МОНГОЛ УЛСЫН ИХ СУРГУУЛЬ
МОНГОЛ ХЭЛ СОЁЛЫН СУРГУУЛЬ
эрдэм шинжилгээний бичиг

Том. XXXI (328) 2010 1-5

ВЕБ - МЭДЛЭГИЙН ХӨМРӨГ БОЛОХ НЬ

Ч.Алтангэрэл[1]
Ж.Пүрэв[2]

Веб бол компьютерийн хулганы товшилтоор бүхэл бүтэн мэдээллийн, мэдлэгийн ертөнцийг хэрэглэгчийн өмнө нээх "хязгааргүй", чөлөөтэй эх үүсвэр юм. Веб нь 1994 оноос Интернетийн салшгүй хэсэг болж хөгжихийн зэрэгцээ их хэмжээний материал өөртөө агуулдаг нь түүний хамгийн гол давуу тал юм. Keller ба Lapata (2003) нарын судалгаагаар 2003 онд Google-ээр 98 гаруй тэрбум үгийн индексийг үүсгэжээ. Google зэрэг хайлтын хөдөлгүүр системийн бий болгосон хурдан, хялбар хандалтын үр дүнд Веб нь мэдээллийн эх үүсвэрийн хувьд анхаарал татах болсон.

Анхлан газарзүйн байрлалын хувьд алслагдсан хэрэглэгчид мэдээлэл хуваалцах зорилгоор гарч ирсэн веб нь одоо энгийн жижиг хэрэглээнээс эхлээд маш том бизнес, үйлдвэрлэлийн төлөвлөлт зохицуулалт, хамтын ажилллагааны хэрэглээг түгээх орчин болтлоо хөгжжээ. Аялал жуулчлал, үйлдвэрлэл, банк, боловсролын байгууллага болон засгийн газар үүнийг өөрсдийн үйл ажиллагааг сайжруулж нэмэгдүүлэх зорилгоор вебэд суурилсан хэрэглээг ашиглаж байна (Whitten N, 1995; Adam K,2003).

Түүнчлэн орчин үеийн хэрэглээний системүүд нь мэдээлэлд бус мэдлэгт тулгуурласан чиглэл рүү хандаж байгаа бөгөөд вебээс мэдлэгийг төрөл бүрийн арга техникээр гарган авч ашиглаж байна. Тэдгээр нь энгийн статистик үзүүлэлтээс эхлээд бүр нарийн боловсронгуй тооцоонд тулгуурласан байдаг.

Эдгээр мэдлэгийн нэг хэлбэр болох үгийн утгазүйн сүлжээ (Lexical Semantic Network) нь хүний хэл эзэмших гол 2 элемент болох үгийн сан болон энэхүү сангийн хэсгүүдийн хоорондох уялдаа холбоог илэрхийлдэг. Энэ нь мэдээллийн хайлт, семантик вэб, машин орчуулга, хэлний сургалт, боловсрол зэрэг маш олон мэдлэгт тулгуурласан системийн үндсэн суурь болон хэрэглэгдэж байгаа бөгөөд үүнийг гараар үүсгэх ажил маш их цаг хугацаа, хөдөлмөр шаарддаг. Жишээ нь Англий Princeton WordNet-ийг судалгааны нэг баг 15 жилийн турш гараар хийж сайжруулжээ. Бусад өндөр хөгжилтэй улс орнууд өөрсдийн үгийн утгазүйн сүлжээг (Англи: Princeton WordNet, Герман: GermaNet, Солонгос: KorLex UWIN etc.) үүсгэн улмаар хооронд нь холбон тив дэлхийн хэмжээнд олон хэлний сүлжээг (Европийн орнуудын EuroWordNet, Балканы орнуудын BalkaNet, Азийн орнуудын Asian WordNet etc.) үүсгэн ашиглаж байна. Энэхүү олон хэлний сүлжээ нь машин орчуулга, хэл дамнан мэдээлэл хайх (Cross Lingual Information Retrieval), хэлний боловсрол, сургалт (language education) зэрэг олон салбарт зайлшгүй чухал хэрэгцээтэй байдаг. Иймээс Монгол хэлний үгийн

[1] Доктор. МУИС-ийн Мэдээлэл технологийн сургууль.
[2] Доктор. МУИС-ийн Мэдээлэл технологийн сургууль.

утгазүйн сүлжээг (хагас) автоматаар, богино хугацаанд, зардал багатайгаар, чанартай үүсгэх боломжийг судлах нь нийгэм эдийн засаг, инновацийн чухал ач холбогдолтой асуудал болоод байна. Дэлхий нийтээр сонирхон судалж байгаа ийм төрлийн мэдлэгийг автоматаар үүсгэх, хүний гараар хийсэн мэдлэгтэй харьцуулж үзэх, ингэснээр хүний субьектив оролцоо хэр байгааг илэрхийлэх, хүний мэдлэгийг хэрхэн загварчлах зэрэг ерөнхий суурь судалгаануудыг хийхэд зарим улс орон, эрдэмтэд вебийг мэдлэг хайх үндсэн эх, сууриа болгон ашиглаж байна.

Вебийг мэдлэгийн эх болгосон өөр нэг хэрэглээ нь машин орчуулгад хэрэглэх параллел корпус үүсгэхэд хэрэглэсэн байдал юм. Машин орчуулгын үндсэн аргуудын нэг болох статистик аргад хөрвүүлэг хийх хоёр хэлний параллел материалын санг их хэмжээгээр ашигладаг бөгөөд энэ санг цуглуулах нь мөн л маш их цаг хугацаа хөдөлмөр шаардсан ажил болдог. Жишээлбэл энэхүү ажлыг хөнгөвчлөх болон илүү их мэдлэгийг олж авахын тулд Хойд Техасын их сургууль веб хуудаснаас параллел материалын санг автоматаар цуглуулах Бавилон системийг хөгжүүлсэн байна.

Монгол хэлний хувьд вебийг энэ түвшинд хүртэл ашиглах боломж нь одоогоор дутмаг байгаа нь вебийн агуулга, тогтвортой ажиллагаа болон веб дээрх бичвэрийн чанартай холбон тайлбарлаж болох юм.

Үгийг зөв бичсэн эсэхийг шалгах жишээг энгийн статистик аргаар авч үзье. Google хайлтаар "*морины*" гэдэг үгийг Хүснэгт 1-д үзүүлсэн шигээр олон янзаар бичиж хэрэглэснийг илрүүлж харж болно. Эндээс харахад морины гэдэг зөв бичсэн үг нь хамгийн их давтамжтай байна.

Хүснэгт 1: "морины" гэдэг үгийн хэрэглээ

	давтамж
морины	65600
морьны	1190
морний	291
мориний	268
морьний	104

Веб нь хэл шинжлэл, хэл боловсруулалт, хиймэл оюун ухаан болон бусад олон салбарын судлаачдын хувьд жинхэнэ хэлний өгөгдөл буюу корпус болж байна. Хэл шинжлэлийн судалгаанд хэрэглэх бичгийн материалын сан буюу корпусыг гол төлөв сонины өгүүлэл, ном зохиол зэрэг хэвлэмэл эх бичгээс үүсгэдэг уламжлалтай. Харин мэдээллийн эх үүсвэр болох Вебийн үсрэнгүй өсөлтөөс улбаалан түүнийг хэл боловсруулалтын бодлогын сургалтын өгөгдөл хэлбэрээр улам ихээр хэрэглэх болсон байна.

Материалын санг Вебээс бүтээх нь хэвлэмэл зүйл ашиглахаас олон давуу талтай. Веб өгөгдөл нь машин уншиж авч чадах цахим хэлбэрт байдаг. Гэтэл хэвлэмэл материал нь бүхэлдээ цахим хэлбэрээр байдаггүй.

Сургалтын өгөгдлийн хэмжээ өсөн нэмэгдэх бүр NLP (natural language processing буюу хэл боловсруулалт) системийн бүтээмж сайжирдаг. Banko ба Brill (2001) нар сургах өгөгдлийн хэмжээг, жишээ нь 1 тэрбум үг хүртэл нэмэгдүүлснээр зөв бичих зүйн алдааг илрүүлэх нарийвчлал нэмэгдэж байсныг өөрсдийн туршилтаар тогтоосон байна.

Интернэтэд байрлах мэдээ, мэдээлэл, бусад зүйл нь Монгол улсын мэдээллийн өмчийн нэгээхэн хэсэг, нийгмийн баялаг тул тэдгээрийг эхнээс нь алдаагүй зөв бэлдэх ёстой. Ингэснээр цаашдын боловсруулалтын ажил дөхөм болно.

Веб бичвэрийг янз бүрийн зохиогчид бүтээдэг. Хэвлэмэл материалыг бодвол веб материалыг түүний зөв эсэхийн талаар бага анхаарч хямдхан, хурданаар бүтээж болдог. Интернэт болон бусад цахим эх үүсвэрээс цуглуулах явцад алдааг илрүүлж зөв болгоход ихээхэн цаг зарцуулдаг. 5 сая үгтэй материалын санг цэвэрлэхэд[3] 3 судлаач бүхэл бүтэн 5 сарын хөдөлмөр зарцуулсан ч бүрэн дуусгаагүй. Бидний хийсэн туршилтаас харахад давхцсан байдлаар интернет дэх Монгол бичвэр доторхи үгийн 20-иод хувь нь алдаатай байна.

Хүснэгт 2: Туршилтанд ашигласан зарим бичвэрийн статистик

#	Words in text	Incorrect words
1.	683	83
2.	725	114
3.	726	257
4.	805	85
5.	772	104
6.	939	208
7.	726	118
8.	810	243
9.	770	122
10.	729	200
	768 (100%)	153 (19.9%)

Энэ гарч байгаа алдааг ерөнхийд нь дараах байдлаар ангилж болно.

- Зөв бичгийн дүрмийн алдаа

  Энэ нь бичвэр оруулж буй хүний 100%-ийн алдаа байна.

	Зөв бичгийн дүрмийн алдаа	
1	*"ь"-ээр төгссөн үгэнд дагавар, нөхцөл залгахад "ь" нь "и"-ээр солигдох, эс солигдох дүрэм*	яривал→ярьвал, шуугидаг→шуугьдаг
2	*Эгшиг зохицох ёсыг зөрчиж болох и, ий, ы, эй эгшгүүдийг дагаж орох эгшиг нь тухайн эгшгийн өмнөх эгшиггэй зохицох дүрэм*	төвөгтэйхэн→төвөгтэйхөн, төвийнхэнд→төвийнхөнд, нөхөдтэйгээ →нөхөдтэйгөө, хөрөнгөтэйдээ→хөрөнгөтэйдөө, өвчтэйгээс→өвчтэйгөөс
3	*Үйл үгийн үндэс үүсгэх "–л" дагаварт шаардагдах балархай эгшгийг өмнө нь, хойно нь бичих дүрэм*	шарласан→шаралсан, цэгцлэж→цэгцэлж, хурдлаж→хурдалж, нийтлэнэ→нийтэлнэ, наслана→насална, наминчлана→наминчилна, дээдлэж→дээдэлж
4	*Үйл үгийн үндэст нөхцөлдүүлэн холбох "–ж, –ч", "–жээ, –чээ" нөхцөлийг залгах дүрэм*	өгөж→өгч, дарчээ→даржээ
5	*Үйл үгийн үндэс үүсгэх "–р" дагаврыг залгах дүрэм*	новшроно→новширно
6	*Балархай эгшиг гээгдэх дүрэм*	мэхэлдэгийг→мэхэлдгийг, мэдээлсэнээр→мэдээлснээр. биелүүхэд→биелүүлэхэд, хрэгжүүлж→хэрэгжүүлж, гавшайлан→гавшгайлан

- Код хольсон

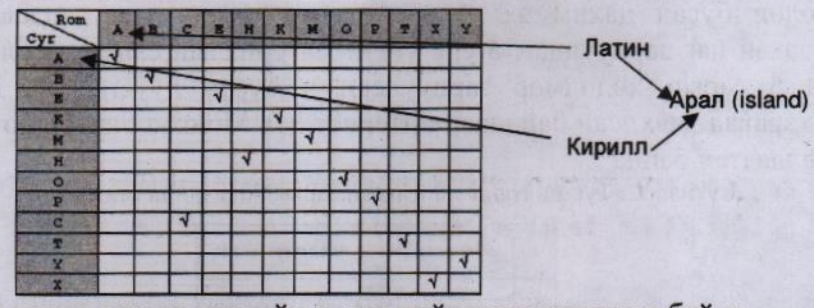
  ASCII, Unicode кодыг хольж бичсэн: бітээгдэхііінээр→бүтээгдэхүүнээр\

  Энэ нь ихэвчлэн тухайн бичвэрийг оруулж буй хэрэглэгчийн гарын драйвераас үүдсэн алдаа бөгөөд, юникодоор (ср1251-ээр) бичсэн бичвэрийг засахдаа ср1251 (юникод) драйвераар засах үед гардаг нилээн түгээмэл алдаа.

---

[3] Бичвэрийг цэвэрлэх гэдэг нь Веб дээр байгаа боловсруулаагүй HTML хуудас бүхий материалыг NLP алгоритм, программ буулгахтай холбоотой бүхий л процессыг хамардаг. Кодлолын асуудлыг нэг мөр, тухайлбал UTF 8 болгохоос гадна HTML/XML тэмдэглээг арилгах, бэлдэх явцад гарсан зөв бичих зүйн дүрмийн алдааг засах, тасалсан үгийг нийлүүлэх шаардлага гардаг.

- Хэлбэр ижил латин болон кирилл үсгүүдийг хольж хэрэглэсэн: Арал гэдэг үгийн эхний а үсэг нь латин, сүүлийн а үсэг нь кириллээр бичигдсэн.



- Эдгээрээс гадна дараах төрлийн алдаа нийтлэг тохиолдож байна.

№	Алдааны төрөл	Жишээ
	Дундаа зураастай үг	бай-на→байна, бай-галь→байгаль
	Бичлэг ижил үсэг, цифрийг хольсон	15ооо→15000, бөгөөд→бөгөөд
	Ойролцоо хэлбэртэй үсгийг сольсон	
	“э” үсгийг”з” үсгээр сольсон	тзмүүжин→тэмүүжин, үззгдлийг→үзэгдлийг, бүтэзлч→бүтээлч
	“ь”, ”ъ” тэмдгийг хооронд нь сольсон	тавъя→тавья, гарья→гаръя, өгъс→өгьс
	”ө”, “е” үсгийг хооронд нь сольсон	терүүлсэн→төрүүлсэн, өвгөн→өвгөн, хеөгөөд→хөөгөөд, еерсдөө→өөрсдөө, мерсөдес→мерседес
	“и”, “й” үсгийг хооронд нь сольсон	ажйллаж→ажиллаж, наим→найм, аимгийн→аймгийн, байгалиин→байгалийн, саихан→сайхан, хөгжилтэи→хөгжилтэй
	“ү”, “у” үсгийг хооронд нь сольсон	бурийн→бүрийн, дүурсэн→дүүрсэн, дугаар→дугаар
	“н”, “и” үсгийг хооронд нь сольсон	миийй→миний, гншүүн→гишүүн
	“т” үсгийг “г” үсгээр сольсон	тоггнолыг→тогтнолыг, эмэгтэйчүүд→эмэгтэйчүүд, тоггнолыг→тогтнолыг
	“с” үсгийг “е” үсгээр сольсон	буеад→бусад
	“ын” үсгийн нийлцийг (нийлэмж) “ьш” үсгийн нийлцээр сольсон	ардьш→ардын
	“ын” үсгийн нийлцийг “ьн” үсгийн нийлцээр сольсон	байгууллагьн→байгууллагын, улсьн→улсын
	“ч” үсгийг “н” үсгээр сольсон	надавхийг→чадавхийг
	“н” үсгийг “ң” үсгээр сольсон	сарың→сарын, бүрэлдэхүүң→бүрэлдэхүүн
	“н” үгийг “в” үсгээр сольсон	ахуйв→ахуйн
	“ц” үсгийг “д” үсгээр сольсон	духуйлгаж →цухуйлгаж

Дээрх алдааны сүүлийн 2 төрөлд сканердаж оруулсан веб бичвэрийн алдаа зонхилж байна. Ялангуяа үсэг солисон алдаан дээр бичвэр оруулагч алдаж бичиж болох хэдий ч илт боломжгүй тохиолдол байна. Жишээлбэл, *тэмүүжин* гэдгийг яаж алдаж бичсэн ч *тзмүүжин* гэж бичихгүй, алдлаа гэхэд э-гийн ойролцоо байрлах *ж*, *н* үсгийг сольж *тнмүүжин* эсвэл *тжмүүжин* гэх магадлалтай.

Дүгнэлт

Энэ ажилд Веб нь мэдлэгийн эх үүсвэр, хөмрөг болох талаар танилцуулж Монгол вебийн агуулгын өнөөгийн байдалд чанарын талаас нь шинжилгээ дүгнэлт өгч асуудал дэвшүүлэхийг зорилоо.

Судалгаанаас харахад Веб дэх монгол бичвэр алдаа ихтэй, кодийн хувьд нэгэн жигд бус байна. Эхийг үүсгэж байх явцдаа алдаагүй зөв оруулах нь хожмын судалгаа, хэрэглээнд зориулан цэвэрлэх ажилд шаардагдах хугацаа, хөдөлмөрийг хэмнэх сайн талтай.

Интернэтэд байрлах мэдээ, мэдээлэл, бусад зүйл нь Монгол улсын мэдээллийн өмчийн нэгээхэн хэсэг, орчин үеийн мэдлэгт суурилсан нийгмийн баялаг тул тэдгээрийг эхнээс нь алдаагүй зөв бэлдэх ёстой.

Ном зүй

Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, 29, 29(3):333–347.

Frank Keller and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.

Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the ACL*, pages 26–33, Toulouse, France, 9–11 July.

Whitten N (1995) Managing Software Development Projects - Formula for Success, John Wiley & Sons, NY, USA

Michael Mohler and Rada Mihalcea, BABYLON ParallelTextBuilder: Gathering Parallel Texts for Low-Density Languages,

SUMMARY

In this paper we aimed at introducing the web as a source of knowledge (such as lexical semantic network) and language resource (such as machine translation corpus) and at analyzing contents quality issues of Mongolian web.

As a result of the research, texts in the web have many orthographic errors and irregular code formats. When creating sources, inputting correct texts is very important thing that can reduce the time and effort to clean for further research and usage.

Any artifact such as news, information and publication etc. on the web are the intellectual properties of Mongolia and are resources of the knowledge based modern society. That is why from the very beginning they should be prepared and dealt with carefully and correctly.