Монгол хэлний нэрийн бүлгийн задлуурыг модны сангаас үүсгэх нь

Чагнаагийн Алтангэрэл[1]*, Дагвасүмбэрэлийн Энхжаргал[2], Базаржавын Пүрэвсүрэн[2], Мөнхжаргалын Золжаргал[1], Чүлтэмсүрэнгийн Баярцацрал[1], Нямдаваагийн Оюундарь[1]

[1] Монгол Улсын Их Сургууль, Хэрэглээний шинжлэх ухаан, инженерчлэлийн сургууль, Мэдээлэл, компьютерийн ухааны тэнхим
[2] Монгол Улсын Их Сургууль, Шинжлэх ухааны сургууль, Европ судлалын тэнхим

*altangerel@num.edu.mn
Хүлээн авсан: 2018.03.30, засварласан: 2018.05.28, зөвшөөрсөн: 2018.06.01

Хураангуй

Өгүүлбэрийн бүлгийн задлал нь өгүүлбэрийн гишүүдийг дэд бүлэгт (нэрийн, үйлийн болон бусад) хуваадаг үйлийг хэлдэг. Үүнийг хийдэг хэрэгсэл нь эх хэл боловсруулалтын үндсэн суурь хэрэгслүүдийн нэг бөгөөд үүний дотор нэрийн бүлгийн задлуур нь бичвэрээс нэр томьёо, мэдээлэл, мэдлэг гарган авах болон машин орчуулга зэрэгт хэрэглээ, ач холбогдол ихтэй, чухал хэрэгсэл юм. Энэхүү ажлаар тэмдэглэгээт модны сангаас нэрийн бүлгийг ялгаж нэрийн бүлгийн хөмрөг үүсгэх, түүн дээр сургасан нэрийн бүлгийн таниур загварыг үнэлэх, улмаар нэрийн бүлгийн сайн чанартай хөмрөг үүсгэх боломжийг судаллаа. Модны сангаас автоматаар үүсгэсэн нэрийн бүлгийн хөмрөгт тулгуурлан 89% нарийвчлалтай ажиллах нэрийн бүлгийн задлуурын загварыг гаргаж авлаа.

Түлхүүр үг: Хэлний нөөц, модны сан, нэрийн бүлгийн задлуур

1. Удиртгал

Эх хэл боловсруулалтын систем нь бичвэрийг ямар хэлээр бичигдсэнийг таних хэл илрүүлэгчээс эхлээд олон салаа утгатай үгийн тухайн өгүүлбэрт хэрэглэгдэж буй утгыг тэмдэглэх салаа утга таниур хүртэл хэл боловсруулалтын янз бүрийн түвшний олон хэрэгслийг агуулдаг. Системийн эдгээр бүрдэл хэрэгслүүдийг цогцоор хөгжүүлэх нь аливаа хэлний хувьд чухал ач холбогдолтой ажил юм. Ялангуяа энэ нь Монгол гэх мэтийн залгамал бөгөөд нарийн бүтэц зүй тогтолтой хэлний хувьд томоохон сорил байдаг.

Монгол хэлний хувьд Англи гэх мэтийн бусад хэлтэй харьцуулахад хэл боловсруулалтын хувьд маш бага судлагдсан, цахим хэлбэрт оруулан үүсгэсэн бичвэр, толь бичиг, гэх мэт хэлний нөөц багатай хэл юм. Монгол улсад кирилл бичигт зориулсан хэл боловсруулалтын судалгааны ажлууд голчлон хийгдэж байна (J.Purev and Ch.Altangerel, 2004). Харин монгол бичигт зориулсан үгзүйн задлуур, өгүүлбэрзүйн задлуур, нэрлэсэн нэгж таниур зэрэг судалгааг Өвөрмонголын Их Сургуульд эрчимтэй хийж байна (S.Sarula 2014, S. Loglo 2013).

Хэлний бүтэц зүй тогтлыг цахимд томьёолох судалгаанд хэрэглэгддэг дүрмийн болон эмпирик гэсэн үндсэн хоёр үзэл баримтлал байдаг бөгөөд эмпирик буюу өгөгдөлд суурилсан арга нь дүрэмд суурилсан аргаас илүү их хэрэглэгдэж орчин үеийн төлөөлөгч нь болж байна. Учир нь гэвэл хэлний тэрхүү нарийн зүй тогтол бүрийг дүрмээр нэг бүрчлэн тогтоож өгөх нь маш их хөрөнгө, хүч шаардсан ажил байдаг байна. Харин өгөгдөлд суурилсан арга нь хэлний цахим баримт, бүртгэл дээр суурилсан, түүнээс зүй тогтлыг автоматаар гарган ашигладаг харьцангуй зардал багатай илүү динамик шинжтэй юм. Өгөгдөлд суурилсан аргаар эдгээр системийг хөгжүүлэхийн тулд тухайн хэлний нөөц болох хэлний төрөл бүрийн корпус чухал үүрэгтэй. Үүний дотор өгүүлбэрийн бүтцийг тэмдэглэсэн корпус болох модны сан тун чухал нөөц бөгөөд Монгол хэлтэй ойролцоо Түрк, Мажар гэх мэт хэлэнд модны сангаас нэрийн бүлгийн хөмрөгийг үүсгэн хэрэглэх талаар ажлууд хийгдэж байсан (Yıldız нар 2015, Recski, G. 2014).

Энэхүү судалгааны ажлын хүрээнд Монгол хэлний өгүүлбэрзүйн задлуур, тэр дотроо нэрийн бүлгийг таних хэрэгсэл, түүнд хэрэглэгдэх хэлний гол нөөц болох модны санг шинжилж

түүнээс үүсгэсэн нэрийн бүлгийн хөмрөгийг модны сангийн тэмдэглэгээний чанарыг үнэлэхэд хэрэглэж болох таамаг дэвшүүлэн туршилт хийлээ.

2. Материал, Арга зүй

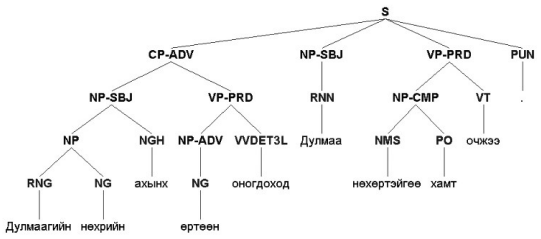
Өгөгдөлд суурилсан өгүүлбэрийн задлуур программыг сургах үндсэн сан болох Модны сан, түүний (Dependency treebank болон Structural treebank) төрлийн талаар мөн монгол хэлний өгүүлбэрийн онцлогийг судалж, Penn Treebank загвараар модны санг Монгол хэл дээр үүсгэхэд баримтлах мөрдлөгөөг (annotation guideline) боловсруулав. Үгийн аймгийн тэмдэглэгээг илүү нарийвчлан тодорхойлж, өгүүлбэрзүйн түвшний тэмдэглэгээг үүргийн тэмдэглэгээний хамт тодорхойлов.

Хүснэгт 1. Гараар тэмдэглэсэн модны сангийн хэмжээ

Бичвэрийн төрөл	Файлын тоо	Өгүүлбэрийн тоо	Хэмжээ*
Уран зохиол	12	4,509	~42K мөр/ 31K
Сонин	9	2,208	~35K мөр/ 33K
Нийт	21	6,617	

*нийт мөрийн тоо / үгийн тоо

Энэхүү боловсруулсан мөрдлөгөө ашиглан үгийн аймгийн тэмдэглэгээт МУИС-ийн материалын сангаас 40 гаруй мянган үгтэй, 6,700 гаруй өгүүлбэрийг хоёр хэл шинжээч гараар тэмдэглэж модны санг үүсгэсэн. Хэл шинжээч бүр нэг нэг төрлийн бичвэрийг тэмдэглэсэн. Өөрөөр хэлбэл давхардан тэмдэглэсэн бичвэр байхгүй гэсэн үг. Энэ нь нэг талаар адил бичвэрийг давхар тэмдэглэж хөдөлмөр үрэхгүй ч нөгөө талаар тэмдэглэгчдийн хоорондын нийцлийг шууд бодох боломжгүй болгодог. Иймээс тэмдэглэгч бүрийн ажилласан бичвэр дээр бичвэрийн төрөл дамжсан туршилтаар нийцлийг үнэлэхийг зорилоо. Хүснэгт 1-д модны сангийн хэмжээг бичвэрийн төрөл бүрээр жагсаан үзүүллээ.



Зураг 1. Модны сан дахь тэмдэглэсэн өгүүлбэрийн жишээ. *Модлог хэлбэрээр.*

Өгүүлбэрийн задлуурын хамгийн энгийн хувилбар болох нэрийн бүлгийн задлуур нь оролтын өгүүлбэрээс өгүүлбэрийн модны гүн бүтцийг үүсгэхгүйгээр нэг түвшин дэх нэрийн бүлгийг ялган тэмдэглэдэг хэрэгсэл юм. Жишээ нь Зураг 1-д үзүүлсэн жишээ өгүүлбэрээс дараах дөрвөн нэрийн бүлэг тодорхойлж болно. Эдгээр нь:

[Дулмаагийн_RNG нөхрийн_NG] |
[Дулмаагийн_RNG нөхрийн_NG ахынх_NGH],
[өртөөн_NG],
[Дулмаа_RNN],
[нөхөртэйгөө_NMS хамт_PO] болно.

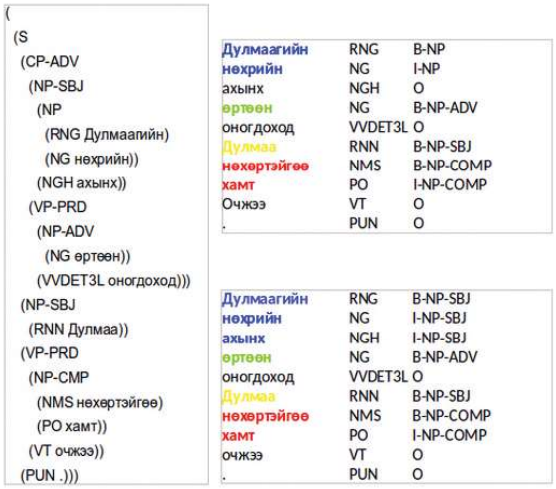
туршилтад үгийн аймгийн тэмдэглэгээг дангаар нь нэг онцлог болгон авах эсвэл дээд түвшний болон доод түвшний хоёр онцлог болгон хувааж авсан үед задлуурын гүйцэтгэлийг үнэлж үзлээ. Мөн нэрийн бүлгийн тэмдэглэгээг өгүүлбэрийн үүргээр нарийн ялган тэмдэглэх эсвэл нарийн ялгалгүй тэмдэглэх үед задлуурын гүйцэтгэлийг 10 нугалаат туршилтаар үнэллээ.

Модны сангаас автоматаар үүсгэсэн нэрийн бүлгийн хөмрөгөөр сургасан нэрийн бүлгийн таниур хэрэгслийн нарийвчлал нь тухайн модны сангийн тэмдэглээний чанар болон жигд байдлыг илэрхийлнэ. Мөн бичвэрийн төрөл бүр дээр сургасан таниурын гүйцэтгэлээр тэмдэглэгчдийн хоорондын санал нийцлийг харьцуулах болон бичвэрийн төрөл хоорондын онцлог ялгааг гаргаж болно гэж таамаглалаа.

Нэрийн бүлгийн задлуурын оновчтой загварыг гаргахын тулд нэрийн бүлгийн шатлал, үүргийн болон үгийн аймгийн тэмдэглэгээний олон хувилбараар туршиж харьцуулав.

Нэрийн бүлгийн шатлал

Модны сангаас нэрийн бүлгийг түүний шатлалаас хамаарч хоёр янзаар ялгаж болно. Үүний эхнийх нь хамгийн доод түвшний буюу модны 3 дугаар түвшин дэх нэрийн бүлгүүдийг ялгах.



Зураг 2. Нэрийн бүлгийг ялгах 2 өөр шатлалын жишээ. *3 дахь түвшин болонбүрэн түвшингээр.*

Энэ аргаар ялгавал *Дулмаагийн_RNG нөхрийн_NG ахынх_NGH* гэсэн нэрийн шаталсан тэмдэглээнээс зөвхөн *Дулмаагийн_RNG нөхрийн_NG* гэсэн нэрийн хэсэг гарах бөгөөд *ахынх_NGH* гэсэн хэсэг нь орхигдоно. Жишээг дараах зургаар үзүүлэв.

Нэрийн бүлгийг цогцоор нь авах удаах тохиолдолд *Дулмаагийн_RNG нөхрийн_NG ахынх_NGH* гэсэн нэрийн хэсэг ялгарна. Энэхүү цогц нэрийн нийлэмжээр авах тохиолдолд илүү урт буюу олон үгтэй нэрийн бүлэг үүснэ. Хүснэгт 2-т нэрийн бүлгийн 2 янзын шатлалд харгалзах статистикийг үзүүллээ.

Дээрх хүснэгтээс харахад энэхүү модны санд тэмдэг нэрийн үүрэгтэй нэрийн бүлгийн (NP-ADJ) тоо нь нэрийн бүлгийн шатлалыг өндөрсгөхөд огцом цөөрч байна. Энэ нь тэмдэг нэрийн үүргээр орсон энэ бүлэг бусад цогц нэрийн бүлгийн гишүүн болж орж байгааг харуулж байна. Мөн өгүүлэхүүний үүргээр орж байгаа бүлгийн (NP-SBJ) тоо ихсэж, түүний гишүүдийн дундаж тоо нь нэмэгдэж байна. Энэ нь өгүүлэхүүний бүлэг нь бас их шатлалтайгаар, нийлмэл байдлаар тэмдэглэгдсэн гэдгийг харуулж байна.

Хүснэгт 2. Нэрийн бүлгийн шатлал болон дундаж урт

NP төрөл*	3 түвшний NP тоо	3 түвшний Дундаж урт	Цогц түвшний NP	Цогц түвшний Дундаж урт
NP-ADJ	459	2.38	12	2.92
NP-ADV	1,315	2.34	1,257	2.45
NP-COMP	2,560	2.00	2,512	2.58
NP-CONJ	11	1.91	11	1.90
NP-OBJ	2,595	1.92	2,630	2.51
NP-PRD	208	2.28	313	3.78
NP-SBJ	4,822	1.91	5,870	2.59

NP-ADJ: тэмдэг нэрийн, NP-ADV: байцын, NP-COMP: шууд бус тусагдахуун,NP-OBJ: тусагдахуун, NP-PRD: өгүүлэгдэхүүн, NP-SBJ: өгүүлэхүүний үүргээр орсон нэрийн бүлэг тус тус болно.

Үүргийн болон үгийн аймгийн тэмдэглэгээ

Нэрийн бүлгийн тэмдэглэгээг түүний үүргийг оролцуулан илүү нарийвчлалтайгаар эсвэл үүргийн тэмдэглэгээг хэрэгсэхгүйгээр ерөнхийлөн хоёр янзаар тэмдэглэж болно. Онолын хувьд үүргийн тэмдэглэгээг агуулсан илүү нарийвчилсан нэрийн бүлгийн задлуурын ялгах ангиллын тоо их тул гүйцэтгэлийн нарийвчлал нь ерөнхийлсөн нэрийн бүлгийн задлууртай харьцуулахад бага байна.

Модны сангийн үгийн аймгийн тэмдэглэгээ нь 2 түвшинтэй тул үүнийг мөн нарийвчилсан болон ерөнхийлсөн байдлаар 2 янзаар туршиж харьцуулав. Энэ нь нэрийн бүлгийг тодорхойлоход үгийн хувирах нөхцөлийн оролцоо хэр нөлөөтэй байгааг харуулах юм.

Ашигласан хэрэгсэл

Ийнхүү Модны сангийн гүн бүтцийг нэрийн бүлгийн тэмдэглэгээ рүү хөрвүүлсэн санг сургалтын өгөгдөл болгон туршлаа. Нэрийн бүлгийг таних загвараар дарааллыг таамаглах нөхцөлт санамсаргүй талбар (J. Lafferty нар 2001) -ийг хэрэгжүүлсэн CRF++ хэрэгсэл ашиглав. Модны сангийн гүн бүтцээс нэрийн бүлгийг ялгах, сургалт болон тестийн өгөгдлийг хуваах, туршилтын үр дүнг харьцуулах линукс орчны скрийпт бичиж ашигласан.

3. Үр дүн

Дээрх аргаар бэлдсэн өгөгдөл дээр CRF хэрэгслээр эхлээд нэрийн бүлгийн үүргийн болон тэмплейтийн туршилт хийж дараа нь сургаж 10 нугалаат туршилтын үр дүнг гаргав.

3.1. Нэрийн бүлгийн үүргээр ялгах нь

Нэрийн бүлгийн үүргээр ялган сургах нь ерөнхий сургахаас ойролцоогоор 10 хувь харьцангуй бага нарийвчлалтай, F1 оноо ойролцоогоор 13 хувь бага байгааг Зураг 3-т дэлгэрэнгүй харуулсан байна.

Хүснэгт 3. Нэрийн бүлгийн задлуурын туршилтын үр дүн

Хувилбар*	Задлуурын Нарийвчлал (%)	Задлуурын F1 оноо
YAT=1, Үүрэг=-, Түвшин=3	85.44	68.40
YAT=2, Үүрэг=-, Түвшин=3	85.45	68.84
YAT=1, Үүрэг=-, Түвшин=3+	89.03	69.69
YAT=2, Үүрэг=-, Түвшин=3+	89.07	69.96

*YAT = (1: Үгийн аймгийн тэмдэглэгээг салгаагүй, 2: Үгийн аймгийн тэмдэглэгээг салгасан), Үүрэг = (+: Нэрийн бүлгийн үүргийг нарийвчилсан, -: Нэрийн бүлгийн үүргийг нарийвчлаагүй), Түвшин = (3: Нэрийн бүлгийн суурь түвшинд ялгасан, 3+: Нэрийн бүлгийг дээд түвшинд ялгасан)

Загвар 1 нь үгийн аймгийн ерөнхий тэмдэглэгээг ашигласан нэг дүрэмт загвар (unigram template) бөгөөд нийт 618,189 онцлогтой байв. Харин Загвар 2 нь үгийн аймгийн ерөнхий тэмдэглэгээг ашигласан хоёр дүрэмт загвар (bigram template) бөгөөд нийт 1,730,487 онцлогтой байв. Загвар 1 нь сайн үр дүнтэй байгаа тул дараах туршилтуудад ашиглалаа.

3.2. 10 нугалаат туршилт

Туршилтыг үгийн аймгийн тэмдэглэгээний 2 янз (1: үгийн аймгийн тэмдэглэгээг салгаагүй, 0: үгийн аймгийн тэмдэглэгээг салгасан), нэрийн бүлгийн түвшний 2 янз (3: нэрийн бүлгийн шатлал 3 дугаар түвшнээр ялгасан, 3+: нэрийн бүлгийн цогц түвшнээр ялгасан), нэрийн бүлгийн үүргийн тэмдэглэгээний 1 янз (-: үүргийг ялгаагүй) байдлаар хослуулан туршив.

Нэрийн бүлгийн үүргийг нарийвчлан авч сургах нь харьцангуй түвэгтэй байгаа нь өмнөх туршилтаас харагдсан тул үүргийг ерөнхийлөн дан NP-ээр авсан болно.

Туршилтын үр дүн Хүснэгт 3-аас харахад үгийн аймгийн тэмдэглэгээг салгасан, нэрийн бүлгийг цогц буюу дээд түвшинд ялган тэмдэглэсэн хувилбар дээр хамгийн сайн үр дүн үзүүлж байна.

3.3. Тэмдэглэгч болон бичвэрийн төрөл бүрээр

Туршилтыг мөн бичвэрийн төрөл бүр дээр болон төрөл дамнуулан туршив.

FineP-CourseC
template1

processed 5497 tokens with	found	1127	phrases	correct	794		
accuracy	87.17%	precision	71.02%	recall	69.38%	FB1	70.16
NP	precision	71.02%	recall	69.38%	FB1	70.16	

template2

processed 5497 tokens with	found	1127	phrases	correct	786		
accuracy	86.79%	precision	70.36%	recall	68.72%	FB1	69.5
NP	precision	70.36%	recall	68.72%	FB1	69.5	

FineP-fineC
template1

processed 5483 tokens with	found	1000	phrases	correct	601		
accuracy	77.90%	precision	61.02%	recall	53.20%	FB1	56.82
NP-CONJ	precision	0.00%	recall	0.00%	FB1	0	
NP-ADV	precision	50.70%	recall	31.61%	FB1	38.17	
NP-SBJ	precision	65.38%	recall	67.34%	FB1	66.3	
NP-OBJ	precision	73.35%	recall	73.80%	FB1	73.55	
NP-ADJ	precision	0.00%	recall	0.00%	FB1	0	
NP-SBJ-1	precision	20.00%	recall	1.15%	FB1	2.18	
NP-1	precision	0.00%	recall	0.00%	FB1	0	
NP	precision	29.48%	recall	19.48%	FB1	22.67	
NP-PRD	precision	41.81%	recall	9.72%	FB1	14.82	
NP-CMP	precision	49.20%	recall	38.21%	FB1	41.86	
NP-COMP	precision	30.38%	recall	26.35%	FB1	26.67	

template2

processed 5497 tokens with	found	991	phrases	correct	580		
accuracy	76.82%	precision	59.53%	recall	51.17%	FB1	55.01
NP-CONJ	precision	0.00%	recall	0.00%	FB1	0	
NP-ADV	precision	45.31%	recall	27.00%	FB1	33.33	
NP-SBJ	precision	63.15%	recall	65.49%	FB1	64.25	
NP-OBJ	precision	72.38%	recall	71.96%	FB1	72.14	
NP-ADJ	precision	0.00%	recall	0.00%	FB1	0	
NP-SBJ-1	precision	10.00%	recall	0.53%	FB1	1	
NP-1	precision	0.00%	recall	0.00%	FB1	0	
NP	precision	27.62%	recall	17.69%	FB1	20.57	
NP-PRD	precision	32.52%	recall	7.57%	FB1	11.97	
NP-CMP	precision	48.04%	recall	32.88%	FB1	38.2	
NP-COMP	precision	30.08%	recall	25.71%	FB1	26.72	

Зураг 3. Нэрийн бүлгийн үүргээр ялгасан туршилтын үр дүн

Хүснэгт 4. Бичвэрийн төрөл бүр дэх 10 нугалаат туршилтын үр дүн

Бичвэрийн төрөл*	Задлуурын Нарийвчлал (%)	Задлуурын F1 оноо
Уран зохиол	90.63	77.49
Сонин	81.82	62.46

Бичвэрийн төрөл дамнан туршихдаа нэг төрөл дээр сургасан загвараа нөгөө төрлийн сан дээр туршиж нарийвчлалын оноог бодсон болно.

Хүснэгт 5. Бичвэрийн төрөл дамнасан туршилтын үр дүн

Сургасан бичвэрийн төрөл	Шалгасан бичвэрийн төрөл	Задлуурын Нарийвчлал (%)	Задлуурын F1 оноо
Уран зохиол	Сонин	73.03	50.80
Сонин	Уран зохиол	83.14	59.77

4. Хэлэлцүүлэг

Модны сангаас нэрийн бүлгийг автоматаар ялгаж нэрийн бүлгийн хөмрөг үүсгэж түүнийг шинжих нь модны сангийн тэмдэглэгээний чанарын дам үнэлгээ юм. Өөрөөр хэлбэл сайн чанартай тэмдэглэсэн модны сангаас сайн нэрийн бүлгийн хөмрөг үүсгэж болно. Улмаар сайн нэрийн бүлгийн хөмрөг нь түүгээр сургасан нэрийн бүлгийн задлуур нь нарийвчлал сайтай байна. Энэ таамаглалаар үзвэл нийт модны сангийн чанар маань 89 хувийн нарийвчлалтай байгаа нь боломжийн үр дүн гэж үзэж байна.

Анхаарах асуудал нь F1 оноо нь 70 хувь орчим байгаа нь бас цаашид илүү сайжруулах хэрэгтэйг харуулж байна. Энэ оноо бага байгаагийн нэг шалтгаан нь тэмдэглэгч хэл шинжээчдийн хоорондын тэмдэглэгээний нийцэл (inter-annotator agreement) гэж үзэж байна. Энэ нь Хүснэгт 4 болон Хүснэгт 5-д үзүүлсэн туршилтын үр дүнгээс харагдаж байна. Тэмдэглэгч бүр харгалзах бичвэрийн төрлийг тэмдэглэсэн тул бичвэрийн төрөл бүрээр тусад нь хийсэн туршилт нь тухайн нэг тэмдэглэгчийн тэмдэглэгээний чанартай шууд холбоотой гэж үзэж болно. Энэ нь нэг хүн дангаараа хийвэл жигд чанартай хийж болно гэдгийг харуулж байна. Харин олон тэмдэглэгч оролцвол тэдгээрийн нийцэл буурч байгааг Хүснэгт 5-аас харж болно. Хэрвээ хоёр тэмдэглэгчийн нийцэл өндөр байвал бичвэрийн төрөл дамнасан туршилтын үр дүн харьцангуй өндөр гарах юм.

Удаах шалтгаан нь Монгол хэлний өгүүлбэрийн гишүүдийн онцлог мөн зарим эздрээтэй тохиолдлыг нарийн тодорхойлсон

тэмдэглэгээний мөрдлөгөө юм. Өөрөөр хэлбэл тэмдэглэгээний мөрдлөгөөг сайжруулж нэг мөр болгох шаардлагатай.

5. Дүгнэлт

Энэхүү судалгааны ажлаар Монгол өгүүлбэрийн модны сангаас нэрийн бүлгийг автоматаар ялгаж нэрийн бүлгийн хөмрөг үүсгэж чанарыг үнэлэв. Ингэхдээ нэрийн бүлгийн хөмрөгөөс төрөл бүрийн загвараар сургаж 89 хувийн нарийвчлалтай ажиллах нэрийн бүлгийн загварыг гаргалаа. Модны санг тэмдэглэгчдийн нийцлийг нэгдсэн тэмдэглэсэн бичвэргүйгээр, тэдний тэмдэглэсэн өөр өөр бичвэр дээрээс үнэлж дүгнэв.

Талархал

Бакалаврын судалгааны ажлынхаа хүрээнд модны гүн бүтцээс нэрийн бүлгийг автоматаар ялгах, үгийн аймгийн тэмдэглэгээг салгах скрийпт бичиж, туршилтад тусалсан Ч.Очиргарьдад талархлаа илэрхийлье.

Зохиогчийн оролцоо

Ч.А, Д.Э, М.З туршилтыг зохиомжилж, Д.Э болон Н.О тэмдэглэгээг хийж М.З сургалтын хэрэгсэл, скрийпт бичиж, Ч.А туршилтыг хийж Ч.А өгүүллийг бичив.

Санхүүжилт

Энэхүү судалгааны ажлыг МУИС-ийн залуу судлаачийн №P2016-1118 төслөөс санхүүжүүлэв. The research has received funding from the Mongolian Science and Technology Fund under grant agreement SSA_024/2016.

Ном зүй

1. P.Jaimai, T.Zundui, A.Chagnaa, and O.Che ol-Young, “PC-KIMMO-based De-scription of Mongolian Morphology,” Int. J. Inf. Process. Syst., vol. 1, no. 1, pp. 41–48, 2007.
2. Z.Munkhjargal and P.Jaimai, “Mongolian Trigram Part-of-Speech Tagger,” in The 7th International Conference on Multimedia Information Technology and Applica Tions (MITA 2011), 2011, no. July, pp. 6–9.
3. P.Jaimai and O.Chimeddorj, “Part of Speech Tagging for Mongolian Corpus,” in Proceedings of the 7th Workshop on Asian Language Resources, 2009, no. August, pp. 103–106.
4. Б.Нэргүй and О.Корфф, “Монгол хэлний хэл шинжлэлд хоёр түвшинт морфологийн аргын хэрэглээ,” 1996.
5. S.Sarula, “Construction of a Mongolian Dependen cy Treebank,” Int. J. Knowl. www, vol. 5, no. 2, pp. 32–42, 2014.
6. S.Loglo, “A Rule-Based Mongolian Dependency Parsing Model,” Int. J. Knowl. www, vol. 4, no. 3, pp. 27–37, 2013.
7. J.Purev and Ch. Altangerel, “An outline of Mongolian syntax.”, МУИС, МТС эрдэм шинжилгээний бичиг (I), 64-98, 2004
8. B.Batjargal, G.Khaltarkhuu, and A.Maeda, “An Approach to Named Entity Extraction from Mongolian Historical Documents,” 2015 Int. Conf. Cult. Comput. (Culture Comput., pp. 205–206, 2015.
9. B.Bataa and K.Altangerel, “Word sense disambiguation in Mongolian language,” in 2012 7th International Forum on Strategic Technology (IFOST), 2012, pp. 1–4.
10. Recski, G. (2014). Hungarian Noun Phrase Extraction Using Rule-based and Hybrid Methods. Acta Cybern., 21(3), 461-479.
11. H.Shen and A. Sarkar (2005). Voting between multiple data representations for text chunking. Proceedings of the Eighteenth Meeting of the Canadian Society for Computational Intelligence, Canadian AI 2005.
12. J.Lafferty, A. McCallum, and F.Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data, In Proc. of ICML, pp.282-289, 2001
13. Z.H.Huang, W. Xu, and K. Yu (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. In arXiv:1508.01991. 2015.
14. Yıldız O.T., Solak E., Ehsani R., Gцrgьn O. (2015) Chunking in Turkish with Conditional Random Fields. In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2015. Lecture Notes in Computer Science, vol 9041. Springer, Cham
15. Yıldız O.T., Solak E., Çandır Ş., Ehsani R., Görgün O. (2016) Constructing a Turkish Constituency Parse TreeBank. In: Abdelrahman O., Gelenbe E., Gor-bil G., Lent R. (eds) Information Sciences and Systems 2015. Lecture Notes in Electrical Engineering, vol 363. Springer, Cham
16. https://taku910.github.io/crfpp/

Noun Phrase Chunker Development from Mongolian Treebank

Altangerel Chagnaa[1]*, Enkhjargal Dagvasumberel[2], Purevsuren Bazarjav[2],
Zoljargal Munkhjargal[1], Bayartsatsral Chultemsuren[1] and Oyundari Nyamdavaa[1]

[1] Department of Information and Computer Science, School of Engineering and Applied Sciences, National University of Mongolia, Ikh surguuliin gudamj-3, Ulaanbaatar-14201
[2] Department of European Studies, School of Sciences, National University of Mongolia, Ikh surguuliin gudamj-3, Ulaanbaatar-14201

*altangerel@num.edu.mn
Received on 03.30.2018; revised on 05.28.2018; accepted on 06.01.2018

Abstract

Shallow sentence parsing or chunking is one of fundamental tools in natural language processing. It simply divides sentence tokens into noun, verb or other phrases. The most used one of thes is a noun phrase chunker which plays the key role in term extraction, information extraction and maa-chine translation etc. This paper introduces a noun phrase corpus extracted from manually created Mongolian constituency treebank, and evaluated this corpus by evaluating a model trained with this corpus. Automatically created noun phrase corpus gives a noun phrase chunker which has about 89% precision.

Key words: Language resource, Treebank, Noun phrase chunking