Нутгийн Мэдлэгийн Цөмийг хамтын ажиллагаат олны хүчээр үүсгэх систем

Зундуйн Цолмон[1], Бямбадоржийн Эрдэнэбилэг[2], Ганболдын Амарсанаа[3]
Мэдээлэл, компьютерийн ухааны тэнхим
ХШУИС, Монгол Улсын Их Сургууль
Улаанбаатар, Монгол улс
Цахим шуудан: {tsolmonz[1], amarsanaag[3]}@num.edu.mn, erdenebileg.byambadorj@gmail.com[2]

***Хураангуй*—Энэ ажлаар үгийн утгазүйн цахим сан - Нутгийн Мэдлэгийн Цөмийг (НМЦ) (Local Knowledge Core) хамтын ажиллагаат олны хүчээр үүсгэх аргачлалыг хэрэгжүүлсэн нээлттэй эхийн системийг танилцуулах болно. Энэ аргачлал олон хэлээр илэрхийлсэн утгазүйн уялдаа холбоо бүхий багц ойлголтуудыг ямар нэг хэл рүү нутагшуулахад оролцогчдыг зохион байгуулж тэдний хувь нэмрийг үр дүнтэй нэгтгэж утгазүйн шинэ цахим сан үүсгэх зорилготой. Бид ойлголтыг нутагшуулах гурван-шатлалт хүний оюуны даалгаврыг, мөн оролцогчдод даалгаварыг оновчтой үүсгэж өгөх алгоритмыг зохиож хэрэгжүүлсэн.**

I. УДИРТГАЛ

Семантик вэб [1] технологийг хөгжүүлэхэд өнөөгийн Вэб дэх бүтэцлэгдээгүй (тухайлбал, түүхий бичвэр), хагас бүтэцлэгдсэн болон бүтэцлэгдсэн өгөгдлийг илүү нарийн тодорхойлж үгийн утгазүйн түвшинд мета-өгөгдлөөр баяжуулж машинд ойлгогдох мэдлэг болгон хадгалах шаардлагатай байна. Ингэснээр эх хэлийг ойлгохуй (Natural Language Understanding), мэдээллийг утга агуулгаар хайх, бүтэцлэгдээгүй өгөгдлийг нэгтгэх зэрэг семантик технологийг илүү сайжруулах боломж бүрдэнэ. Ийм мэдлэгийг аливаа хэл соёлоор илэрхийлэх ойлголт (хот, суурин газар г.м), тэдгээрийн хоорондох утгазүйн холбоо хамаарлыг (хот *бол* суурин газар г.м) дүрсэлсэн онтологиор илэрхийлж болдог. Тодруулбал, ертөнц дээр орших нийтлэг ойлголтуудыг хэл соёл, орон нутаг бүрд өөр өөр үгээр – хэлний үгийн сангийн нэгжээр илэрхийлэх юм. Олон хэлээр илэрхийлэх нэгдмэл онтологи үүсгэхийн тулд ойлголтуудыг хэлнээс үл хамаарах логик давхаргад формаль хэлээр, ойлголтуудыг илэрхийлэх үгийн санг хэлний давхаргад дүрсэлдэг. Иймд аливаа ойлголтыг олон хэл соёлын үгийн сангийн нэгжээр илэрхийлж хооронд нь харгацуулах шаардлагатай. Үүнийг онтологи нутагшуулалтын Ontology localization ишлэл аргаар шийдвэрлэх боломжтой.

Ийм төрлийн хэлний нөөцийг үгийн сан зүйч, сэтгэц-хэл шинжээчид хамгийн өндөр чанартай [2] үүсгэж чадна. Гэвч ертөнц дээрх хэдэн зуун мянган ойлголтуудыг аль нэг хэл рүү нутагшуулах нь цаг хугацаа, өртөг их шаардана. Ийм мэргэжлийн ур чадвар, туршлага шаардсан ажлыг хамтын ажиллагаат олны хүчийг (collaborative crowdsourcing) ашиглан гүйцэтгэх судалгаа сүүлийн үед эрчимжиж байна. Тухайлбал, DBPedia[1] онтологийн ойлголтуудыг Япон хэл рүү орчуулах [3], Ертөнцийн Мэдлэгийн Сангийн (EMC) ойлголтуудыг Монгол хэл рүү нутагшуулах [4], BabelNet[2] онтологийн утгазүйн холбоосуудыг үнэлэх видео тоглоом [5], Орос хэлний Wordnet хөгжүүлэх [6] зэрэг судалгаа онтологийг олны хүчээр нутагшуулах оролдлогууд юм. Эдгээр ажлууд ихэвчлэн үгсийг тэмдэглэх, орчуулах зэрэг хэл шинжлэлийн мэдлэг шаардсан даалгавруудыг оролцогчдод өгч гүйцэтгүүлдэг. Мэргэжлийн хүмүүсийн хийж чадах ажлыг мэргэжлийн бус оролцогчдоор гүйцэтгэхийн тулд даалгаврыг хэсэгчлэн хувааж улам хялбар, энгийн болгох; тухайн оролцогчийн сайн гүйцэтгэж чадах даалгаврыг олж оноох зэрэг асуудал хангалттай судлагдаагүй байна.

Энэ ажлын зорилго нь НМЦ-ийн ойлголтуудыг хамтын ажиллагаат олны хүчээр нутагшуулах аргачлалыг, мөн оролцогчдод нутагшуулалтын даалгавар үүсгэж оноох алгоритмыг хэрэгжүүлсэн системийг хөгжүүлэх юм. Аргачлал нь ойлголтыг илэрхийлэх ойролцоо утгатай үгс - синсетийг орчуулах; орчуулгын үр дүнд үүссэн үгсийг засах; ойлголтыг оновчтой илэрхийлэх үгсийг үнэлэх зэрэг 3 шатлалт *хүний оюуны даалгавараас* (human intelligence task) бүрдэнэ.

Оролцогч даалгавар гүйцэтгэхдээ тодорхой сэдвийн багцаас сонгож тухайн сэдэвтэйгээ холбогдох ойлголтуудтай ажиллана. Ингэснээр оролцогчдод даалгаврууд санамсаргүйгээр хуваарилахаас зайлсхийх юм.

Бид хөгжүүлсэн програм хангамжаа серверт байршуулсан[3] бөгөөд оролцогчид ажиллахад бэлэн болсон байгаа. Бид энэ програм хангамжийг нээлттэй эхийнх[4] болгосон нь энэ төрлийн ажилд оруулж буй хувь нэмэр бөгөөд хэн ч дотоод системдээ суулгаад өөрчлөн ашиглахад чөлөөтэй.

Энэ өгүүллийн II хэсэгт холбогдох ажлуудын дэлгэрэнгүй, III бүлэгт НМЦ-ийн тухай, IV бүлэгт хүний оюуны даалгаврын зохиомж, V бүлэгт даалгавар үүсгэх алгоритмыг, VI бүлэгт хэрэгжүүлэлтийн тухай, эцэст нь VII бүлэгт ажлын дүгнэлтийг тус тус хавсаргалаа.

[1] http://wiki.dbpedia.org/services-resources/ontology
[2] http://babelnet.org/
[3] http://lkc.num.edu.mn
[4] https://github.com/brainylark/lkc2

II. ХОЛБОГДОХ АЖЛУУД

Мэргэжлийн бус оролцогчдоор зорилтот хэлэнд нутагшуулалтын аргаар хэлний нөөц үүсгэх нь хэд хэдэн дэд бүлгийн ажлуудтай уялдана. Аргачлал болоод зорилгын хувьд хамгийн ойрхон тусах нь Амарсанаа нарын гүйцэтгэсэн судалгаа юм [4]. Энэ ажлаар Англи хэлний 99 ойлголтыг Англи-Монгол хос хэлтэй 44 веб хэрэглэгчдээр хийлгэхэд 77 хувийн нарийвчлалтай 78 ойлголтыг нь орчуулсан. Үүгээр нутагшуулалтад веб хэрэглэгчдээр орчуулгын ажил хийлгэх нь боломжийн үр дүнтэй болохыг харуулжээ.

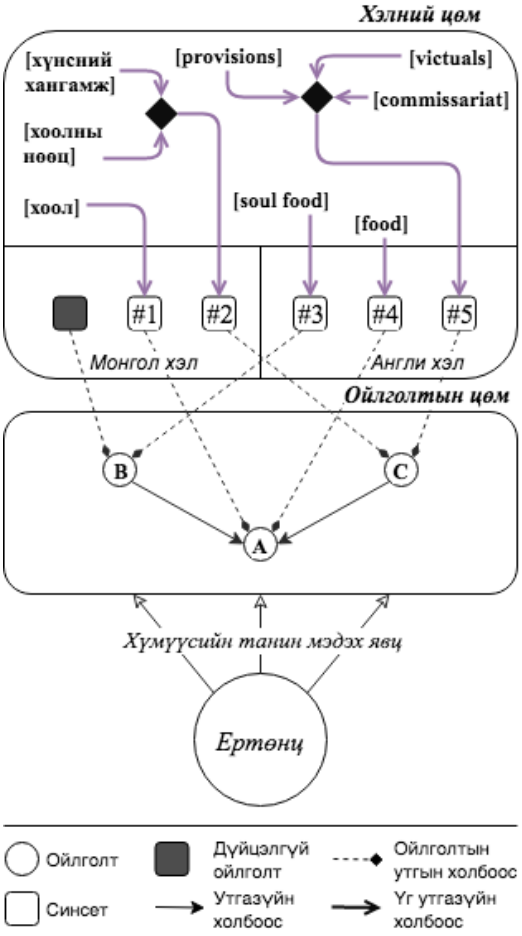
Мөн DBPedia [3] мэдлэгийн сангаас Англи хэлний онтологийн тодорхой хэсгийг Япон хэл рүү олны хүчээр орчуулах аргаар нутагшуулах туршилтын судалгааг Лансер нар [7] хийсэн байдаг. Энэ туршилтаар орчуулгын болон үнэлгээний хоёр үет ажлын дараалал бидний боловсруулж буй аргачлалтай төстэй. Цаашлаад сайн орчуулгыг сонгохын тулд судлаачид олон төрлийн орчуулгыг оролцогчдоор хийлгэн, аль орчуулга хамгийн сайн чанартайг таамаглах зорилготой загваруудыг зохиомжилдог [8], [9], [10] хандлага энэ ажилд төдийгүй бидний аргачлалд тусгагдсан байгаа нь нийтлэг чанар юм.

Орос хэлний Wordnet хөгжүүлэх зорилгоор Браславски нар [6] оролцогчдыг Wiktionary болон Wikipedia зэрэг үгсийн сангаас бэлэн байсан нөөцөөр хангаж, өгөгдсөн толгой үгсэд тэндээс ойролцоо үгсийг сонгуулах, эсвэл өөрөө үг үүсгэх замаар ойлголтын синсетийг үүсгэх туршилт явуулсж 800 орчуулгын үр дүнгийн 434 нь маш сайн, 253 нь боломжийн ба 113 нь тааруу гэж үнэлэгдсэн нь тухайн ажлын дарааллыг чанартай зохиомжлогдсонг харуулж байна.

Оролцогчдын хэлний түвшинд тохирсон даалгаруудыг хуваарилах нь олны хүчээр үүсгэх орчуулгыг чанаржуулна гэж Амбати нар [11] дүгнэсэн байдаг. Тэд туршилтандаа Тэлүгү-Англи, Тэлүгү-Энэтхэг тус бүр 100 өгүүлбэрүүдийг уламжлалт ба өөрсдийн боловсруулсан буюу хамтарсан аргачлалаараа туршилт явуулсан. Үр дүнд нь BLEU [12] хэмжигдэхүүнээрээ хамтарсан аргачлал нь уламжлалт аргачлалаасаа 6.60 дахин илүү оноо дүрсэлсэн билээ.

sloWCrowd ашиглан хэл шинжлэлийн даалгаруудад зориулсан олны хүчийг ашиглах нээлттэй эхийн системийг ашиглан, Фишер нар [13] автоматаар хөгжүүлсэн Словен хэлний ВордНетийн алдааг оролцогчдоор цохож тэмдэглүүлэх туршилт явуулсан байдаг. Улмаар оролцогчдын дундаж нарийвчлал нь 80.12%-аар дүгнэгдсэн нь хэлний бэрх, төвөгтэй утгазүйн даалгаврын хувьд өндөр үзүүлэлт байж. Үүнээс гадна, Клубика ба Лжубезик нар [14] оролцогчдоор Хорват хэлний стандарт бичвэрийн цуглуулгын үнэлүүлж байхдаа нэг оролцогч 90%, гурван оролцогчийн олонхийн хариулт нь дунджаар 97% нарийвчлалтай байсныг тогтоожээ.

Зөвхөн хэл шинжлэлийн хүрээгээр хязгаарлагдалгүй, олны хүчийг ашигласан орчуулгын ажлууд аливаа системийн бүрэлдхүүн хэсэг болж ажилласаар байгаа



Зураг 1. Ертөнцийн Мэдлэгийн Сангийн загвар

билээ. Жишээлбэл, сайн дурын ажилчид Facebook[5] вебсайтыг олон хэлүүд рүү нутагшуулсан [15]. Мөн Duolingo[6] платформ янз бүрийн баримтын орчуулгыг оролцогчдоор хийлгэдэг байсан ба тэдгээрийг хэл сурах тоглоомын нөөц болгон хувиргасан [16].

Дээрх ажлуудад хэрэгжүүлсэн аргачлалууд нь *орчуулах-үнэлэх* ба *орчуулах-засварлах* ангиудад хуваагдана. Гараар үнэлэх ажлыг оролцогчид маш сайн гүйцэтгэдэг нь сайн орчуулгыг ялган авахад зайлшгүй хэрэгтэй. Харамсалтай нь олон сонголттой үед үнэлээчид тун цөөхөн обьектийг зөрж бөглөхөд санал нийцлийн коеффициент асар ихээр унадаг. Үнэлээчдэд үгийн сангийн алдаатай, ойлголтод тохиромжгүй үгс орчуулгын шатаас шууд нэгтгэгдэн ирэх нь энэ асуудалтай тулгарах үндэс юм. Тийм учраас бидний дэвшүүлж буй гурван шатлалт *орчуулах-засварлах-үнэлэх* ажлын урсгал нь уг асуудлыг шийдвэрлэхэд нэмэр болно гэж итгэж байгаа билээ.

III. НУТГИЙН МЭДЛЭГИЙН ЦӨМ

A. Ойлголт ба Синсет

Олон зууны турш хүмүүс бид орчлон ертөнцийг харж, сонсож, таньж мэдэж ирсэн. Энэ ертөнцөд олон дахин ажиглагдсаны үндсэн дээр бидний нэрлэж, зааж хэвшсэн ямар ч зүйлс нь **ойлголт** юм. Жиүнхилиа нар [17] ойлголтыг *ажиглагдаж байгаа зүйлийн оюуны дүрслэл* гэж тодорхойлсон байдаг.

Нэг **үг** олон **утга** илэрхийлж болдог. Харин утга ба ойлголт нь нэг-ба-нэг холбоосоор хамааралтай. Миллерийн томьёолол нь [18] ”*хэрвээ хүн тайлбарыг нь уншаад ойлголтоо авчихсан бол түүнийг тодорхойлоход ойролцоо үгс л хангалттай*” гэсэн дифференциал онолд түшиглэдэг. Жишээлбэл, {’сонсох’, ’чагнах’, ’дуулах’} зэрэг үгс манай хэлэнд нэг ойлголтыг заана. Иймд нэг ойлголтыг илэрхийлж байгаа ойролцоо үгсийг **синсет** гэнэ.

Принстон Ворднет [7] ба олон хэлний ВордНет [8] ойлголтуудаа синсетээр дүрсэлдэг. Гэвч тэдгээрийн Англи хэл рүү хэт хазайсан онтологи нь бусад орны өв соёлын онцгой ойлголтуудыг хамруулж чадахгүй. Олон хэлний нөөц нь универсаль шинж чанараа авч үлдэхийн тулд ойлголтыг үгийн сангаас салгасан архитектуртай байх хэрэгтэй. Жишээ нь, ’монгол гэр’ ойлголт нь манайхаас өөр хэл соёлд байхгүй ч гэсэн гарцаагүй ертөнцийн ойлголтуудын нэг учраас хэлний нөөцөд орох учиртай.

Энэ шаардлагыг хангах мэдлэгийн сан нь **Ертөнцийн Мэдлэгийн Сан** (ЕМС) (Universal Knowledge Core) [19] аж. ЕМС нь хэлнээс хамаарахгүй ойлголтууд ба ойлголтуудын утгазүйн холбоо хамаарлыг формаль хэл дээр буулгасан **ойлголтын цөм** (ОЦ) (Concept Core), ойлголтуудыг ялгаатай хэлүүдэд тохирсон үгийн сангаар нь илэрхийлэх **эх хэлний цөмийг** (ЭХЦ) (Natural Language Core) агуулсан онтологи билээ.

Зураг 1 дээр ЕМС-ийн жишээ загварыг үзүүлэв. ОЦ-д A, B ба C гурван ойлголтыг үзүүлжээ. Ойлголт хоорондын утгазүйн холбоос нь ойлголтуудын эцэг-хүү хамаарлыг дүрслэнэ (B бол A, C бол A). Энд жишээ нь C ойлголтыг ЭХЦ-д Монгол хэлээр #2, Англи хэлээр #5 дугаартай синсетээр тус тус илэрхийлж байна. Тодорхой үгсийг үг-утгазүйн холбоосоор синсеттэй холбосон нь тухайн ойлголтыг төлөөлөх синсетийн үгс гэдгийг харуулна. Энгийнээр, {’хүнсний хангамж’, ’хоолны нөөц’} (C ойлголт) бол {’хоол’} (A ойлголт) гэсэн логикийг ОЦ ба ЭХЦ-ийн давхаргад ангилсан байдал юм.

ЭХЦ-д оршиж байгаа аль нэг хэлний синсетүүдийг багтаасан сан нь тухайн хэлнийхээ **Нутгийн Мэдлэгийн Цөм** болно. Иймд Монгол НМЦ-г үүсгэх бол чухамдаа ОЦ-д орших ойлголтуудыг заах монгол синсетүүдийг үүсгэх ажил. Харин үүсгэх аргачлал нь аль хэдийн үүссэн (эсвэл үүсэж байгаа) хэлнүүдийн синсетийн үгсийг Монгол хэл рүү орчуулан буулгах замаар гүйцэтгэгдэнэ. Энд *орчуулан буулгана* гэдэг нь эх хэлний (Англи ба бусад) синсетийн үгсийг зорилтот (Монгол) хэл рүү үг бүрийг нь орчуулна гэсэн үг биш. Харин тухайн синсетийн илэрхийлж байгаа ойлголтыг манай хэл соёлд ямар үгээр нэрлэдэг болохыг *онооож бичих* гэсэн утгатай юм.

B. Анхдагч ойлголтууд

Олны хүчийг ашиглах нь төрөл бүрийн хүмүүсийн оюуны ашгийг хүртэж буй явдал. Үр шимтэй орчуулга гарган авахад тухайн оролцогчид тохирсон ойлголтыг оноох нь зүйн хэрэг. Гэвч оролцогчдоо таньж мэдсэний үндсэн дээр даалгавар үүсгэж, хуваарилагчийг автоматжуулах нь ямар ч өгөгдөлгүй эхлэх системийн хувьд ярвигтай. Тиймээс оролцогч өөрийн сайн гүйцэтгэл харуулж чадна гэж дүгнэсэн сэдэвтэйгээ холбоотой ойлголтуудаа сонгож ажиллах нь үр дүнтэй.

ОЦ-д бүх нэр үгийн ойлголт удамшдаг 25 анхдагч ойлголтууд байдаг. Тэдгээрийг 25 ширхэг дэд модуудын үндэс гэж төсөөлбөл зохино. Үндсээс доош удамшиж явах зангилаанууд нь эцэг зангилааныхаа илүү онцлог шинжийг тусгасан ойлголтуудыг заана. Жишээлбэл, 6 гүнтэй айг авч үзвэл, ***ачааны тэрэг***–*бол*–***машин***–*бол*–***моторт унаа***–*бол*–***жолоодлоготой унаа***–*бол*–***унаа***–*бол*–***тээврийн хэрэгсэл***–*бол*–***бүтээгдэхүүн*** нь модны нэг зам юм. Бид үндэс ойлголтыг өөрчлөн **ай** гэж нэрлэсэн. Оролцогч айг өөрөө сонгох нь харьцангуй боломжийн гүйцэтгэл үзүүлэх магадлалтайгаас гадна, нэр үгийн төстэй ойлголтуудын ялгааг гаргаж ажиллахад хэрэгтэй.

Дараагийн бүлэгт оролцогчдын системд гүйцэтгэх *орчуулах*, *засварлах*, *үнэлэх* даалгавруудыг дэлгэрэнгүй тайлбарлах болно. Харин оролцогчдод айн ойлголтуудыг ямар дараалаар үүсгэж өгөхийг V хэсэгт тайлагнасан.

IV. ХҮНИЙ ОЮУНЫ ДААЛГАВАР

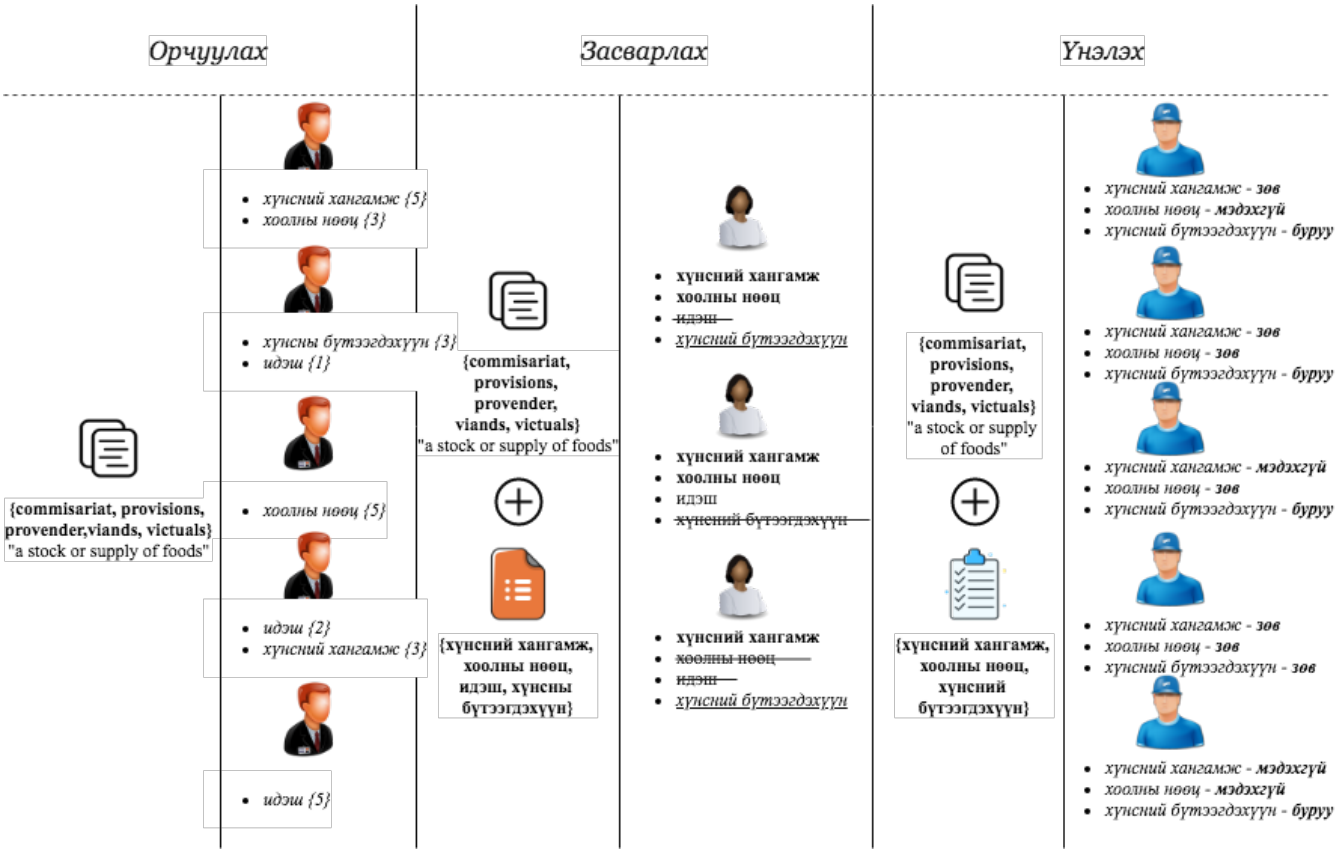
A. Орчуулах

Өмнөх бүлэгт дурьдсанчлан, гадны хэлүүд дээр байгаа синсетийн үгсийг Монгол руу орчуулах даалгавруудыг үүсгэж, оролцогчдын гүйцэтгэлийн үр дүнг нь хадгалах нь манай системийн зорилго. Бид даалгаврыг үүсгэхдээ Англи, Испани, Итали, Хятад ба Индонез зэрэг хэлүүдийн синсетүүдийг эх хэлний хувьд авч үүсгэсэн. Ингэснээр оролцогч зөвхөн Англи хэлнээс орчуулах хязгаарлалтыг үгүй болгож байгаа юм.

Хүний оюуны даалгаврын (ХОД) зохиомж нь ээдрээтэй, олон хэмжээст байх нь оролцогчдоос сайн үр дүн гаргадаггүй [20], харин тэдгээр даалгавруудыг харьцангуй энгийн, хялбар үе шатуудад хуваах нь бүтээмж өндөртэйг онцолсон судалгаанууд байдаг [21], [22].

Дээрх учир шалтгаанд суурилж, системд гүйцэтгэх даалгавруудыг задалж **орчуулах**, **засварлах** ба **үнэлэх** гурван төрөлд ангилсан. Даалгавруудыг тодорхой тооны оролцогчид гүйцэтгэснээр систем тэдгээрийн үр дүнг нэгтгэсэн дараагийн шатны шинэ даалгаврыг үүсгэдэг. Жишээлбэл, нэг синсетийг 5 оролцогч орчуулахад тэдний хариултыг нэгтгэж нэг засварын даалгавар

[5]https://www.facebook.com
[6]https://www.duolingo.com
[7]http://wordnetweb.princeton.edu/perl/webwn
[8]http://multiwordnet.fbk.eu/



Зураг 2. Нэг синсетийг гурван шатлалаар боловсруулахад оролцогчдын үзүүлэх хувь нэмэр

үүсгэнэ. Засварын даалгавар 3 оролцогчоор засагдаж нэг үнэлгээний даалгавар үүснэ. Үнэлгээний даалгавар тус бүрийг 5 оролцогч гүйцэтгэнэ. Өөрөөр хэлбэл, нэг орчуулгын даалгаврын хувьд 13 ялгаатай нэгж ХОД-ын үр дүнг олж авч байна (Зураг 2).

Оролцогч өөрийн сонирхсон нэр үгийн айг сонгоход систем оролцогчид урьдчилан үүсгэсэн даалгаврууд аас хуваарилна. Жишээлбэл, оролцогч орчуулгын даалгавар дотроос ’food’ айг сонгоход систем {’commissariat’, ’provisions’, ’provender’, ’viands’, ’victuals’} үгстэй, ’a stock or supply of foods’ тайлбартай синсетийг хуваарилсан гэж үзье. Энэ үед оролцогчид гурван сонголт байгаа: 1) Синсетийн үгсийг орчуулаад илгээх; 2) Даалгаврыг алгасах; 3) Тухайн ойлголт Монгол хэлэнд байхгүй гэж үзэх буюу дүйцэлгүй ойлголтоор тэмдэглэх. Гурван сонголт гурвуулаа оролцогчийг дараагийн даалгаварт хөтлөнө. Хэрэв оролцогч {’хүнсний хангамж’, ’хоолны нөөц’} гэсэн үгсээр синсетийг Монгол хэлэнд төлөөлүүлсэн гэж үзье. Үр дүнг илгээхийн өмнө оролцогч өөрийн хувийн үнэлгээг үг бүрийн хувьд өгөх шаардлагатай. Хувийн үнэлгээ 1-5 хүртэлх онооноос тогтоно. Энэ нь оролцогч тэр үг тухайн синсетийг илэрхийлж чадна гэдэгт өөрөө хэр итгэлтэй байгаагаа дүрслэх үнэлгээ юм. Орчуулгыг илгээснээр дараагийн орчуулгын даалгавар хуваарилагдаж ирэхэд оролцогч саяны болоод түүнээс урагш орчуулгуудаа сөхөж харж болно. Энэ нь ижил төстэй ойлголтуудад тодорхой үгсийг давтамжтай ашиглахаас сэргийлэх буюу Монгол хэлэнд төрөл ойлголтуудын ялгааг тодруулахад нэмэртэй хэмээн таамаглаж байна.

B. Засварлах

Систем орчуулгын шатанд бий болсон үр дүнгүүдийг нэгтгэснээр засварын даалгаврыг үүсгэдэг. Синсетийг энэ фазаар боловсруулснаар үгзүйн алдаа, ойлголтыг илэрхийлэхэд тохиромжгүй үгс болон хөнөөлт хэрэглэгчдийн ташаа өгөгдлүүдээс ангижрана гэж итгэж байгаа. Жишээ нь: оролцогчид өмнөх шатанд жишээ болгож авсан синсет, мөн орчуулгын үр дүнгүүдээс үүссэн Монгол синсетийн үгс {’хоолны нөөц’, ’хүнсний хангамж’, ’идэш’, ’хүнсны бүтээгдэхүүн’} хуваарилагдаж иржээ. Энэ нөхцөлд оролцогч, алгасах ба дүйцэлгүй ойлголтоор тэмдэглэхийг эс тооцвол, боломжит 3 үйлдэл гүйцэтгэнэ: 1) синсетийн үг(с) алдаатай бол засаж бичих; 2) синсетийн үг(с) алдаагүй бол засах шаардлагагүй гэж тэмдэглэх; 3) синсетийн үг ойлголтыг илэрхийлэхэд тохиромжгүй бол хасах. Синсетийн үгс болон тайлбарыг үзвэл ’хоолны нөөц’, ’хүнсний хангамж’ нь ойлголтыг төлөөлж чадахын зэрэгцээ үгзүйн алдаагүй учраас хэвээр үлдэнэ. Харин ’идэш’ нь мал амьтны тэжээлийг нэрлэхэд түлхүү хэрэглэдэг учраас хасаж болно. Сүүлчийн үг алдаатай бичигдсэн учраас засаж бичвэл зохих юм.

C. Үнэлэх

Үнэлэх даалгавар нь синсетийн орчуулсан үгсийг гаргаж авах сүүлийн шат. Энэ шатанд оролцогчид өөрт оноогдсон даалгаварт байгаа синсетийн үг тус бүрийг дараах 3 тэмдэглэгээний аль нэгээр дугуйлна: 1) Зөв; 2) Буруу; 3) Мэдэхгүй. Өмнө жишээ болгож авсан синсетийн үнэлгээний даалгаварт {'хоолны нөөц', 'хүнсний хангамж', 'хүнсний бүтээгдэхүүн'} үгс иржээ гэж төсөөлье. Оролцогчийг Монгол хэлний дундаж хавьцаа мэдлэгтэй гэж дүгнэвэл 'хоолны нөөц' гэдэг тийм ч танил сонсогдохгүй учраас мэдэхгүйгээр дугуйлах нь зүйн хэрэг. Харин 'хүнсний хангамж' гэдэг үг бидний өдөр тутмын амьдралд тодорхой хэмжээнд хэрэглэгддэг, дуулддаг тул зөвөөр ангилахад нийцнэ. Нөгөөтэйгүүр, 'хүнсний бүтээгдэхүүн' нь хоол хүнсээр хангах нөөц гэх ойлголтыг заахад тохиромжгүй тул буруу гэж оноовол зохиж байна.

V. НУТАГШУУЛАЛТЫН ДААЛГАВАР ҮҮСГЭХ АЛГОРИТМ

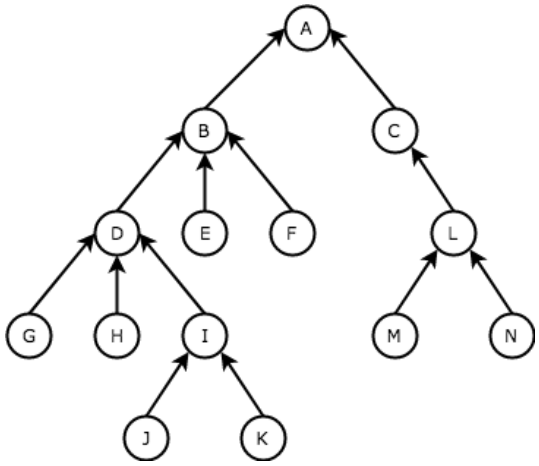
Засварлах болон үнэлэх даалгаврууд үүсэх дараалал нь өмнөх шатнаас нэгтгэгдэж ирэх дарааллаараа үүснэ. Жишээ нь, засварын даалгаврууд аль дарааллаар үүссэн нь хамаагүй, харин түрүүлж 3 оролцогчоор гүйцэтгэгдсэн даалгавар нь үнэлгээний даалгаврууд руу цуварч орно. Харин орчуулах даалгаврыг айн модноос үүсгэнэ.

Алгоритм 1 Даалгавар үүсгэх алгоритм

```
function traverse(parentId) ▷ parentId–эцэг ойлголтын дугаар
    children ← getChildConceptIds(parentId)
    childCount ← children.count
    if childCount == 0 then
        return
    for i = 0 to childCount do
        generateTask(children_i) ▷ Орчуулгын даалгавар үүсгэх
    for i = 0 to childCount do
        traverse(children_i)
```

Даалгавар үүсгэх дарааллыг алгоритм 1 дээр үзүүлжээ. Төсөөтэй ойлголтуудаас дараалан даалгавар үүсгэхийн тулд модоор аялах *түвшний* болон *гүний* нэвтрэлтийн алгоритмыг хослуулсан. Энэ функц анх дуудагдахад *parentId* нь айн үндэс зангилааны ойлголтын дугаар байна. Эцэг ойлголтын хүүхэд ойлголт бүрт даалгавар үүсгээд, дараагаар хүүхэд бүрийг нь эцэг ойлголт гэж үзээд функц өөрийгөө дуудна.

Жишээ модноос даалгаврууд үүсгэж байгаа дарааллыг зураг 3-т харууллаа. Энд A зангилааг үндэс ойлголт гэж үзвэл эхлээд B, C зангилаагаар гүйнэ. Дараа нь B зангилааны хүүхдүүдээр гүйгээд эхний хүүхэд болох D зангилааны хүүхдүүдээр аялна. H, I зангилаанууд хүүхэдгүй тул буцаад J зангилааны хүүхдүүд M, N дээр очих маягаар явна.



Зураг 3. Модноос даалгавар үүсэх дараалал

Нэг эцэгтэй ойлголтуудаас дараалан даалгавар үүсгэх хандлага нь оролцогчдыг ижил төстэй ойлголтуудтай ажиллуулан, тэдний нарийн зааг ялгааг илэрхийлэх үгсийг оноолгоход оршино.

VI. ХЭРЭГЖҮҮЛЭЛТ

A. Програм хангамжийн архитектур

Бидний хөгжүүлсэн системийн програм хангамжийн архитектур нь 7 бүрэлдэхүүн хэсгээс тогтоно. Үүнд:

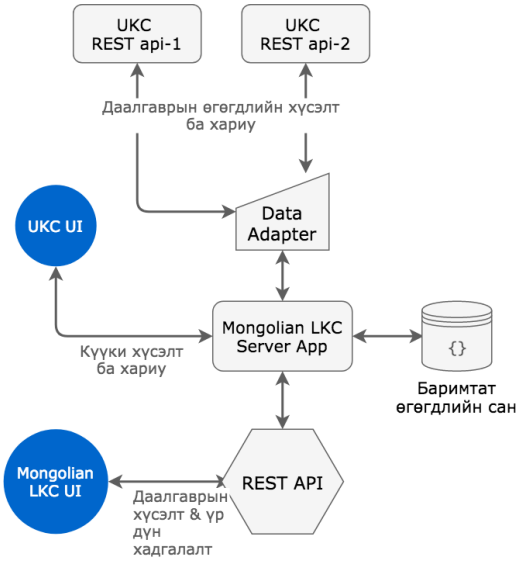
1) **UKC REST-1**: Модны эцэг хүү ойлголтуудын дугаарыг авах API; модоор гүйхэд ашиглагдана
2) **UKC REST-2**: Аливаа ойлголтын дугаарт харьяалагдах олон хэлний синсетийг авах API
3) **Data Adapter**: UKC REST API руу хүсэлт явуулах, ирсэн өгөгдлийг даалгаврын загварын дагуу хувиргах бүрэлдэхүүн
4) **Mongolian LKC Server App**: Сервер дээр ажиллаж буй back-end аппликейшн
5) **UKC UI**: UKC–ийн хэрэглэгчийн интерфейс; нэвтэрч орон күүкиг нь олж авснаар UKC REST API руу хандах эрхтэй болно
6) **REST API**: Системийн өгөгдлүүдийг UI руу илгээх, хүлээн авах интерфейсийг програмчилсан системийн API
7) **Mongolian LKC UI**: REST API ашиглан backend-ээс даалгаврын өгөгдөл авах, үр дүн хадгалах, нэвтрэх бүхий хэрэглэгчийн интерфейс

зэрэг багтана (Зураг 4–г харна уу).

B. Ашигласан технологи

Системийг бүхэлд нь MEAN (MongoDB, ExpressJS, Angular, NodeJS) фреймвөркээр кодлов. Цаашид даалгаврын өгөгдлийн бүтэц боломжит шаардлагуудын өөрчлөлтөөс үүдэн өөрчлөлт ороход өгөгдлийн хадгалалтыг уян хатан шинж чанартай байлгах зорилгоор NoSQL санг сонгосон.

Орчин үед Javascript хэлээр дан ганц front-end гэлтгүй back-end талыг ч програмчилдаг болжээ. Маш хялбар



Зураг 4. Mongolian LKC системийн програм хангамжийн архитектур

аргаар, хурдан хугацаанд бидний шаардлагад нийцсэн RESTful API-г хөгжүүлэх үүднээс ExpressJS фреймворкээр кодлосон юм. Цаашлаад, баримтат өгөгдлийн сан, гадны API-тай харьцахад зориулсан олон олон модулиуд нь ашиглахад тун энгийн агаад өндөр хүчин чадалтай юм.

Эцэст нь системийн front–end-ийг Angular 5 front-end архитектураар, TypeScript хэлээр програмчилсан. Angular–ын *component*, *service*, *data–binding*, *event–binding* зэрэг элементүүд нь front–end хэсгийг цэгцтэй, салангид зохион байгуулах хамгийн оновчтой суурь юм. Мөн Angular Material загвараар харагдацийг Google Translate Community–тай төстэй болгож зохиомжилсон билээ.

VII. ДҮГНЭЛТ

Энэ ажлаар Монгол НМЦ-ийн синсетүүдийг үүсгэх аргачлалыг хэрэгжүүлсэн програм хангамжийн системийг хөгжүүллээ. Бид үүгээр олны хүчийг ашигласан нутагшуулалтын ажлуудад түгээмэл орхигддог оролцогчдод зориулсан даалгавар үүсгэх аргачлалыг, мөн нутагшуулалтын цогц ажлыг цааш улам жижигчилсэн гурван шатлалт хүний оюуны даалгаврыг зохиомжилж хэрэгжүүлсэн юм. Улмаар бид системд оролцогчдыг ажиллуулсны дүнд үүсэх өгөгдлүүдэд шинжилгээ хийн синсетүүдийн чанарыг үнэлэх, цаашлаад аргачлалаа үнэлэх сонирхолтой байгаа. Түүгээр зогсохгүй, синсетийн тайлбарыг орчуулах, синсетийн үгс аль утгаараа өргөн хэрэглэгддэгийг эрэмбэлэх синсетийн үгийн эрэмбэ тогтоох зэрэг ажлууд нь дараа дараагийн шатанд яригдах асуудлууд билээ.

ТАЛАРХАЛ

Энэ судалгааг Монгол улсын их сургуулийн Залуу судлаачийн P2017-2383 дугаартай төслөөс санхүүжүүлсэнд талархаж байна.

Заалт

[1] T. Berners-Lee, J. Hendler, and O. Lassila, “The semantic web,” *Scientific American*, vol. 284, pp. 34–43, May 2001.
[2] N. Savage, “Gaining wisdom from crowds,” *Commun. ACM*, vol. 55, pp. 13–15, Mar. 2012.
[3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, “Dbpedia: A nucleus for a web of open data,” pp. 722–735, 2007.
[4] Амарсанаа and Алтангэрэл, “Онтологи нутагшуулахад олны хүчийг ашиглах туршилт,” in *Монголын Мэдээллийн Технологийн Эрдэм Шинжилгээний Хурал*, 2015.
[5] D. Vannella, D. Jurgens, D. Scarfini, D. Toscani, and R. Navigli, “Validating and extending semantic knowledge bases using video games with a purpose,” in *ACL*, 2014.
[6] P. Braslavski, D. Ustalov, M. Mukhin, and Y. Kiselev, “Yarn: Spinning-in-progress,” 2017.
[7] B. Lanser, C. Unger, and P. Cimiano, “Crowdsourcing Ontology Lexicons,” 2016.
[8] O. F. Zaidan and C. Callison-Burch, “Crowdsourcing translation: Professional quality from non-professionals,” pp. 1220–1229, 2011.
[9] M. Benjamin and P. Radetzky, “Multilingual lexicography with a focus on less-resourced languages: Data mining, expert input, crowdsourcing, and gamification,” 2014.
[10] M. G. Rui Yan, E. Pavlick, and C. Callison-Burch, “Are two heads better than one? crowdsourced translation via a two-step collaboration of non-professional translators and editors,” *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, pp. 1134–1144.
[11] V. Ambati, S. Vogel, and J. Carbonell, “Collaborative workflow for crowdsourcing translation,” pp. 1191–1194, 2012.
[12] K. Papineni, S. Roukos, T. Ward, and W. jing Zhu, “Bleu: a method for automatic evaluation of machine translation,” pp. 311–318, 2002.
[13] D. Fišer, A. Tavčar, and T. Erjavec, “slowcrowd: A crowdsourcing tool for lexicographic tasks,” 2014.
[14] N. L. Filip Klubicka, “Using crowdsourcing in building a morphosyntactically annotated and lemmatized silver standard corpus of croatian,” 2014.
[15] TechCrunch, “Facebook taps users to create translated versions of site. spanish, french and german available now,” 2008.
[16] L. von Ahn, “Duolingo: Learn a language for free while helping to translate the web..”
[17] F. Giunchiglia and M. Fumagalli, “Concepts as (recognition) abilities,” in *Formal Ontology in Information Systems: Proceedings of the 9th International Conference (FOIS 2016)*, p. 153, 2016.
[18] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, “Wordnet: An on-line lexical database,” *International Journal of Lexicography*, vol. 3, pp. 235–244, 1990.
[19] K. B. Fausto Giunchiglia and A. A. Freihat, “One world-seven thousand languages,” in *19th International Conference on Computational Linguistics and Intelligent Text Processing*, 2018.
[20] P. G. Iztok Kosem and S. Krek, “Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing,” in *Electronic Lexicography in the 21st Century: Thinking Outside the Paper. Proceedings of the eLex*, pp. 17–19.
[21] C. Biemann and V. Nygaard, “Crowdsourcing wordnet,” in *Proceedings of the 5th Global WordNet Conference*, (Mumbai, India), 2010.
[22] A. Rumshisky, “Crowdsourcing word sense definition,” in *Proceedings of the 5th Linguistic Annotation Workshop*, LAW V ’11, (Stroudsburg, PA, USA), pp. 74–81, Association for Computational Linguistics, 2011.