Онтологи нутагшуулахад олны хүчийг ашиглах туршилт

Ганболдын Амарсанаа[1], Чагнаагийн Алтангэрэл[2]
Мэдээлэл, компьютерийн ухааны тэнхим
ХШУИС, Монгол Улсын Их Сургууль
Улаанбаатар, Монгол улс
{amarsanaag[1], altangerel[2]}@num.edu.mn

***Хураангуй*—Мэдлэгийн сангийн их хэмжээний өгөгдлийг олны хүчээр бүрдүүлэх арга өргөн нэвтэрч байна. Гэвч олон хэл соёлын мэдлэг шингээсэн мэдлэгийн сангийн онтологийг нутагшуулахад олны хүчийг ашигласан туршлага ховор байдаг. Энэ ажлаар онтологи нутагшуулахад олны хүчийг ашиглах боломжтой эсэхийг судлах юм. Бид онтологийн нэр томъёог орчуулах хүний оюуны даалгавар боловсруулж хос хэлтэй сайн дурын вэб хэрэглэгчдээр Англи хэлээр илэрхийлсэн онтологийг Монгол хэлнээ нутагшуулах туршилтыг хийсэн.**

Түлхүүр үгс—олны хүч; онтологи нутагшуулалт; мэдлэгийн сан

I. Удиртгал

Семантик вэбэд хэрэглэх мэдлэгийн санг үүсгэхэд олон хэлний мэдлэгийг шингээж ялгамжийг (diversity) мэддэг байх асуудал чухал тавигдаж эхэллээ [1]. Учир нь олон хэл соёл дунд нийтлэг ашиглах ойлголтууд байхад зөвхөн тухайн хэл, соёлд хэрэглэдэг ойлголтууд бас байдаг бөгөөд үүнийг ялгамж гэдэг. Ийм ялгамжийг шингээсэн мэдлэгийн сан сая ертөнцийн бүх мэдлэгийг боловсруулах боломжтой болно. Хэдийгээр олон хэлний мэдлэгийн сангууд YAGO [2], DBpedia [3] зэрэг нь ВикиПедиагийн хагас бүтэцлэгдсэн өгөгдлийг задалж гаргасан ч ялгамжийг нарийн тусгаж чадаагүй.

Харин UKC [4], [5] мэдлэгийн санг үүсгэхдээ онтологийг нутагшуулах гар аргаар ялгамжийг барьж авах асуудлыг туршжээ [6]. Гэвч их хэмжээний мэдлэгийн санг мэргэжлийн хүмүүсээр гар аргаар үүсгэх нь цаг хугацаа, зардал их шаарддаг бөгөөд хэлний цахим нөөц багатай хэл соёлын хувьд автомат аргыг ашиглах боломжгүй. Иймд бага нөөцтэй хэл соёлын мэдлэгийн санг олны хүчээр үүсгэх боломжийг судлах нь энэ ажлын гол зорилго юм. Олны хүч гэдэг нь олон нийтийн нэг хүн бүрийн хувь нэмрийг их хэмжээгээр цуглуулж хүний оюуны ашгийг хүртэх үйл явцыг хэлнэ. Бид энэ ажлаар UKC мэдлэгийн сангийн орон зайн айн онтологийн нэр томъёог Англи хэлээс Монгол хэлрүү орчуулах зэрэг хүний оюуны даалгавар (human intelligence task) боловсруулж вэб хэрэглэгчдээр гүйцэтгүүлж мэдлэгийн санг олны хүчээр нутагшуулах арга зүйг туршсан анхан шатны үр дүн гаргасан.

Энэ өгүүллийн 2-р бүлэгт орчуулгын ажлыг олны хүчээр хийх судалгааны ажлуудыг эргэцүүлж тайлбарласан. 3-р бүлэгт олны хүчээр мэдлэгийн санг нутагшуулах хүний оюуны даалгаврын зохиомжийг харуулсан. 4, 5-р бүлэгт хийсэн туршилт, энэ судалгааны дүгнэлтийг тусгалаа.

II. Олны Хүч ба Орчуулга

Хамтын ажиллагаат орчуулгын ажлын урсгалын зохиомж [7], номын сангийн каталог бүрдүүлэхэд орчуулга ашиглах [8], мэргэжлийн бус орчуулгын чанарын удирдлагын загвар болон орчуулгын чанарыг шалгах [9], [10] зэрэг ажлууд нь олны хүчээр орчуулга хийх, орчуулгын чанарыг үнэлэх сэдвийг хөндсөн байна.

Хамтын ажиллагаат орчуулга нь эхлээд өгүүлбэрийн үг бүрээр орчуулаад дараа нь хос хэлтэй хэрэглэгчдээр өгүүлбэрийг гүйцээж орчуулуулдаг ба эцэст нь орчуулах хэлний хэрэглэгчид тухайн өгүүлбэрт тохирох хамгийн зөв орчуулгыг үнэлж тогтоожээ. Энэ ажлаар өгүүлбэрийг бүхэлд нь нэг орчуулагчаар орчуулуулах бус өгүүлбэрийн бүрэлдэхүүн хэсэг тус бүрт нягт хамтын ажиллагаатай орчуулгыг хийсэн нь гол ололт юм. Мөн энэ арга нь машин орчуулгын чанарыг үнэлэхэд өргөн хэрэглэдэг BLUE (Bilingual Evaluation Understudy) аргаас арай илүү чанартай байсан.

Олны хүчээр хийх орчуулгын чанарын удирдлагын загварт хүний уншиж ойлгох чадварыг ашигласан ба эхээс асуулт асууж зөв хариулж байгааг нь тооцоолжээ. Энэ арга өмнөх бүлэгт дурдсан бидний шаардлагыг хангах боломжтой. Учир нь ойлголтыг илэрхийлэх үгийг орчуулах нь агуулгаас ихээхэн хамаардаг бөгөөд хүмүүс тухайн үгийг харахад ямар ойлголтыг төсөөлж байгаагаар нь шалгаж болно. Өөрөөр хэлбэл, орчуулга нь хэр ойлгомжтой байгааг үнэлэх нь хамгийн зөвийг харьцуулж үнэлэх бусад аргаас илүү үр дүнтэй.

Орчуулгыг гар аргаар үнэлэх нь мэргэжлийн орчуулагч шиг өндөр чанартай орчуулга гарган авч болдог, үүнийг олны хүчний платформ ашиглан хямд зардлаар гүйцэтгэх боломжтойг нотлон харуулсан байна. Энэ ажилд судалсан гараар үнэлэх аргуудыг хамгийн алдартай хүний оюуны хөдөлмөрийн онлайн зах болох Amazon’s Mechanical Turk

(заримдаа MTurk гэдэг) платформыг ашиглан амжилттай туршжээ.

Дээрх аргууд олны хүчний орчуулгыг гар аргаар үнэлэх болон дэс дараатай олон давтагдах жижиг даалгаруудад хувааж гүйцэтгүүлэхэд хямд зардлаар чанартай үр дүн гарган авч болохыг баталж байна.

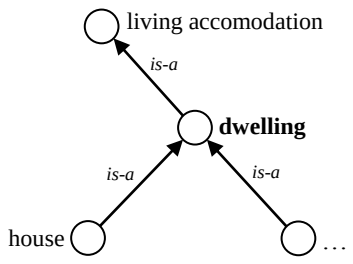
III. Олны Хүчээр Онтологи Нутагшуулах Арга

Онтологи нутагшуулах гэдэг нь мэдлэгийн зарим хэсгийг ямар нэг хэл соёлруу нийцүүлэн буулгах үйл явцыг хэлнэ. Ийм ажлыг сэтгэц хэл шинжлэлийн мэргэжилтнүүд хийдэг. Тэгвэл онтологийг олны хүчээр, мэргэжлийн бус хүмүүсээр онтологийг нутагшуулах боломжийг энэ ажлаар судалсан. Бидний ашиглах UKC мэдлэгийн сан нэгнээсээ үл хамаарах үгийн сангийн өгөгдлийн сангуудыг агуулдаг бөгөөд шаталсан хэлбэрээр зохион байгуулсан синсетийн сан ВөрдНэтийг бүхэлд нь MultiWordNet-ийн Итали хэлний хэсэгтэй нэгтгэж үүссэн [4]. Энд үгийн сангийн нэг өгөгдлийн сан мэдлэгийн санг бүхэлд нь илэрхийлдэг. Мэдлэгийн сангийн нэг ойлголтыг нэг синсетээр илэрхийлэх ба синсет бол ямар нэг хэлэнд тэр ойлголтыг илэрхийлж чадах ижил утгатай үгсийн олонлог. Жишээ нь, *хот* гэх ойлголтыг Англи хэл дээр *city, metropolis, urban center*, Итали хэл дээр *città* (дуудлага: chit'a) зэрэг үгсийн олонлогууд хэл тус бүртээ илэрхийлнэ. Иймд тухайн ойлголт ялгаатай хэлүүд дээр нэг утгыг илэрхийлэх ба өөр өөр хэлний хэд хэдэн синсеттэй холбогддог. Синсет бүр өөрийн илэрхийлэх ойлголтыг тайлбарлах тайлбартай байна. Ийм онцлог өгөгдлийн сан тухайн ойлголтод тохирох хамгийн оновчтой орчуулгыг шаарддаг. Учир нь нэг үг олон утгатай байж болох ба үгийн утгыг машин орчуулгаас илүү хүн хамгийн сайн салгаж чадна. Ийм учраас энэ ажилд хүний оюуныг зайлшгүй ашиглах шаардлагатай юм.

A. *Синсет ба хүний оюуны даалгаврын зохиомж*

Аль нэг хэлний синсетэд тохирох өөр хэлний синсетийг (зорилгын синсет гэх) тодорхойлохдоо эх синсетийг орчуулах аргаар гүйцэтгэж болно. Үүний тулд зорилгын синсетэд үгийн сангийн нэгж(үүд)ийг оноож бичих шаардлагатай. Ийм ажлыг олон тооны вэб хэрэглэгчдээр синсетийг орчуулах хүний оюуны даалгавар гүйцэтгүүлж хамгийн тохирох синсетийг олж болно. Үгийн сангийн нэгж гэдэг нь толь бичгийн нэг толгой үг эсвэл нэг хэлц үг эсвэл нэг холбоо үгийг хэлнэ. Жишээ нь Зур. 1-д харуулсан *dwelling* ойлголтыг илэрхийлэх Англи синсетийг дараах байдлаар харуулав.

dwelling, home, domicile, abode, dwelling house, habitation (housing that someone is living in)



Зур. 1. Орон зайн онтлогийн нэг дэд хэсэг

Энэ ойлголтыг Монгол хэлэнд дараах синсетээр нутагшуулж болно.

гэр орон, орон гэр, гэр, орон байр, оршин суугаа газар (хүн амьдарч байгаа байр)

Онтологийг нутагшуулах явцад орчуулах боломжгүй нэр томъёо гарч ирдэг ба үүнийг дүйцэлгүй үг (lexical gap) гэдэг. Дүйцэлгүй үг нь тухайн ойлголтыг нутагшуулж буй хэл соёл байхгүй эсвэл түүнийг үгийн сангийн нэгжээр илэрхийлж хэвшээгүйг илтгэж байдаг. Жишээ нь, *submarine furrow (a closed, linear, narrow, shallow depression)* Монгол хэлэнд дүйцэлгүй үг *(усан доорх урт, нарийхан, бага зэргийн хотгор газар)* болно.

Иймд вэб хэрэглэгчдээс тухайн ойлголтыг илэрхийлэх синсет оноох эсвэл дүйцэлгүй үг тэмдэглэгээ хийхийг асууж синсет бүрийн хувьд олон хэрэглэгчээс олон үгийн сангийн нэгжийг хүлээж авна. Хэрэв вэб хэрэглэгч синсет оноож мэдэхгүй байвал тухай даалгаврыг гүйцэтгэлгүй алгасах боломжтой. Дараа нь тухайн синсетийн давхардаагүй үгийн сангийн нэгж бүрийг хэр зөв орчуулга болсныг дахин вэб хэрэглэгчдээр үнэлүүлэх бөгөөд үүнийг бас хүний оюуны даалгавраар хийлгэнэ.

B. *Орчуулгыг үнэлэх ба үр дүнг нэгтгэх*

Олны хийсэн ажлыг үнэлэх аргуудыг чанарын болон тоон гэж ангилдаг. Орчуулгын үр дүнд үүссэн үгийн сангийн нэгжийг *Зөв, Буруу, Мэдэхгүй* гэж ангилал хийхэд чанарын аргыг ашиглах нь тохиромжтой байдаг. Тодорхой тооны хүмүүс үгийн сангийн нэгжийг үнэлэх үед үнэлгээнүүд хоорондоо хэр нийцтэй байгааг Fleiss' kappa [11] статистик хэмжүүрийг ашиглаж болно. Ингэж нэг синсет бүрийн хувьд Fleiss' kappa корреляцийн коэффициентийг тооцож (1) синсетийн орчуулга хэр зөв болсныг ерөнхийд нь дүгнэх боломжтой.

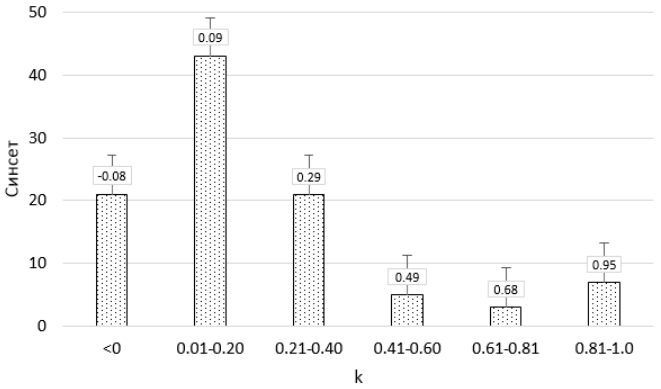
$$k = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e} \quad (1)$$

Энд $\overline{P} - \overline{P}_e$ нь хүрчихээд байгаа санал нийлэмжийн зэрэг, $1 - \overline{P}_e$ нь хүрч болох санал нийлэмжийн зэрэг. к=1 тохиолдолд санал бүрэн нийлсэн, к ≤ 0 үед санал ер нийлээгүй гэж үзнэ. Коэффициентийн утга 0.01-1.0 хооронд байвал тухайн синсетийг үнэлэгчдийн санал ямар нэг байдлаар нийцтэй гэж үздэг. Гэхдээ энэ хооронд хэдийг сонгож авах даалгавраас хамаарч өөр байж болно. Өөрөөр хэлбэл коэффициентийн утга санал өгөх зүйлсийн тоо, ангилалын тооноос хамаарч харилцан адилгүй байдаг. Иймд бидний туршилтанд үүнээс бага байхад тухайн синсетийг үнэлсэн вэб хэрэглэгчид синсетийн үгийн сангийн нэгжүүдийн зөв бурууг зөв ялгаж чадсан гэж ойлгож болно.

Дараа нь сайн нийцтэй үнэлэгдсэн синсетээс дийлэнх олонхийн саналаар *Зөв* үнэлгээ авсан үгийн сангийн нэгжүүдийг ялган гаргаж авах замаар эх синсетэд хамгийн оновчтой тохирох зорилгын синсетийг тодорхойлж үр дүнг нэгтгэнэ.

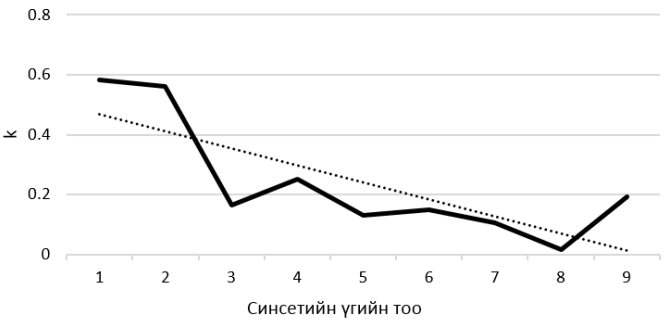
IV. Туршилт

Олны хүчийг ашиглахад тулгардаг эхний асуулт бол хэрэглэгчдийг хэрхэн зохион байгуулах байдаг [12]. Энэ ажлаар Англи хэлээр илэрхийлсэн 99 ойлголтыг Англи-Монгол хос хэлтэй 44 вэб хэрэглэгчдийн хүчийг ашиглан Монгол хэлнээ нутагшуулсан. Нэг синсетийг 10 өөр вэб хэрэглэгчээс асууж нийт 990 синсетийг орчуулуулахад давхардсан тоогоор 639 (давхардаагүй 364) үгийн сангийн нэгж, 20 дүйцэлгүй ойлголт, 411 орчуулалгүй алгассан гэсэн хариу хүлээж авсан юм.



Зур. 2. Синсетийн санал нийлсэн байдал

Зураг 2-т каппа коэфициентийн тайлал бүрт хэчнээн синсет харгалзаж байгааг харуулав. Ямар нэг байдлаар санал нийцтэй гарсан синсетүүдийн эзлэх хувь 80% байна. Мөн хамгийн бага нийцэлтэй гарсан 43 синсетийн дундаж каппа 0.09 байв. Энэ нь синсетийн үгийн тоо ихсэх тусам нийцэл багасдагтай холбоотой (Зураг 3).



Зур. 3. Fleiss' каппагийн дундаж утга

Каппагийн утга 0-ээс их байх синсетүүдийг ялгаж авсны дараа 381 үгийн сангийн нэгж үлдсэн. Дараа нь синсет бүрийн үгийн сангийн нэгжүүдээс тухайн синсетэд үнэлгээ өгсөн вэб хэрэглэгчдийн дийлэнх олонхийн саналаар *Зөв* үнэлгээ авсан үгийн сангийн нэгжүүдийг сонгож авахад 145 үгийн сангийн нэгж үлдсэн. Эдгээр үгсийг Англи хэл, ерөнхий мэдлэгийн зохих түвшинтэй хүнээр хянан тохиолдуулж эцсийн үр дүнг 112 үгтэйгээр гаргаж авсан.

Туршилтын харахад үгийн сангийн нэгжийн оронд тайлбар бичих, үг үсгийн алдаатай бичих зэрэг механик алдаанууд их гаргасан байсан нь олны хүчээр нутагшуулах програм хангамжийн зохиомж, мөн даалгавраар хориглох зүйлсийг нарийн тусгаж өгсөн байх хэрэгтэйг харуулж өглөө. Мөн дүйцэлгүй ойлголт нэг ч гараагүй нь вэб хэрэглэгчид дүйцэлгүй ойлголтыг нарийн тодорхойлж чадахгүй байна.

Дүгнэлт

Энэ ажлаар онтологийг нутагшуулахад олны хүчийг ашиглах туршилтыг амжилттай гүйцэтгэлээ. Туршилтын үр дүнд олны хүчээр нутагшуулсан ойлголтуудыг илэрхийлэх Монгол үгийн сангийн нэгжүүдийг Англи хэл, ерөнхий мэдлэгийн зохих түвшинтэй хүнээр хянан тохиолдуулахад олны хүчээр орчуулсан үгийн сангийн нэгж 77 хувийн нарийвчлалтайгаар 99 ойлголтоос 78-ыг орчуулсан байна. Үүнийг ямар ч шалгуур тавиагүй вэб хэрэглэгчдээр гүйцэтгүүлсэн гэхэд боломжийн үзүүлэлт болж чадна. Учир нь үр дүнг шалгах асуулт оруулах, ажлын явцын дунд синсетийг үнэлүүлж тухайн хэрэглэгчийн оруулсан нэмрийн чанарыг тооцох зэргээр мэхлэгч хэрэглэгчийг илрүүлэх, эсрэгээрээ чанартай хувь нэмэр оруулдаг хэрэглэгчдийг урамшуулах замаар шаардлага хангасан хэрэглэгчдээр хийлгэж үр дүнг өсгөж болно.

Заалт

[1] A. Ganbold, F. Farazi, M. Reyad, O. Nyamdavaa, and F. Giunchiglia, "Managing Language Diversity Across Cultures : the English-Mongolian Case Study," *Int. J. Adv. Life Sci.*, vol. 6, no. 3, pp. 167–176, 2014.

[2] F. M. Suchanek, G. Kasneci, and G. Weikum, "YAGO: A Large Ontology from Wikipedia and WordNet," *Web Semant. Sci. Serv. Agents World Wide Web*, vol. 6, no. 3, pp. 203–217, Sep. 2008.

[3] C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia : A Nucleus for a Web of Open Data," *Nucleus*, vol. 4825, pp. 722–735, 2007.

[4] F. Giunchiglia, V. Maltese, and D. Biswanath, "Domains and context: first steps towards managing diversity in knowledge," *Web Semant. Sci. Serv. Agents World Wide Web*, vol. 12–13, pp. 53–63, 2012.

[5] F. Giunchiglia, B. Dutta, V. Maltese, and F. Farazi, "A facet-based methodology for the construction of a large-scale geospatial ontology," *J. Data Semant.*, vol. 1, no. 1, pp. 57–73, 2012.

[6] A. Ganbold, F. Farazi, and F. Giunchiglia, "An Experiment in Managing Language Diversity Across Cultures," in *eKNOW 2014 : The Sixth International Conference on Information, Process, and Knowledge Management*, 2014, no. c, pp. 51–57.

[7] Vamshi Ambati, Stephan Vogel, and Jaime Carbonell, "Collaborative workflow for crowdsourcing translation," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 2012, pp. 1191–1194.

[8] J. Corney, A. Lynn, C. Torres, P. Di Maio, W. Regli, G. Forbes, and L. Tobin, "Towards crowdsourcing translation tasks in library cataloguing, a pilot study," in *Digital Ecosystems and Technologies DEST 2010 4th IEEE International Conference on*, 2010, pp. 572–577.

[9] C. Callison-Burch, "Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk," *Lang. Speech*, vol. 1, no. August, p. 286, 2009.

[10] Omar F. Zaidan and Chris Callison-Burch, "Crowdsourcing translation: professional quality from non-professionals," in *HLT'11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, 2011, pp. 1220–1229.

[11] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, no. 5, pp. 378–382, 1971.

[12] A. Doan, R. Ramakrishnan, and A. Y. Halevy, "Crowdsourcing systems on the World-Wide Web," *Commun. ACM*, vol. 54, no. 4, p. 86, 2011.