Монгол Хэлний Үгзүйд Патрик Шоне ба Даниел Журавскийн Аргыг Ашиглах

Б.Батзолбоо*, Б. Гүндсамбуу**, Б. Удвалцэцэг***
Програмчлалын технологийн профессорын баг
ШУТИС, Компьютерийн Техник Менежментийн Сургууль
{*b.batzolboo, **gundsambuu.bold, ***udvaltsetseg}@csms.edu.mn

***Хураангуй*— Энэ судалгааны ажлын хүрээнд үгзүйн орчин үеийн хандлагуудыг судалж, тэдгээрийг монгол хэлний үгзүйд хэрэгжүүлэх боломжийг эрэлхийлсэн юм. Монгол хэлний бүтэцтэй төстэй гэж үздэг Турк хэлэнд туршигдсан байдлыг үндэслэн мөн Монгол хэлний үгзүйн онцлог, шинж чанаруудыг харгалзан удирдамжгүйгээр (unsupervised) сургах аргын гол төлөөлөл болох Патрик Шоне болон Даниел Журавский нарын аргыг сонгож, хэрэгжүүлэв. Судалгааны ажлын үр дүн нь боловсруулаагүй текст корпус (буюу хөмрөг)-ийг оруулахад үгийн үндсэнд залгах боломжтой залгаврууд олонлогийг гаргахад оршино.**

Түлхүүр үг*— *Хувирдаг үгзүй, удирдамжгүй сургах арга, залгаврууд олонлог, Патрик Шоне болон Даниел Журавскийн арга

I. Оршил

“Монголын ард түмний түүх, соёлын дурсгалт зүйл, шинжлэх ухаан, оюуны өв төрийн хамгаалалтад байна” [1] гэж үндсэн хуульд бий. Өвөг дээдсээс уламжлан ирсэн бидний хэл аялгуу маань Монголын төдийгүй нийт хүн төрөлхтний оюуны соёлын салшгүй нэг хэсэг, үнэт өв мөн билээ. Дэлхий дээр бие даасан, тусгаар улс орон, ард түмнүүд оршсоор байх цагт эх хэл бичгийн асуудал байсаар байх нь гарцаагүй зүйл билээ. Харин тухайн улсын болон хэлний том жижиг зэрэг хүчин зүйлсээс шалтгаалж өөр өөрийн онцлогт тохирсон хэлний бодлоготой байдаг байна.

Мэдээллийг автоматаар боловсруулах, хэрэглэхийн ач холбогдол нь мэдээллийн эрин зуунд шилжиж буй өнөө үед улам их өссөөр байна. Дэлхий нийтэд интернэт болон бусад цахим ертөнцөд нийтлэгдсэн мэдээллүүдийг өөрийн төрөлх хэл дээрээ хөрвүүлэх програм хангамжийг Англи, Орос, Герман, Франц, Хятад, Япон, Солонгос зэрэг улсуудад зохион хэрэглэж байна. Монгол улсын хувьд энэхүү автомат орчуулгын системд эх хэлээ оруулах нь зайлшгүй шаардлагатай зүйлүүдийн нэг билээ. Үүний тулд монгол хэлийг компьютерээр боловсруулах чиглэлд олон судалгаа хийх шаардлагатай бөгөөд одоогийн байдлаар манай улсын их дээд сургууль болон судалгаа шинжилгээний байгууллагуудад хийгдсэн ажлын түвшин нь автомат орчуулгын системийг бүтнээр нь хийхэд хүрэхгүй ч үгзүйн зарим түвшинд хангалттай хийгдсэн гэж үзэж байна [2].

Автоматаар үгзүйн шинжилгээ хийх анхны ажлууд 1960-аад оны үеэс машин орчуулгын системд хийгдэж байв [3]. Эдгээрийн дийлэнх нь хэлний мэдээллийг хатуу кодоор оруулан бичиж, хэтэрхий энгийн түр аргацаасан аргуудыг ашигласан байдаг. Сүүлийн 10 гаруй жилд судлаачид удирдамжгүй сургах аргаас багахан хэмжээний удирдамжтай (supervised) үгзүйн статистик загвар бий болгох аргыг сонирхох болжээ. 1990-ээд оноос өмнө үгзүйн зөв бичих дүрмийг хүний оролцоотой гар аргаар оруулж байсан боловч 2000 оноос хойш машин сургах аргаар зөв бичгийн дүрмийг цуглуулах оролдлогууд хийгдэж байв.

Алтай язгуурын хэлүүд болох Турк, Монгол зэрэг залгамал хэлүүдэд үг харьцангуй олон бүтээврийн дарааллаас бүтдэг. Иймээс удирдамжтай статистик аргыг хувирдаг үгзүйд хэрэглэхэд хэцүү байдаг [4]. Иймээс энэ ажлын зорилго нь үгзүйн шинжилгээнд удирдамжгүй сургах аргуудыг Монгол хэлэнд хэрэгжүүлэх боломжийг судлах юм.

II. Машин сургах арга

Аливаа асуудлыг компьютерээр шийдэхийн тулд алгоритм шаардлагатай байдаг. Алгоритм бол оролтоос гаралт руу хувиргах үйлдлийг гүйцэтгэх алхамуудын дараалал юм.

Машин сургах арга нь жишээ өгөгдлүүд болон хуримтлуулсан мэдлэгтээ тулгуурлан гүйцэтгэлийн шалгуурыг оновчлох компьютерийн тооцоолол юм [5]. Машин сургах аргад хэд хэдэн параметрээр тодорхойлогдох загвар байх бөгөөд загвар нь ирээдүйн таамаглалуудыг тодорхойлохын тулд урьдчилсан эсвэл дүрсэлсэн шинжтэй байж болдог.

Удирдамжтай сургах аргын гол зорилго бол зөв үр дүнг агуулсан гаралтыг гаргахын тулд оролтын өгөгдлөөс зааварлагчийн тусламжтайгаар сурах явдал юм. Удирдамжгүй сургах аргад зааварлагч шаардлагагүй, зөвхөн оролтын өгөгдөл хэрэгтэй байдаг бөгөөд гол зорилго нь оролтын өгөгдлөөс зүй тогтлыг олох явдал юм. Энэхүү ажлын хүрээнд Жон Голдсмит, Патрик Шоне болон Даниел Журавскийн, удирдамжгүй сургах аргуудыг авч үзлээ.

III. Компьютерийн хэл шинжлэл дэх үгзүй

Хүн төрөлхтний хэл нь тодорхой тооны нэгж хэсгүүдээс бүрдэх бөгөөд тэдгээр нь нарийн зүй тогтлоор холбоотой бүхэл бүтэн дохионы тогтолцоо юм. Ийм нарийн тогтолцоотой хэлний нэгжүүд нь нэг нь нөгөөтэйгөө эсрэгцэхийн зэрэгцээ бие биеэ нөхөж, давхарлан шаталж тогтсон байдаг. Өөрөөр хэлбэл, дээд түвшний нэгж нь доод түвшний нэгжээ багтаасан, доод түвшний нэгж нь дээд түвшний нэгждээ багтсан шинжтэй байна [6]. Хэлний нэгжүүдийн гол цөм нь үг бөгөөд хэлний аль ч салбарт үгийг олон талаас нь өөр өөр зорилгоор судалдаг.

XX зууны хагасаас эхлэн хэл шинжлэлийн ухаанд цахим хэл шинжлэл гэх шинэ салбар үүсэн бий болжээ. Компьютерийн хэл шинжлэл нь эх хэлийг компьютерээр хэрхэн боловсруулах арга техникийг судалдаг. Үг бол эх хэлний хамгийн бага нэгж юм [7]. Иймээс цахим үгзүй нь эх хэлийг компьютерээр боловсруулах эхний ажлуудын нэг билээ. Үгзүй нь үгсийн дотоод бүтцийг судалдаг ба үг нь бүтээврүүдээс тогтоно. Бүтээвэр гэдэг нь үгийн бүтцийн цааш үл задрах хамгийн бага утгат нэгжийг хэлнэ [8]. Үгийн язгуур, үндэс, залгавар нь бүтээврийн төрөл болно. Үгийн язгуур гэдэг нь үгийн гол цөм утга агуулдаг бөгөөд цаашид үл задрах бүтээвэр юм. Харин үгийн үндэс нь залгавар бүтээвэр авч хувилж болдог бүтээвэр юм. Залгавар бүтээвэр нь язгуур болон үндэс бүтээвэрт залгагдаж шинэ үг бүтээдэг [8].

IV. Жон Голдсмитийн арга

Жон Голдсмитийн хамгийн бага илэрхийлэх урт (minimum description length) арга нь дагавар залгах хувилбаруудыг сурдаг арга юм [9]. Энэ аргаар зохиосон систем нэг хэлний боловсруулаагүй текст корпусыг оруулахад, үгсэд тохирох сигнатуруудын олонлогийг гаргаж өгдөг. Сигнатур гэдэг нь энгийнээр өгөгдсөн үндсүүдэд залгах залгаврууудын олонлог юм. Жишээлбэл, Жон Голдсмитийн шинжлэгчээр англи хэлний корпусыг оруулахад blow, bomb, broadcast, drink, dwell, farm, feel гэсэн үндсүүд бүгд Ø, er, ing, s дагавар авсан тул NULL.er.ing.s (Жон Голдсмитийн тэмдэглэгээ) гэсэн сигнатур үүсчээ.

Энэ арга нь боловсруулаагүй корпусаас сигнатуруудыг ялгахдаа хоёр алхамаар гүйцэтгэдэг. Эхний алхам нэр дэвшигч сигнатурууд болон сигнатурын бүрэлдэхүүнийг тодруулах. Хоёр дахь алхамаар нэр дэвшигчдийг үнэлдэг.

Жон Голдсмит өөрийн аргаа Англи, Франц, Испани, Итали, Латин хэлүүдийн корпус дээр шалгаж хэл бүрийн хувьд хамгийн сайн үнэлгээтэй 10 сигнатурыг жагсаажээ. Англи хэлний хувьд Брауны корпусын эхний 500,000 үгээс бүрдсэн, Франц хэлний хувьд 350,000 үгээс бүрдсэн, Испани хэлний хувьд 124,716 үгээс бүрдсэн, Латин хэлний хувьд 125,000 үгээс бүрдсэн, Итали хэлний хувьд 100,000 үгээс мөн 1,000,000 үгээс бүрдсэн корпусууд дээр шалгажээ.

ХҮСНЭГТ 1. Жон Голдсмитийн тооцоолсон Англи хэлний эхний 10 сигнатурууд

1	NULL.ed.ing.s	accent, afford, attempt
2	‘s.NULL.s	adolescent, amendment
3	NULL.ed.er.ing.s	attack, charm, flow
4	NULL.s	aberration, abstractionist
5	e.ed.es.ing	achiev, compris, describ
6	e.ed.er.es.ing	advertis, enforce, pac
7	NULL.ed.ing	applaud, bloom, cater
8	NULL.er.ing.s	blow, drink, feel
9	NULL.d.s	abbreviate, balance, costume
10	NULL.ed.s	acclaim, bogey, burden

Жон Голдсмит мөн Англи, Франц хэлний хувьд тухайн хэлний корпус тус бүрээс 1000 дараалсан үгийн олонлог сонгон авч үр дүнг үнэлсэн байна [10]. Ингэхдээ сайн, буруу задалсан, задраагүй, хиймэл задлалт гэсэн ангилалд оруулжээ. Үр дүнг дараах хүснэгтээр харуулав.

ХҮСНЭГТ 2. Англи, Франц хэлний үнэлгээ

Ангилал	Англи хэл		Франц хэл	
	Тоо	*Хувь*	*Тоо*	*хувь*
сайн	829	82,9	833	83.3
буруу задалсан	52	5.2	61	6.1
задраагүй	36	3.6	42	4.2
хиймэл задлалт	83	8.3	64	6.4

V. Патрик Шоне ба Даниел Журавскийн арга

Шоне болон Журавский нарын энэхүү арга нь боловсруулаагүй текст корпусаас утгазүй, үгийн зөв бичлэгийн дүрэм, өгүүлбэрзүйн мэдээллийг ашиглан үгзүйн загварчлалын шинжилгээг хийдэг бөгөөд тэдний 2000 оны хамтарсан ажлын үргэлжлэл юм. CELEX (1996 онд Англи, Голланд, Герман хэлнүүдийн үгсийн сангаар байгуулагдсан) корпусыг ашиглан Англи, Голланд, Герман хэлнүүдэд туршигдсан.

Шоне, Журавский нар уг ажилдаа үгзүйн загварчлалын зарим аргуудын асуудалтай талыг дурьдсан байдаг. Тухайлбал, Голдсмитынх шиг аргуудад зөвхөн зөв бичих дүрмийн талаас анхаарлаа хандуулсан байна гэж үзсэн бөгөөд жишээ дурьдвал утгазүйн мэдээллийг нэмж тооцоолохгүйгээр “ally” гэсэн үгийг “all+y” гэж боловсруулахад төвөгтэй байх юм. Голдсмитээс хойших үгзүйн загварчлалын аргуудад үсгийн дүрмийн өөрчлөлтөнд нөлөөлөх ажлууд хийгдээгүй бөгөөд [11] дан ганц утгазүйг дагнан боловсруулалт хийх нь хангалтгүй юм. Үүсмэл үгзүй нь утгазүйн хувьд гол утгаасаа холдсон байж болно.

Тэдний энэхүү алгоритм нь дөрвөн алхамд хуваагддаг. Үүнд:

1. Залгаврууд нэр дэвшигчдийг таамаглах;

2. Нэр дэвшигчдийн хослолуудыг боломжит үгзүйн хувилбарууд бүхий залгаврууд тодорхойлох;

3. Нэр дэвшигдэж байгаа залгаврууд үгс дэх эзлэмжээр нь утгазүйн векторыг тооцоолох;

4. Эдгээр үгстэй ижил утгазүйн вектор бүхий бодит үгзүйн хувилбаруудыг сонгох;

A. Залгаврyудад нэр дэвшигчдийг таамаглах

Энэ нь уг алгоритмын эхний алхамд боломжтой залгаврyудыг олох, ижил үгсийг тодорхойлох юм. Хэрэв w1, w2 гэсэн хоёр үг p-ижил бол:

i. w1-ийн эхний p тэмдэгтүүд w2-ийн эхний p тэмдэгтүүдтэй адил байх

ii. w1-ийн p+1 тэмдэгт w2-ийн тэмдэгттэй адил биш байх

Жишээлбэл, үз, үзлэг үгсийн хувьд эхний хоёр (2-ижил) тэмдэгт адил байна.

B. Боломжит үгзүйн хувилбарууд бүхий залгаврyудыг тодорхойлох

Залгаврyудад нэр дэвшигчдийг тодорхойлсны дараа нэр дэвшигч залгаврyудын хослолуудыг тодорхойлно. Үүнийг бид хослолын дүрэм (rules) гэе. Хоёр үг нь ижил үндэстэй бөгөөд ижил залгаврын дүрэмтэй хэлбэрийг “Боломжит үгзүйн хувилбаруудын хослол” (pair of potential morphological variants (PPMVs)) гэе. Өгөгдсөн дүрэм болох нийтлэг дүрмүүдтэй бүх PPMV-н цуглуулгаас дүрмийн олонлогийг (ruleset) тодорхойлно. Бидний алгоритм нь өгөгдлөөс нэр дэвшүүлж буй дүрэм бүрд дүрмийн олонлогуудыг тодорхойлсон жагсаалтыг бий болгодог.

C. Утгазүйн векторыг тооцоолох

1990 оноос хойш компьютерийн хэл шинжлэлийн судлаачид корпус дахь бичвэр болон үгүүдийн хоорондын утгазүйн холбоосуудын ач холбогдлыг олохын тулд хүний оролцоо шаардлагагүй байж болно гэдгийг нотолж, харуулах болсон [12]. Үүнийг M матрицад ганц утгат задралыг ашиглаж үр дүнг харуулсан юм. M матрицийн M (i, j) оролт бүр нь корпусын j бичвэр дэх i үгийн давтамжийг агуулна. Энэ аргачлал нь Далд утгазүйн шинжилгээ (LSA)-тэй хамааралтай бөгөөд Ландауэр (Landauer 1998); Мэннинг ба Шфитц (Manning and Schfitze 1999) бүтээлд тодорхой тайлбарлагджээ.

Далд утгазүйн шинжилгээ (Latent semantic analysis) нь аливаа бичиг баримт болон текстийн түлхүүр үгэнд нь суурилан өгөгдлийг олж авахын хажуугаар, өгөгдөл хоорондын утгын хамаарал ямар байгааг харуулсан статик үзүүлэлтэнд үндэслэн дүгнэн шинжилдэг.

Энэ шинжилгээ нь үг болон текст хоёрын хоорондын харилцаа хамаарлыг олох асуудлыг шийддэг [13]. Энэхүү арга нь текстэнд давхардаж орсон үгсээр матриц үүсгэдэг [14] [15]. Матриц үүсгэхдээ туслах болон чимэх үгс мөн цифр, тэмдэгтүүдийг матрицын өгөгдлөөс хасаж тооцдог. Далд утгазүйн шинжилгээ нь ойролцоолох загвар бөгөөд төсөөтэй обьектуудыг хамтад нь орон зайгаар нь бүлэглэдэг математикийн нэгэн арга юм.

Бидний судалгааны ажлын хувьд энэ шинжилгээний явцад нөхцөлт нөхцөлт матрицыг байгуулсан бөгөөд уг матрицдаа ганц утгат задралыг (singular value decomposition) ашиглав. Энэ аргыг ашигласнаар матрицын хэмжээ багасах юм. Тодорхойлогч нь тэгтэй тэнцүү матрицыг сингуляр матриц гэнэ [16].

D. Ижил утгазүйн вектор бүхий үгзүйн хувилбаруудыг сонго

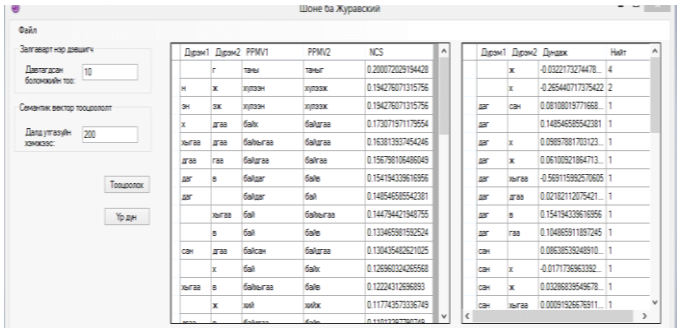
PPMV бүрийг үг бүрийн хоорондох хамааралыг тодорхойлно. Үүнийг тодорхойлохын тулд normalized cosine score (NCS) нэрлэгдэх аргыг ашиглах болно. NCS-г олохын тулд эхлээд утгазүйн вектор бүрийн хоорондох Ω_w косинусыг тооцоолно.

$$\cos(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|} . | \qquad (1)$$

Утгазүйн векторууд нь санамсаргүйгээр сонгогдсон 200 үгээс бүрдэнэ. Энэ утгыг олсноор бид w-ийн хамаарлын утга (μ_w) болон стандарт хазайлт (σ_w)-г олно. W_k үг бүр нь $k \in (1,2)$, векторын бусад 200 үгс санамсаргүйгээр сонгогдоно. W_k болон тооцоолсон 200 үг бүрийн хооронд косинусын утгууд болох утга(μ_k), вариац(σ_k) байна. W_1, W_2 -ийн косинус нь нормальчлагдсан бөгөөд NCS дараах байдлаар тооцоологдоно.

$$NCS(w_1, w_2) = min_{ke(1.2)} \frac{\cos(\Omega_{w1}, \Omega_{w2}) - \mu_k}{\sigma_k} \qquad (2)$$

Санамсаргүй NCS таамаглахдаа $N(\mu_T, \sigma_T^2)$ зөв корреляцийн вариацууд, тархалтын утга адил байх N(0,1) хэвийн тархалт хийнэ. Сэлгэлтүүдийн хослолын хоорондох дотоод өгүүлбэр зүйн орчны төсөөтэй талууд нь бас тооцоологдоно. Тухайн сэдвийн үгсийн анхны тооцоолол нь бага давталттай, их давталттай ч бай аль аль нь санамсаргүйгээр тооцоологддог. Үүнийг боломжит үг зүйн вариацуудын хослол буюу PPMV гэх ба үгсэд зориулсан сэлгэлтийн тал бүрийн гишүүд юм.



Зураг 1. Нормальчилсан косинус

ХҮСНЭГТ 3. ПАТРИК ШОНЕ БА ДАНИЕЛ ЖУРАВСКИЙН АРГУУДЫН ХАРЬЦУУЛСАН ҮР ДҮН

Ангилал	Англи хэл	Герман хэл	Турк хэл	Монгол хэл
	Тоо	*Хувь*	*Тоо*	*хувь*

Шоне ба Журавский	85,2	88,3	58,6	64,17

VI. Дүгнэлт

Энэхүү судалгааны ажлын хүрээнд удирдамжгүй сургах үгзүйн аргуудын төлөөлөл болох Патрик Шоне ба Даниел Журавскийн аргыг судалж, хэрэгжүүлэв. Ингэснээр дараах дүгнэлтэнд хүрэв. Үүнд:

1. Дээр өгүүлсэн арга нь удирдамжгүй үгзүйн дүрмийн илрүүлэлтийн эхний оролдлого биш боловч олон ажилд иш татагдсан байдаг. Шинэ гарч буй аргуудыг үнэлэх нэгэн төрлийн стандарт болж чаджээ.
2. Монгол хэлтэй төстэй залгамал хэлнүүдэд туршигдсан. Нөгөөтэйгүүр, туршигдаж буй хэлнүүдийнхээ хувьд бусад аргуудтай харьцуулбал харьцангуй өндөр үр дүнг үзүүлдэг.
3. Залгамал хэлнүүдийн онцлог болох хувирдаг үгзүйн асуудалд статик үгзүйн загварчлалын аргууд тохиромжгүй байдаг. Тиймээс дээрх аргуудыг залгамал хэлнүүдийн хувьд туршиж үзэх нь илүү үр дүнг үзүүлэх боломжтой.
4. Патрик Шоне ба Даниел Журавскийн алгоритм нь хувирдаг хэлнүүдэд төвлөрч байгаа боловч алгоритмд хүний зүгээс мэдээлэл оруулахгүй, зөвхөн утгазүйн нөлөөллөөр нь тооцоолж байгаа учраас үгийн хуваалтын хувьд онцгой тохиолдол гарч байгаа.
5. Патрик Шоне, Даниел Журавскийн арга нь дараах үндсэн дөрвөн алхамаас бүрдэж байна.
    a. Боломжит залгаврууды тодорхойлох;
    b. Боломжит үгзүйн хувилбарууд бүхий үгийн хослолуудыг олох;
    c. Үгсэд зориулж утгазүйн хувилбаруудыг боловсруулах;
    d. Ижил утгазүйн вектор бүхий хувилбаруудыг сонгох;
6. Уг алгоритмд далд утгазүйн шинжилгээ, ганц утгат задралын аргуудыг давхар ашигласан нь эдгээр аргуудын монгол хэлний үгзүйн хувьд анхны туршилтууд болж байна.
7. Ижил үгзүйн бүтэцтэй үгс ижил утгыг хуваалцаж байдаг. Тиймээс уг алгоритмын үр дүнд PPMV буюу боломжит үгзүйн хувилбаруудны хослолын векторуудын хамааралтай байдалд анхаарлаа хандуулав.
8. Судалгааны ажлын дагуу хийсэн программ 352 үгтэй текст файлыг оруулахад гаралтын үр дүн 64.17%-тай зөв гарч байна. Тус текст файлаас программын үр дүнд давхардсан тоогоор 23 залгавар гарах байсан. Гэтэл программын үр дүнд 16 залгавар гарсан бөгөөд 14 нь зөв үлдсэн 2 нь зөрүүтэй байлаа. Энэ үр дүнд программын зүгээс оруулж өгөх залгаварт нэр дэвших боломжийн тоо, матрицын хэмжээс нөлөөлсөн болно.

Ном зүй

[1] Монгол улсын үндсэн хууль, Улаанбаатар, 1992.

[2] А.Хүдэр, Өгүүлбэр дэх үгийн аймгийг марковын гинжээр тодорхойлох компьютерийн загвар боловсруулах нь, Улаанбаатар, 2013.

[3] R, Roack; R, Sproat, Computational Approaches to Morphology and Syntax, New York: Oxford University Press Inc, 2007.

[4] Б. Батзолбоо, “Кирилл, Монгол бичгийг хөрвүүлэх ба Монгол хэлний үг хувиллыг загварчлах судалгаа,” Докторын зэрэг горилсон нэг сэдэвт бүтээл, pp. 25-29, 2012.

[5] E. Alpaydin, Introduction to Machine Learning, London: The MIT Press, 2009.

[6] Ү. Ариунболд, Д. Төмөртогоо, Г. Цэцэгдарь, Ж. Санжаа ба Ц. Өнөрбаян, Орчин цагийн монгол хэл, Улаанбаатар, 2004

[7] S.R.Spiegler, Machine Learning for the Analysis of Morphologically Complex Languages, United Kingdom: University of Bristol, 2011.

[8] Ц.Өнөрбаян, Орчин цагийн Монгол хэлний үгзүй, Улаанбаатар, 2004.

[9] R, Roack; R, Sproat, Computational Approaches to Morphology and Syntax, New York: Oxford University Press Inc, 2008.

[10] J. Goldsmith, Unsupervised Learning of the Morphology of a Natural Language, Chicago: MIT Press, 2001.

[11] R. Brian ба S. Richard, Computational Approaches to Morphology and Syntax, New York: Oxford University Press, 2007.

[12] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer ба R. Harshman, Indexing by latent semantic analysis, Journal of the American Society for Information Science, 1990.

[13] T. K. Landauer, D. S. McNamara, S. Dennis ба W. K. (eds.), Handbook of Latent Semantic Analysis, Lawrence Erlbaum,, 2007.

[14] Association for Computational Linguistics, Advantages in WSD, 2005.

[15] B. Chen, Latent Semantic Analysis (LSA).

[16] Trucco, Singular Value Decomposition (SVD), Appendix A.6.