Хүн, байгууллага, орон байрыг заасан хэлцийг нэрлэсэн нэгж гэж үзэх бөгөөд тухайн нэгжийг өгүүлбэрээс салгаж дангаар нь тавихад өөрийн зааж буй цор ганц юмыг бүрэн илэрхийлж чаддаг байх ёстой. Тиймээс эдгээр гурван ангиллыг илэрхийлсэн оноосон нэрийг[1] нэрлэсэн нэгж гэж үзнэ.

A. Монгол хэлний оноосон нэр: Амьд Хэл Боловсруулалтын үүднээс

Аливаа оноосон нэр нь хүн төрөлхтөний хөгжлийн нэлээд боловсорсон шатанд буй болсон ба адил төст зүйлийг бие биеэс нь ялгах, тусгайлан нэрлэх шаардлагаас үүссэн бөгөөд ерийн нэрийг оноосон нэр болгодог учраас оноосон нэрэнд үгийн сангийн утга байхгүй, ерийн нэрийн хэлбэр дуудлагыг нь ашиглаж, нэг төрлийн олон юмсаас ялгасан дохионы үүрэгтэй юм [1].

ННТ сайн хөгжсөн англи, герман зэрэг хэлнүүдээс ялгарах монгол хэлний оноосон нэрийн онцлог нь тоо, тийн ялгал, хамаатуулах нөхцлөөр хувилан өгүүлбэрт ихэвчлэн өгүүлэгдэхүүн, тусгадахууны үүрэг гүйцэтгэдэг [1]. Үүнийг машинаар танихад үүсдэг гол асуудал нь нэг оноосон нэр нөхцлөөр хэлбэржиж анх байсан хэлбэрээсээ өөр хэлбэртэй болдог. Тухайлбал машинд “Бат” гэсэн нэрийг оноосон нэр гэж сургавал “Батын”, “Батынхаас”, “Баттайгаа” зэрэг бусад хувилсан хэлбэрийг таньж чадахгүй. Учир нь машин эдгээрийг ялгаатай тэмдэгтийн цуваа гэж үздэг. Шийдэж болох арга нь эдгээр үгийг автоматаар үгзүйн задлал хийж үгийн үндсийг олоод тэмдэглэж болдог. Харамсалтай нь монгол хэлний хувьд үгзүйн задлуур нь бүрэн хийгдээгүй байна.

Хэрэв оноосон нэр олон тооны нөхцөл авбал ерийн нэр болдог [1]. Олон тооны нөхцөлтэй оноосон нэр болон *–ынхан, -ийнхэн* нөхцөл авч олон хүнийг заан нэрлэх утгатай болсон газар орон, байгууллагын нэрийг нэрлэсэн нэгж гэж үзэхгүй. Учир нь оноосон нэр жинхэнэ үүргээрээ байхдаа ямагт ганц тоо заадаг бөгөөд утгын хувьд мөн ганц утга илэрхийлдэг. Олон тооны дагавар авбал “..нэг төрлийн олон юмаас ялгасан дохионы үүрэг” алдагдана. Жишээ нь: *Япон-Япончууд, Увс-Увсынхан* гэх мэт.

Түүнчлэн оноосон нэрийг ерийн нэр болгож болдог. Жишээ нь *монгол хэл, хятад торго*. Ингэж оноосон нэр ерийн нэр болох, ерийн нэрээр оноосон нэрийг нэрлэдэг нь машин хооронд нь зөв ялган танихад төвөгтэй асуудал үүсгэдэг. Үүний тулд машин бүх ертөнцийн талаар мэдлэгтэй байхыг шаарддаг.

[1] Нэг төрлийн олон юмны дотроос аль нэгийг нь онцлон ялгахын тулд тусгайлан оногдуулж өгсөн жинхэнэ нэрийг оноосон нэр гэнэ. (ОЦМХ 2008 хуу. 130)

Бичих хэлбэрийн хувьд, Ц.Дамдинсүрэн, Б.Осор нарын “Монгол Үсгийн Дүрмийн Толь”-ийн (МҮДТ) [2] дүрэм 47.2-т “хүний нэр, амьтан, хот суурин, уул ус зэрэг ертөнц дээр ижилгүй юмын оноосон нэрийг том үсгээр эхэлж бичнэ”, 47.3-т “Дэлхий дахины албан ба олон нийтийн байгууллагын нэрийг хэдэн үгээр бүтсэн бол цөмийг том үсгээр эхэлж бичнэ”, 47.4-т “Улс гүрэн ба улсын төвийн хороо, зөвлөл, яам, холбоо зэрэг байгууллагын нэр хэдэн үгээр бүтсэн бол цөмийг том үсгээр эхэлж бичнэ” гэж заажээ. Гэтэл 47-р дүрэмд тайлбар хийхдээ “Дээрхээс бусад улсын төвийн захиргаанд харьяалагдах албан үйлдвэр, аймаг, сум дүүрэг, соёл шинжлэх ухаан зэрэг газрын нэрийг хэдэн үгээр бүтсэн бол зөвхөн эхний үгийг том үсгээр эхэлж бичнэ.” гэж заасан байдаг. Машины зүгээс үзвэл, оноосон нэрийн эхний үгийг том үсгээр эхэлж бичээд бусдыг нь жижгээр бичсэн тохиолдлыг өгүүлбэрийн бусад ерийн нэрээс ялгах нь төвөгтэй асуудал болж хувирдаг. Жишээ нь: *Зам тээвэр, барилга, хот байгуулалтын яам, Гашуун сухайтын боомт*.

Мөн хамт бичиж заншсан хоёр үгээс бүтсэн оноосон нэрийн хоёрдугаар үг нь эгшгээр эхэлсэн байвал хооронд нь холбоос тавьж хоёр дахь үгийг томоор эхэлж бичнэ. Жишээлбэл: *Баян-Өлгий, Буян-Өлзий, Баруун-Урт.*

Оноосон нэр дотроос хүний нэр харьцангуй зүй тогтолтой бичигддэг бөгөөд овог, нэрийг харьяалахын тийн ялгалаар холбож бичдэг. Жишээ нь: *Цахиагийн Элбэгдорж.* Мөн овгийг хураасан үсгийн ард цэг тавин бичнэ. Жишээ нь: *Д.Нацагдорж, Батж. Батбаяр.*

Харин монгол хүний нэр хоёр болон түүнээс дээш өөр нэрээс бүтэж болдог онцлог нь дурын байдлаар хоршин шинэ нэр үүсгэж болдог. Жишээ нь: *Дамдинсүрэн, Сүрэн* гэх мэт. Ингэж шинэ нэр үүсэх нь мөн л машинд өмнө нь сурсан нэрэнд ороогүй тохиолдол үүсч болно.

МҮДТ-ийн дүрэм 58-д хашилт хэрэглэх тухай “Ном зохиол, сонин сэтгүүл, онгоц хөлөг, үйлдвэр нэгдэл, албан байгууллага зэргийн болзож оноосон нэрийг их төлөв хашилтад хийнэ”. Гэхдээ “Дээр дурдсан зүйлийн ерийн нэрийг хашилтгүй бичнэ” гэж заасан байдаг. Энэ дүрэм нь байгууллагын нэрийг танихад машинд дөхөм болгож байна.

Монгол хэлний нэг онцлог нь өгүүлбэрийн дараалал SOV (subject object verb) хэлбэртэй байдаг учир дундаа ямар нэгэн тусгаарласан тэмдэггүйгээр (цэг, таслал) хоёр оноосон нэр дараалан орох боломжтой юм. Жишээ нь “*Ерөнхийлөгч Элбэгдоржийн Булган аймагт хийсэн айлчлал өнөөдөр өндөрлөлөө.*” гэсэн өгүүлбэрт “*Элбэгдоржийн”* болон “*Булган”* гэдэг хоёр оноосон нэр тус тусдаа тэмдэглэгдэх ёстой байтал үүнийг машин “*Элбэгдоржийн Булган*” нь хамтдаа хүний нэр гэж тэмдэглэх магадлал өндөртэй байгаа юм. Учир нь энэ хэлбэр хүний бүтэн нэрийг бичих дүрэмтэй ижил

бөгөөд “*Элбэгдорж*”, “*Булган*” гэдэг хоёр нэр нь мөн хүний нэрээр түгээмэл хэрэглэгддэг.

Гадаад хэлний оноосон нэрийг орчуулалгүй тэр хэвээр нь хэрэглэдэг ба МҮДТ-ийн 49.2-т “Ойрмог гадаад хэлнээс аваад гадаад үг гэдэг нь нийтэд мэдэгдэж байгаа үгийг бичихэд уг хэлний дуудлага ба бичгийн дүрсийг харгалзана.” гэсэн дүрэм байдаг боловч нэг мөр болгон тэмдэглэсэн, журамласан зүйл одоогоор байхгүйгээс тухайн бичээч өөрийн дуудлагаар бичих, эсвэл цаад хэлний хэлбэрийг кирилл үсгээр галигчлан буулгах, бүр зарим тохиолдолд тухайн хэлний бичгээр нь бичиж байна. Жишээ нь: *Джон Нэш, Жон Нэйш, Барак Обама* (Barack Obama), *Times сэтгүүл, Таймс сэтгүүл*.

B. Олон утгат нэрлэсэн нэгж

Аливаа юмыг оноож нэрлэхдээ ихэвчлэн ерийн нэрийг ашиглах тул ерийн ба оноосон нэрийн утгатай ижил нэр хэлэнд элбэг байдаг [1]. Жишээлбэл: *туяа-Туяа, төмөр-Төмөр, сүх-Сүх.* Түүнчлэн нэг ерийн нэр нь өөр өөр төрлийн зүйлийг оноон заах нь тун элбэг. Тухайлбал: *Булган* гэдэг аймгийн нэр, голын нэр, сумын нэр, хүний нэр, байгууллагын нэр байдаг. Энэ нь өгүүлбэр, эхийн утгаас хамаарна.

Зарим оноосон нэр үгсийн аймгийн олон шинжийг хадгалсан байна. Жишээ нь *Дэлгэр* гэвэл үйл үг, жинхэнэ нэр, тэмдэг нэрийн аль алины шинжтэй [1].

Дээрх оноосон нэрийн олон зүйл заах байдал нь машинаар ялган танихад маш төвөгтэй асуудал үүсгэдэг. Учир нь машин утгазүйн түвшинд хүний хэлийг гүнзгий ойлгож, тэдгээрийг ялган таних асуудал нэг мөр шийдэгдээгүй байна. ННТ-ын ихэнх арга нь эдгээр салаа утгатай үгийн хамт хэрэглэгддэг үг, эхийн төрөл зэргээс хамааруулж ялгадаг [8].

C. Нэрлэсэн нэгж тэмдэглэх мөрдлөг

Оноосон нэрийг гараар тэмдэглэхдээ өмнөх хэсэгт дурьдсан монгол хэлний зөв бичих дүрэм болон үгийн сангийн үүднээс дараах мөрдлөгийг баримтлав.

1) Ерөнхий мөрдлөг:

- Оноосон нэр олон тоо эсвэл харьяалахын тийн ялгал+х/хан[4] хэлбэртэй байвал тухайн ялгаж нэрлэсэн зүйлийг бус түүнд хамаарах хүмүүсийг төлөөлөх учраас тэмдэглэхгүй. Жишээ нь: *Монголчууд, Мобикомынхон, УМХГ-ынх (Улсын Мэргэжлийн Хяналтын Газарынх)* гэх мэт.
- Товчилсон оноосон нэрийн араас нөхцөл орсон байвал нөхцлийг хамтад нь тэмдэглэхгүй. Жишээ нь: *АНУ-ын*, *АТГ-т* гэх мэт.
- Газар орон, байгууллага, хүний оноосон нэр өөр зүйлийн оноосон нэрийн дотор тохиолдвол тэмдэглэхгүй. Жишээ нь: “Монгол орон-Монгол тэмүүлэл” аян, “*Зөвхөн Монголдоо*” дуу, *Сүхбаатарын одон* гэх мэт.

2) Хүний нэр тэмдэглэх мөрдлөг:

- Хэн? гэсэн асуултад хариулагдах ганц хүнийг ялган нэрлэсэн утгатай оноосон нэрийг тэмдэглэнэ.
- Харьяалахын тийн ялгалаар холбогдсон овог нэрийг хамтад нь нэг нэр гэж үзнэ. Жишээ нь: *Цэрэндашийн Дамиран*
- Хоч нэр, бүтэн нэрийг цээрлэн дуудах нэр, олонд танигдсан нэр тухайн хүнийг төлөөлж байвал тэмдэглэнэ. Жишээ нь: *Хатанбаатар Магсаржав, Ноён хутагт Данзанравжаа, Ноён хутагт, Дижи Өвжий, Ри багш, Асашёру* (Сумогийн аварга Д.Дагвадоржийн олонд танигдсан нэр)

3) Орон байрын нэр тэмдэглэх мөрдлөг:

- Хаана? гэсэн асуултад хариулагдах улс гүрэн, хот суурин, газар орон, уул усны оноосон нэрийг тэмдэглэнэ. Жишээ нь: *Монгол улс, Сэлэнгэ аймаг, Туул гол* гэх мэт.
- Орон байрын оноосон нэрийг тэмдэглэхдээ ерийн нэрийг хамт тэмдэглэхгүй. Жишээ нь: *Сэлэнгэ аймаг, Туул гол* гэдгээс *Сэлэнгэ, Туул* гэсэн олон аймаг, гол дотроос ялган нэрлэсэн хэсгийг тэмдэглэнэ.
- Улс орны нэр тодотголын үүргээр орж байвал тэмдэглэхгүй. Жишээ нь: *польш цэрэг, америк бараа.*

4) Байгууллагын нэр тэмдэглэх мөрдлөг:

- Дэлхий дахины болон олон улсын байгууллага, пүүс, компани, нам, эвсэл, хөдөлгөөн, хамтлаг, спортын баг, улсын төв хороо, зөвлөл, яам, холбоодын нэрийг тэмдэглэнэ.
- Дээр дурдсан байгууллагуудын оноосон нэрийг тэмдэглэхдээ ерийн нэрийг хамтад нь тэмдэглэхгүй. Жишээ нь: *Хадгаламж банк, “Иргэний зориг” нам* гэдгээс банк, нам зэргийг тэмдэглэхгүй.
- Байгууллагын төрөл, зорилгыг заах ерөнхий нэр нь тухайн оноосон нэрийнхээ бүрэлдэхүүнд багтаж байвал хамтад нь тэмдэглэнэ. Жишээ нь: *Монгол Ардын Хувьсгалт Нам, Монголын Өмгөөлөгчдийн холбоо*. Учир нь эдгээр ерийн нэргүйгээр *Монгол Ардын Хувьсгалт, Монголын Өмгөөлөгчдийн* зэрэг үгс нь тухайн байгууллагын оноосон нэрийг төлөөлж чадахгүй.

5) Цаг хугацаа тэмдэглэх мөрдлөг:

- Хэзээ? гэсэн асуултад хариулагдах, баримтын огноотой харьцуулаад цаглабарт

тэмдэглэж болохуйц тодорхой, харьцангуй болон үнэмлэхүй цаг заасан хэлцийг тэмдэглэнэ. Жишээ нь: *өглөөний 9 цаг 15 минут, өчигдөр, 2014 он, бямба гараг,*

- Нэг тодорхой цагийг зааж байвал хэдэн үгээс ч бүтсэн байсан нэг нэгж гэж тэмдэглэнэ. Жишээ нь: *1945 оны 10 дугаар сарын 24-ий өглөө 06 цагт*
- Аливаа үйл явдлын эхлэл төгсгөл нь тодорхойгүй үргэлжлэх хугацааг тэмдэглэхгүй. Жишээ нь: *10 хоног, тавар сарын турш* гэх мэт.

6) Тоон илэрхийлэл тэмдэглэх мөрдлөг:

- Мөнгө (MONEY), хувь (PERCENT)-ийн тоо хэмжээг илэрхийлсэн тооны нэрийг тэмдэглэнэ. Жишээ нь: *5 хувь, 5%, 10 сая төгрөг, 1 юань*
- Аливаа мөнгөн дэвсгэртийн нэрийг хамтад нь тэмдэглэнэ. Жишээ нь: *50 төгрөг, 500 америк доллар*

III. Хөмрөг

A. Хөмрөгийн шинж

Анх MUC07 [4], CoNLL02 [6], CoNLL03 [5] олон улсын хурлаар англи, герман, франц, испани хэлэнд зориулсан нээлттэй эх бүхий хөмрөгийг нийтэд ил тавьж, эдгээр дээр түшиглэн дэлхийн олон судалгааны төв олон аргыг боловсруулсан байдаг. Тиймээс бид эдгээр хөмрөгтэй мэдээний салбар, үгийн тоо, нэрлэсэн нэгжийн тоогоороо ижил хөмрөг байгуулахыг зорьсон бөгөөд ингэснээр өмнө нь туршигдсан аргуудыг гадаад хэлний үр дүнтэй харьцуулан судлах боломжийг олгох юм.

Хүснэгт I. Гараар тэмдэглэсэн хөмрөгийн шинж

	Үг	Хэлц
Нийт үг	277,188	-
Хүний нэр	8,316	4,574
Байгууллага	6,689	3,460
Орон байр	5,543	4,977
Огноо	3,271	1,240
Цаг	532	156
Мөнгө	1,969	641
Хувь	434	185

Бид эхний ээлжинд Компьютер Хэл Шинжлэлийн Судалгааны Төвийн (КХШСТ) байгуулсан монгол хэлний үгийн аймгийн тэмдэглэгээт хөмрөгийн [3] мэдээ, мэдээллийн хэсэгт нэрлэсэн нэгжийн тэмдэглэгээг хадаж өгсөн. Тэмдэглэгээ хийхийн өмнө MUC07 болон CoNLL03 хурлын тэмдэглэгээний стандартыг монгол хэлний хэлзүйн онцлогт тохируулан өөрчилж монгол хэлний нэрлэсэн нэгж тэмдэглэх мөрдлөгийг үүсгэсэн (мөрдлөгийн хураангуй *II.C*-д бий).

B. Хөмрөг тэмдэглэх ажлын дараалал

Бид хөмрөгийг гараар тэмдэглэгээ хийх зориулалттай нээлттэй эх бүхий "Brat[2]" хэрэгслийг ашиглан хийсэн. Хөмрөгийн чанар нь хөгжүүлж байгаа арга, технологид шууд нөлөөлөх учир өндөр нарийвчлалтай тэмдэглэх шаардлагатай. Үүний тулд бид тэмдэглэх явцыг гурван алхамд хуваасан: 1) хэл шинжээч ННТ мөрдлөг ашиглан гараар тэмдэглэгээ хийнэ, 2) хэл шинжээчийн тэмдэглэсэн ажлыг шалгагч хянана, 3) хэл шинжээч болон шалгагч хоёрын санал зөрсөн тэмдэглэгээг хэлэлцээд дахин тэмдэглэнэ.

Гараар 277,188 үг бүхий хөмрөгийг тэмдэглэсний дараа уг хөмрөгөө ашиглан *Maximum Entropy* [9], *Conditional Random Field* [7] машин сургалтын аргуудыг сургаж туршилт хийсэн. Уг сургасан автомат тэмдэглүүрээр мэдээний вэб сайтын 2007, 2008, 2009, 2010 оны мэдээнээс түүвэрлэж тэмдэглэсэн. Автомат тэмдэглэгээний дараа хэл шинжээч хянагч хоёр машины тэмдэглэснийг гараар засаж шинэ хөмрөг үүсгэх зарчмаар ажилласан.

IV. Хэлэлцүүлэг

Машин сургалтын аргаар сургасан ННТ програм хүн, байгууллага, газар орны оноосон нэрийг 73%-ийн гүйцэтгэлтэй таньсан бөгөөд таниагүй буюу буруу тэмдэглэсэн 27%-ийг гараар засахад: (1) тэмдэглээгүй нэрийг нэмэх, (2) буруу тэмдэглэснийг засах, (3) илүү тэмдэглэснийг хасах засвар хийсэн.

Хүн, байгууллага, байрлалын оноосон нэр бидний сонгон туршсан хөмрөгийн (gogo.mn сайтын 2010 оны 5 сарын мэдээнээс 78,910 үг бүхий хэсгийг түүвэрлэн авсан) 6,5%-ийг хамарч, түүнээс хүний нэр 2,3%, байрлалын нэр 2,5%, байгууллагын нэр 1,6%-ийг эзэлж байна.

Машин хүний нэрийг хамгийн их буюу 85%, байгууллагын нэрийг хамгийн бага буюу 54% таньж байна. Учир нь монгол хүний нэрийг бичих харьцангуй тогтсон дүрэмтэй байхад байгууллагын нэрийг бичих дүрэм тодорхойгүй бөгөөд кирилл бичигт одоо идэвхтэй хэрэглэгдэж байгаа МҮДТ-ийн том үсгээр бичих, үг хурааж бичих (дүрэм 48), хашилт хэрэглэх (дүрэм 58) дүрмийг төдийлөн баримталдаггүй нь байгууллагын нэрийг дараах хэд хэдэн янзаар бичсэнээс харагдаж байна.

- Томоор бичих: 1) Бүх үгийг томоор: *Монгол Улсын Их Сургууль,* 2) Эхний үсгийг томоор: *Монгол улсын их сургууль.*
- Хашилт хэрэглэх: 1) Хашилттай: *"Зоос" банк,* 2) Хашилтгүй: *Зоос банк,* 3) Ерийн

[2] http://brat.nlplab.org/

нэрийг хамтад нь хашилтад бичих: *“Зоос банк”*

- Товчилж бичих: 1) Бүрэн товчлох: *БСШУЯ,* 2) Хагас товчилж: *БСШУ яамны*, 3) Товчлоогүй: *Боловсрол, соёл, шинжлэх ухааны яам.*
- Гадаад үсгээр бичих: “Cass таун” хотхон, “Hero” энтертайнмент

Байгууллагын оноосон нэрийн 46%-ийг таниагүй буюу буруу тэмдэглэсэн бөгөөд тэдгээрийг шалтгаанаар нь ангилбал 86%-д нь огт тэмдэглээгүй байна.

Хүснэгт II-т газар орны нэрийн таниагүй буюу буруу таньсан 27%-ийг дотор нь дахин ангилсныг харууллаа. Газар орны нэрийг тэмдэглэхэд гарсан алдааны хамгийн их буюу 56% нь оноосон нэрийг таньж тэмдэглээгүй алдаа байна. Энэ нь дараах шалтгаантай байна.

- Дараалан орсон олон газар усны нэрийг таслалаар зааглан бичсэнээс олон газар усны нэрийг нэг бүрчлэн тэмдэглээгүй. Жишээ нь: *Адуунчулуун, Чандгана, Талбулаг*.
- Хашилттай өгүүлбэрийн эхэнд орсон газар усны нэрийг танихгүй байх. Жишээ нь: *“Өмнөговь.., “Хятад...*

ХҮСНЭГТ II. ГАЗАР ОРНЫ НЭРИЙГ БУРУУ ТЭМДЭГЛЭСЭН БАЙДАЛ

Алдааны төрөл	**Хувь(%)**
Өмнөх үгтэй нь хамт	16
Хүний нэр	13
Хүний нэрийн хэсэг	10
Байгууллага	4
Байгууллагын хэсэг	1
Тэмдэглээгүй	56

Газар орны нэрийг танихад гарсан алдааны 13%-ийг хүний нэр гэж тэмдэглэх эзэлж байгаа нь дараах шалтгаантай байна.

- Харьяалахын тийн ялгалаар холбогдсон хүний овог нэртэй ижил хэлбэртэй хоёр өөр байрлалын оноосон нэрийг нэг хүний нэр гэж тэмдэглэх. Жишээ нь: *Дорнодын Баянтүмэн, Суматрагийн Пеканбару,*
- Хоёр үгээс бүтсэн оноосон нэрийн хоёрдугаар үг нь эгшгээр эхэлсэн байвал хооронд нь холбоос тавьж, хоёр дахь үгийг томоор эхэлж бичих дүрмээр бичсэн газрын нэрийг хүний нэр гэж тэмдэглэх. Жишээ нь: *Баруун-Урт, Замын-Үүд* гэх мэт.

Хүний нэр танихад гарсан алдааны хамгийн их буюу 50%-ийг тэмдэглээгүй шалтгаанууд нь:

- Өгүүлбэрийн эхэнд орсон, овог буюу овгийн хураасан үсэггүй нэг үгээс бүтсэн хүний нэрийг үл таних. Жишээ нь: *Чингис хаан харь аймгуудыг дайлах..., Мянгаа захирлын...* гэх мэт.
- Хүний нэрийн хамгийн түгээмэл бүтэц нь овгийг хураасан үсгийн ард цэг тавих (*Ц.Элбэгдорж*), овог нэрийг харьяалахын тийн ялгалаар холбох (*Цахиагийн Элбэгдорж*) юм. Энэ хоёр тохиолдол машин алдалгүй тэмдэглэх боловч монгол хүний нэрийг овоггүй дан бичих, дуудах нэр буюу олонд танигдсан нэр, нэг үгээр бүтсэн нэр зэрэг тохиолдолд танихгүй байх нь бий. Жишээ нь: *Ганаа, Мээдээ, дуучин ВХ.*
- Гадаад хүний нэрийг бичих нь монгол хүний нэр бичихээс их ялгаатай. Тиймээс гадаад хүний нэрийг танихгүй байх тохиолдол байна. Жишээ нь: *Ким Ю На, Зой Салдана* гэх мэт. Гэвч зарим тохиолдолд машин хам сэдвээс нь таньж байгаа нь ажиглагдаж байна. Сайд, дарга, ерөнхийлөгч зэрэг үгсийн дараа орсон том үсгээр эхэлсэн үгийг хүний нэр гэж тэмдэглэсэн тохиолдол цөөнгүй байна. Жишээ нь: “*Бодлого судалгааны хүрээлэнгийн Ван Хулин, төрийн зөвлөлийн гишүүн Дай Бингуо*” зэргийн оноосон нэрийг тэмдэглээгүй бол “*гадаад хэргийн сайд Ян Жэчи, худалдааны сайд Чэн Дэмин*” зэрэг тохиолдолд хүний нэр гэж тэмдэглэсэн байна.

V. ДҮГНЭЛТ

Судалгааны ажлаар монгол хэлний нэрлэсэн нэгжийн хөмрөгийг байгуулсан бөгөөд КХШСТ-ийн 5 сая үгтэй үгийн аймгийн тэмдэглэгээт хөмрөгийн мэдээ, мэдээллийн 277,188 үг бүхий хэсгийг гараар тэмдэглэсэн. Үүнд хүний нэр 4,574, байгууллагын нэр 3,460, орон байрын нэр 4,977, огноо 1,240, цаг 156, мөнгө 641, хувь 185 тэмдэглэсэн байна. Хөмрөгийг чанартай үүсгэхийн тулд монгол хэлний оноосон нэрийн онцлог, зөв бичихзүйн дүрэм зэргийг сайтар судалж, монгол хэлний хэлзүйн онцлогт тохируулан нэрлэсэн нэгж тэмдэглэх мөрдлөгийг үүсгэв.

Гараар тэмдэглэсэн хөмрөгт тулгуурлаж статистик машин сургалтын аргаар монгол хэлний нэрлэсэн нэгж таниур програмын эхний хувилбарыг туршсан бөгөөд хагас автоматаар 78,910 үгтэй мэдээний эхийг тэмдэглэлээ. Хагас автоматаар тэмдэглэх явцад орчин цагийн монгол хэлний бичгийн хэлний зөв бичгийн нийтлэг алдаа, хэлзүйн онцлог машин сургалтын аргад хэрхэн нөлөөлж буйг судалсан бөгөөд энэ нь цаашдын ННТ ажилд чухал ач холбогдолтой юм. ННТ-ын хөмрөг үүсгэх, түүн дээр тулгуурлан судалгаа хийх ажил нь монгол хэлэнд анх удаа хийгдэж байна. Энэ ажил нь цаашид ННТ олон аргыг турших, харьцуулах суурь судалгаа, материал болно гэдэгт бид найдаж байна.

ЗААЛТ

[1] Д.Төмөртогоо нар, “Орчин цагийн монгол хэл” Улаанбаатар, 2008.

[2] Ц.Дамдинсүрэн, Б.Осор “Монгол үсгийн дүрмийн толь” Улаанбаатар, 1983.

[3] Purev Jaimai, Odbayar Chimeddorj, “Part of Speech Tagging for Mongolian Corpus”, 7th Workshop on Asian Language Resources, Proceeding of the Workshop, Singapore, 2009.

[4] N. Chinchor, P. Robinson, “Named Entity Task Definition”, Message Understanding Conference 1997.

[5] Erik F. Tjong Kim Sang and Fien De Meulder, “Introduction to the conll-2003 shared task: Language-independent named entity recognition”, In Walter Daelemans and Miles Osborne, editors, Proceedings of CoNLL-2003, pages 142–147. Edmonton, Canada, 2003.

[6] David Nadeau and Satoshi Sekine, “A survey of named entity recognition and classification”, LinguisticaeInvestigationes, 30(1):3–26, January 2007, Publisher: John Benjamins Publishing Company.

[7] Jenny Rose Finkel, Trond Grenager, and Christopher Manning, “Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling”, Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370.

[8] Eszter Simon and Andres Kornai, “Approaches to Hungarian Named Entity Recognition”, Thesis of PhD, Budapest University of Technology and Economics, Budapest, 2013.

[9] Silviu Cucerzan and David Yarowsky, “Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence”, Proceedings of the 1999 Joint SIGDAT.

[10] G Szarvas, R Farkas, L Felfoldi, A Kocsor, J Csirik, “A highly accurate Named Entity corpus for Hungarian”, Proceedings of International Conference on Language Resources and Evaluation, 2006.