Монгол Хэлний Нэрлэсэн Нэгж Таниур

Машин сургалтын аргаар

Мөнхжаргалын Золжаргал*, Нямдаваагийн Оюундарь*, Чагнаагийн Алтангэрэл*, Габор Белла**
* Компьютер Хэл Шинжлэлийн Судалгааны Төв, Монгол Улсын Их Сургууль, Улаанбаатар, Монгол
** KnowDive групп, Трэнтогийн Их Сургууль, Трэнто, Итали
zoljargal@num.edu.mn, oyundari.m@gmail.com, altangerel@num.edu.mn, gabor.bella@unitn.it

***Хураангуй*—Нэрлэсэн нэгж гэж өгүүлбэрт байгаа хүн, байгууллага, орон байрын нэр, тоо, огноо, цаг, мөнгө зэрэг утгат хэсгүүдийг хэлдэг. Эдгээр утгат хэсгүүдийг автоматаар таних нь мэдээллийн оновчтой хайлтад чухал ач холбогдолтой юм. Судалгааны ажлын хүрээнд монгол хэлэнд анх удаа нэрлэсэн нэгж таниур хэрэгсэл програм үүсгэх, түүнд шаардлагатай монгол хэлний нэрлэсэн нэгжийн хөмрөг байгуулахыг зорьсон болно. Нэрлэсэн нэгж таниур нь англи, герман, франц зэрэг өндөр хөгжилтэй орны хэлнүүдэд хүний гараар тэмдэглэсэнтэй ижил төвшинд очихуйц автоматаар таньдаг. Харин монгол хэлний үг нь олон хувилалтай байдаг учир эдгээр хэлнүүдэд ашигласан аргыг шууд ашиглавал муу үр дүнд хүрэх нь ойлгомжтой. Тиймээс бид эдгээр хэлэнд ашиглаж байгаа арга тус бүрийг монгол хэлэнд туршиж, хооронд нь харьцуулж, генетик алгоритм ашиглан холихын дээр монгол хэлний онцлогт тохируулан өөрчлөхийг хичээлээ.**

Түлхүүр үг—Хэл Боловсруулалт (Natural Language Processing); Нэрлэсэн Нэгж (Named Entity); Нэрлэсэн Нэгж Танилт (Named Entity Recognition); Машин Сургалт (Machine Learning); Support Vector Machine; Conditional Random Field; Maximum Entropy; Genetic Algorithm;

I. Удиртгал

Бичвэрэн мэдээллийн хэмжээ өдөр ирэх тусам ихсэж хүний ойлгох болон боловсруулах чадамжаас давж гараад байна. Нөгөө талаас автоматаар мэдээлэл задлах (Information Extraction) технологи нь ийн хүний чадамжаас давсан мэдээлэл дундаас оновчлон сонгоход туслах зорилготой бөгөөд сүүлийн арваад жилд маш өндөр төвшинд хүрчээ. Бичвэрээс мэдээллийг ялгахад голчлон хүний нэр, байгууллагын нэр, газар ус буюу байрлалын нэр, мөнгө, цаг хугацаа илэрхийлсэн үгс дээр тулгуурладаг ба энэ нь ямар төрлийн мэдээлэл гаргаж авах вэ гэсэн системийн зорилгоос хамаарч олон төрөл байдаг. Ийнхүү өгүүлбэрт байгаа нэг бүхэл зүйлийг илэрхийлсэн нэгжийг таньдаг програмыг Нэрлэсэн Нэгж Таниур (ННТ) (Named Entity Recognizer) [2, 3, 13] гэнэ. ННТ нь англи, герман, франц зэрэг технологи өндөр хөгжсөн хэлнүүдэд хүн гараар тэмдэглэсэнтэй дүйхүйц ажилладаг бөгөөд монгол зэрэг нөөц багатай залгамал хэлнүүдэд энэ төрлийн програмыг хөгжүүлэх нь сонирхолтой судалгааны сэдэв юм [18].

ННТ-ийн аргуудыг 1) толинд суурилсан, 2) дүрэмд суурилсан, 3) стохастик машин сургалтын, 4) холимог гэж ангилдаг. Толинд суурилсан арга нь өмнө бэлдсэн үгийн сангаас хайлт хийх зарчмаар танилт хийдэг. Дүрэмд суурилсан арга нь зөв бичгийн дүрмийг (хүний нэрийг томоор эхлүүлэх г.м.) баримталж хайлт хийдэг. Уг хоёр аргыг монгол хэлэнд ашиглах нь жагсаалтад байхгүй үг, эсвэл дүрэм тодорхойлоогүй хэлц тааралдвал таньж чадахгүйн дээр залгамал хэлэнд тохиромжгүй юм. Харин машин сургалтад суурилсан арга нь хүний гараар тэмдэглэсэн материалын сангаас нэрлэсэн нэгжийн тодорхой онцлогт тулгуурлан сургалт хийгээд түүгээрээ тааварлах зарчмаар ажилладаг. Энэ арга нь одоогийн байдлаар дэлхийн олон хэлэнд хамгийн өндөр үр дүнг үзүүлдэг [2, 3] учир бид судалгааны ажилдаа энэ төрлийн аргуудыг ашиглалаа.

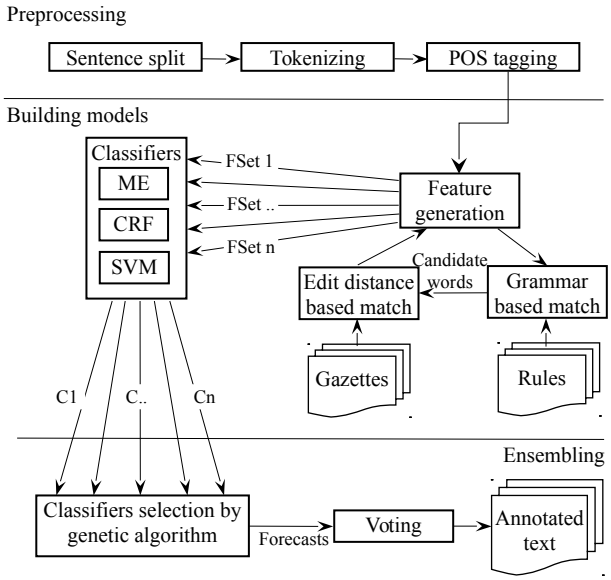
Машин сургалтын аргыг хэрэгжүүлсэн олон төрлийн алгоритм байдаг ба тэдгээрээс мөн хамгийн өндөр үр дүнтэй ажилладаг *Maximum Entropy (ME)* [20], *Support Vector Machine (SVM)* [25], *Conditional Random Field (CRF)* [26] -ыг ашигласан. Гэвч машин сургалтын арга нь сургалтын өгөгдөлд байгаа үгийн давтаж дээр тулгуурладаг учир эдгээр аргуудыг шууд ашиглах нь монгол зэрэг нэг үг хэдэн зуун хэлбэртэй байдаг залгамал хэлэнд мөн л муу үр дүн үзүүлэх нь ойлгомжтой юм. Тиймээс бид монгол хүний, газар усны, байгууллагын нэрийн толиноос өгүүлбэрт айгаа үгсийг ойролцоолон хайж [5, 6] машин сургалтын аргад нэмэлт мэдээлэл өгөх зарчмаар сайжруулахыг оролдлоо.

Нөгөөтэйгүүр өөр өөр онцлогийн олонлогт үндэслэсэн өөр өөр машин сургалтын алгоритмуудыг хольж ашиглах нь үр дүнг сайжруулдаг [19, 22, 23, 24,] учир бид мөн ажлын хүрээнд генетик алгоритмаар [21, 23, 24] хамгийн боломжит нийлэмжийг олохыг оролдсон.

Нэгэнт бид машин сургалтын аргыг ашигласан учир түүнд шаардагдах монгол хэлний анхны нэрлэсэн нэгжийн материалын санг үүсгэсэн [1].

II. ННТ Систем

Өөр өөр онцлогийн вектор тулгуурласан ялгаатай машин сургалтын аргууд нь асуудлыг өөр өнцгөөс харах ба эдгээрээс хамгийн өндөр үр дүнд хүрч байгаа аргуудыг олох, тэдгээрийг хооронд нь хольж хамгийн боломжит өндөр үр дүнд хүрэхийг бид оролдсон юм. Зураг 1-д туршилтын ерөнхий үйл ажиллагааг харуулав.



Зураг 1. ННТ модель үүсгэх ажлын урсгал. Fset- 5 онцлогийн векторын дэд олонлогууд, С- ангилагч.

III. Урьдчилсан боловсруулалт

Боловсруулалтын эхний алхам бол оролтын бичвэрийг өгүүлбэр өгүүлбэрээр салгана. Өгүүлбэрээр салгасны дараа өгүүлбэр дотор өгүүлбэрийн утгат хэсгүүдийг салгана (tokenization). Өгүүлбэр болон утгат хэсэгт тулгуурлан тухайн утгат хэсэг ямар үгийн аймагт хамааруулах байгааг тэмдэглэх бөгөөд эдгээр нь ННТ-ын онцлогийн вектор хэрэглэгдэх юм.

IV. Загвар үүсгэх

Машин сургалтын зарчимд тулгуурласан ангилагчийн алгоритмууд нь сургалтын өгөгдлөөс тодорхой онцлогийн вектор ялган авч түүн дээрээ магадлан тааварладаг. Онцлогийн вектор нь алгоритмын танилтын хувьд шууд нөлөөлдөг учир бид онцлогийн векторыг илэрхийлэх нэрлэсэн нэгжийн шинжийг [15] сонгохдоо нийтлэг хэрэглэгддэг шинжүүдээс нэг бүрчлэн туршиж өсөлт үзүүлж байгаа шинжүүдийг сонгосон [1]. Үүнд :

1. Зөв бичгийн: үг, томоор эхэлсэн эсэх, томоор бичсэн эсэх, жижгээр бичсэн эсэх, багадаа нэг дундуур зураас агуулсан эсэх, зөвхөн тэмдэгтээс тогтсон эсэх, тэмдэгт агуулсан эсэх.
2. Үгийн хэлбэр: урт үлгэр (long pattern: томоор бичсэн үсгийг “X”, жижигээр бичсэн үсгийг “x” болгоно. Ж. Баруун=Xxxxxx), богино үлгэр (ижил тэмдэгтийг давтахгүй. Ж. Баруун=Xx), тэмдэгт (symbol) агуулсан эсэх.
3. Угтвар, залгавар: үгийн эхний 4 үсэг, үсгийн 3-грам.
4. Хам мэдээлэл: үгийн аймаг, дээд болон доод төвшний үгийн аймаг [4], өгүүлбэр дэх байрлал (зөвхөн эхлэлд, дунд эсвэл төгсгөлд), хашилтан дотор бичигдсэн эсэх [16].
5. Толины мэдээлэл: Тухайн үг хүн, газар ус, байгууллагын нэрийн толинд байгаа эсэх (ойролцоолон хайх, V бүлэг).

Дээрх 5 бүлэг онцлогийн олонлогийн нөлөөг шалгахын тулд боломжит бүх дэд олонлогуудаар (2^5-1=31 загвар-model) ангилагч бүрийг сургасан. Хамгийн өндөр үзүүлэлттэй 15-аас 20 загвар ойролцоо үзүүлэлттэй байсан тул ялгаа ихтэй эхний 5 загварыг авч ангилагчуудыг холихдоо ашигласан (VI хэсгийг үз).

V. Ойролцоолон хайх алгоритмууд

Монгол хэлний үг нь өгүүлбэрт хувирсан хэлбэрээр тохиолдох нь элбэг учир толиноос уг хувирсан хэлбэрийг хайхдаа ойролцоолон хайх алгоритмуудыг туршсан. Эдгээрт *Levenshtein* [7], *Smith-Waterman* [8], *Jaro* [9], *Jaro-Winkler* [10], *Variation of Common Prefix* [6], *Fleggi-Shunter* [12], *Monge-Elkan* [11], *SoftTFIDF* [6] багтана.

Өгүүлбэрт байгаа бүх үгийг ойролцоолох алгоритмаар хайх нь компьютерт тооцоолол их шаардах учир монгол хэлний оноосон нэр бичих дүрмийн дагуу үлгэр (pattern) гаргаж зөвхөн оноосон нэр байж болох үгнүүдийг толиноос хайна. Оноосон нэрийн үлгэр нь:

Хүний нэр:

- *-ийн, -ын, -н төгсөгөлтэй томоор эхэлсэн үг + Томоор эхэлсэн үг. Ж.н: **Г**анболд**ын А**мар.*
- *Томоор эхэлсэн + цэг + томоор эхэлсэн үг. Ж.н: **Б**а. **Д**орж, **Ч.А**лтангэрэл .*

Байрлалын нэр:

- *Томоор эхэлсэн үг + байрлалын зүүлт (гудамж, талбай, хот, гол, улс). Ж.н: **Б**улган **гол**.*

Байгууллагын нэр:

- *Томоор эхэлсэн үгийн дараалал + байгууллагын зүүлт (ххк, холбоо, компани). Ж.н: Эм Си Эс ххк.*
- *Томоор эхэлсэн үг + жижгээр бичсэн үгийн жагсаалт + байгууллагын зүүлт. Ж.н: **М**онголын **з**алуучуудын **холбоо**.*
- *Бүгдийг томоор бичсэн үг. Ж.н: МУИС.*

VI. Холимог ангилагч

Бид туршилтандаа нээлттэй эхийн OpenNLP 1.5.3 (Maximum Entropy), CRF++ 0.53 (Conditional Random Field), YamCha 0.33 (Support Vector Machine) ангилагчийн платформуудыг ашигласан.

Олон ангилагчийг холихдоо хамгийн их санал нийлж байгаа тэмдэглэгээг тухайн нэрлэсэн нэгжид оноох энгийн заримыг ашигласан.

A. Генетик алгоритм

Ангилагч тус бүр дэх 31 загвараас хамгийн өндөр үр дүнтэй тав таван ангилагчийг сонгон авч нийт 3 ангилагчийн 15 загвар дээр холих үйлдлийг хийсэн.

Тэгэхдээ эдгээр 15 загвараас хамгийн боломжид хослолыг олохдоо бүх боломжит хувилбарыг туршиж үзвэл 2^15-1 ширхэг туршилт хийх шаардлагатай болно. Тиймээс бид Генетик алгоритм ашиглан энэ асуудлыг шийдэхийг зорьсон юм.

Бидний туршилтад хромосом нь 15 тэмдэгтээс (10101 01010 10101) тогтох хоёртын цуваа илэрхийлэгдэх бөгөөд эхний 5 бит нь *Maximum Entropy* -ын 5 загвар, дараах нь *Conditional Random Field* -ын, сүүлийн 5 нь *Support Vector Machine* -ийн загваруудыг төлөөлнө. Хэрэв бит 1 байвал тухайн ангилагч сонгогдсон, үгүй бол сонгогдоогүй гэсэн үг юм.

Бидний хэрэгжүүлсэн генетик алгоритмын алхам, параметр, аргачлалыг доор жагсаав:

1. Санамсаргүй битийн цуваатай 50 ширхэг хромосом буюу эхний үеийг үүсгэнэ *P(0)*.
2. Хромосом бүр дээр тохирлын функцийг (fitness function) ажиллуулна. Тохирлын функц нь хоёртын цуваагаар илэрхийлэгдсэн ангилагчуудыг холиод F1 [3] оноог бодно. Хамгийн их F1 оноотой хромосомыг хамгийн тохиромжит хромосом байна гэж үзнэ.
3. Тохиромжит оноогоор нь эрэмбэлээд түүнээсээ хос болох хромосомуудыг сонгоно.
4. Сонгосон бүх хосоос солбиж (crossover) дараагийн үе *P(1)* – ийг үүсгэнэ. *P(1)* дэх бүх хосыг хооронд нь элемент бүрээр нь солбиж шинэ хромосомуудыг үүсгэнэ.
5. Бүх шинээр үүссэн хромосомуудыг эргүүлэх (flip) аргаар мутацид оруулна.
6. 2-оос 5 дугаар алхмыг 80 удаа давтана.

Хромосомын тоо, үеийн тоог Ekbal нарын ажилтай [23] ижил авсан.

VII. Туршилт, Үр Дүн

Уг ажлын хүрээнд бид монгол хэлний нэрлэсэн нэгжийн тэмдэглэгээт хөмрөгийг үүсгэсэн бөгөөд [1] -ээс дэлгэрэнгүйг харна уу.

Тэмдэглэгээт материалын сангаас 10 хувийг нь санамсаргүйгээр сонгож туршилтын өгөгдөлд, үлдсэн 90 хувийг сургалтын өгөгдөлд ашигласан. Туршилтын оноог CoNNL2003 [3] хурлын аргазүйгээр хэмжсэн.

A. Ойролцоолон хайх алгоритмын туршилт

4,382 хувилсан хүний нэрийг 176,343 хувилаагүй нэртэй харьцуулж (харьцуулалтын хос 770 сая) алгоритм тус бүр дээр F1 оноог тооцсон (Хүснэгт 1). *Smith-Waterman* алгоритм хамгийн өндөр оноотой байна. Гэсэн ч уг алгоритм нь үгийн дунд орсон тэмдэгтүүдийг зөв гэж үзэж байсан. Жишээ "*Лувсаншараваас*" гэдэг үгийг "*шарав*" гэсэн үгтэй харьцуулбал ялгаа байхгүй гэж гаргана. Тиймээс нэгнийгээ агуулсан хосыг нэг гэж гаргах учир бодит системд муу үр дүн үзүүлнэ гэж үзэж байна.

Levenshtein алгоритм хамгийн муу оноотой байсан. Энэ нь энгийн үгийн алдаа шалгадаг алгоритм нь залгамал хэлэнд үгийг ойролцоолон хайхад тохиромжгүй гэдгийг харуулж байна. Харин *Jaro-Winkler* алгоритм нь үгийн зүүн хэсэгт байгаа ижил үсгүүдэд баруун хэсэгт байгаа ижил үсгүүдээс их жин оноож түүнийхээ дунджаар тооцоолдог учир үгийн үндэс ижил, нөхцөл дагавар нь ондоо үгнүүдэд илүү тохиромжтой юм. Тиймээс бид *Jaro-Winkler* алгоритмыг *Monge-Elkan* аргатай хамт сонгон ашиглахаар шийдсэн.

Алгоритм	Pre	Re	F1
Нэг үгийг ойролцоолох			
Levenshtein	73.4	71.9	72.6
SmithWaterman	85.1	97.5	90.8
Jaro	93.2	88.9	90.1
JaroWinkler	95.5	92.8	94.1
Common prefix	84.5	83.1	83.8
Олон үгийг ойролцоолох			
MongeElkan & SmithWaterman	82.4	94.7	88.1
MongeElkan & JaroWinkler	89.3	93.8	91.6
Jaccard & Direct matching	79.3	67.5	72.9
Felleggi-Shunter	75.6	70.1	72.7
SoftTFIDF & JaroWinkler	92.2	90.2	91.2
SoftTFIDF & Smith-Waterman	86.8	94.8	90.6

Хүснэгт 1. Ойролцоолох алгоритмуудын туршилтын үр дүн

B. Онцлогийн олонлог сонгох туршилт

Бид 3 ангилагч тус бүрд 31 загвар (нийт 93 загвар) үүсгэсэн бөгөөд Хүснэгт 2. -т хамгийн өндөр F1 оноо бүхий 15 загвар, тэдгээрийн онцлогийн олонлогийг үзүүлэв.

ME		CRF		SVM	
Онцлогийн бүлэг	**F1 оноо**	**Онцлогийн бүлэг**	**F1 оноо**	**Онцлогийн бүлэг**	**F1 оноо**
2, 4, 5	83.69	1, 2, 3, 4, 5	86.94	1, 2, 3, 4, 5	86.66
2, 4	83.04	1, 3, 4, 5	86.91	1, 2, 3, 4	86.57
1, 2, 4	82.92	1, 2, 3, 4	86.75	1, 3, 4	86.56
1, 2, 4, 5	82.82	2, 3, 4, 5	86.75	2, 3, 4	86.16
1, 2, 3, 4	82.81	2, 3, 4	86.62	1, 2, 4, 5	86.11

Хүснэгт 2. Ангилагч тус бүрийн хамгийн өндөр F1 оноо бүхий 5 загвар

Туршилтын үр дүнгээс хархад бүх онцлогийн бүлгийг ашигласан CRF нь хамгийн өндөр үр дүнтэй байна. Ер нь онцлогийн бүлгүүд CRF болон SVM аргуудад бүгд ашиглагдсан байна. Тэгэхдээ эдгээр онцлогийн бүлгүүдийг бүгдийг нь нэг дор ашиглах нь компьютероос их хэмжээний тооцоолол шаардана. Ялангуяа толины онцлог нь их хэмжээний тэмдэгтийн цуваг харьцуулдаг учир маш удаан ажилладаг. Түүнчлэн дээрх хүснэгтээс хархад толины онцлогийг ашиглаагүй CRF ангилагч нь ME болон SVM ангилагчийн хамгийн их онооноос илүү, CRF ангилагчийн хамгийн их онооноос зөвхөн 0.19-өөр бага

байна. Тиймээс толины онцлогийг ашиглахгүй байх чанарын хувьд багахан зөрөөтэй ч хугацааны хувьд давуу талтай нь харагдаж байна.

C. Холимог ангилагчийн үр дүн

Бид нэг ангилагчийн 5 загварыг хооронд нь хольж туршиж үзсэн (Хүснэгт 3). ME-ийн оноо 0.97-оор ганцаараа ажилласан онооноос буурсан байна. CRF-ийн оноо 0.42-оор өссөн бол SVM-ийнх 0.23-аар буурсан байна.

Ангилагч	Pre	Re	F1
ME-ийн холимог	88.86	77.38	82.72
CRF-ийн холимог	90.83	84.14	87.36
SVM-ийн холимог	88.19	84.75	86.43

Хүснэгт 3. Нэг ангилагчийн загваруудын холимог.

Генетик алгоритмын үр дүнд 00001 10011 11111 гэсэн хромосомыг олсон бөгөөд энд CRF болон SVM түлхүү сонгогдсон байна.

	Pre	Re	F1
Генетик алгоритмаар сонгосон	90.59	85.88	88.17
Нэг ангилагчийн холимог	90.83	84.14	87.36
Дан ангилагч	90.14	83.97	86.94
Бүх 15 загварын холимог	87.26	91.13	83.71

Хүснэгт 4. Нийт туршилтын үр дүн.

Туршилтын үр дүнгээс холимог ангилагч нь гүйцэтгэлийг өсгөж байгаа ч ямар нэгэн шүүлтүүргүйгээр шууд холивол үр дүнг бууруулж байгаа нь харагдаж байна. Түүнчлэн генетик алгоритм нь хамгийн зохимжит нийлэмжийг олоход ашиглаж болох нь нотлогдож байна.

Олон ангилагчийг холих нь үр дүнтэй ч танилтын хугацаа, тооцооллын хувьд дан ангилагчаас удаан байх нь ойлгомжтой. Бид энэ удаагийн ажлаар танилтын хугацааг чухалчилж үзээгүй бөгөөд дэлхий нийтэд ашиглагдаж байгаа арга технологиудыг боломжит хувилбаруудаар нь монгол хэлэнд туршиж үзэхийг зорьсон юм.

Цаашид сүүлийн жилүүдэд өргөн хэрэглэж байгаа *Newral Network*, *Deep Learning Method*-ийг турших ажлыг хийх шаардлагатай байна. ННТ нь монгол хэлэнд анх удаа хийгдэж байгаа ба ирээдүйд ННТ-ын бусад арга, сайжруулалтыг туршихад бидний энэ удаагийн судалгааны ажил суурь материал болно гэж найдаж байна.

VIII. Зааалт

[1] М.Золжаргал, Н.Оюундарь, Ч.Алтангэрэл. 2014. Монгол хэлний нэрлэсэн нэгжийн хөмрөг байгуулах нь. ММТ2014 хурлын хураангуй. Улаанбаатар, Монгол.

[2] Chinchor N. and Robinson P. 1997. MUC-7 named entity task definition, Proceedings of the 7th Message Understanding Conference.

[3] Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In Proceedings of CoNLL-2003, Canada.

[4] Purev J. and Odbayar C. 2009. Part of Speech Tagging for Mongolian Corpus. In Proceedings of the 7th Workshop on Asian Language Resources, Singapore.

[5] Cohen W. William, Ravikumar Pradeep and Fienberg E. Stephen. 2003. A Comparison of String Distance Metrics for Name-Matching Tasks, In proceedings of IJCAI-3 Workshop on Information Integration on the Web (IIWeb-03), pages 73-78, Acapulco, Mexico.

[6] Piskorski J., Wieloch K., Pikula M. and Sydow M. 2008. Towards Person Name Matching for Inflective Languages. NLPIX, Beijing, China.

[7] Levenshtein V. 1965. Binary Codes for Correcting Deletions, Insertions, and Reversals, Doklady Akademii Nauk SSSR, 163(4):845-848.

[8] Smith T. and Waterman M. 1981. Identification of Common Molecular Subsequences, Journal of Molecular Biology, 147:195-197.

[9] Jaro M. A. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. Journal of the American Statistical Association 84:414-420.

[10] Winkler W. E. 1999. The state of record linkage and current research problems. Technical report, Statistical Research Division, U.S. Bureau of the Census, Washington, DC.

[11] Monge A. and Elkan C. 1996. The Field Matching Problem: Algorithms and Applications. In Proceedings of Knowledge Discovery and Data Mining 1996, pages 267-270.

[12] Fleggi, I. P., and Sunter, A. B. 1969. A theory for record linkage. Journal of the American Statistical Society 64:1183-1210.

[13] David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. LinguisticaeInvestigationes, 30(1):3–26.

[14] Finkel J. R., Grenager T. and Manning C. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In Proceedings of the 43nd Annual Meeting of ACL.

[15] Eszter Simon and András Kornai. 2013. Approaches to Hungarian Named Entity Recognition. PhD thesis, Budapest University of Technology and Economics.

[16] Silviu Cucerzan and David Yarowsky. 1999. Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence. In Proceedings of the Joint SIGDAT.

[17] Szarvas G., Farkas R. and Kocsor A. 2006. A Multilingual Named Entity Recognition System Using Boosting C4.5 Decision Tree Learning Algorithms, Springer Berlin Heidelberg, page 267-278.

[18] Tur G. and Hakkani-Tur D. 2003. A statistical information extraction system for Turkish. Natural language engineering, 9(2):181-210.

[19] Kucuk D. and Yazici A. 2012. A hybrid named entity recognizer for Turkish. Expert systems with Applications, 39(2012):2733-2742.

[20] Bender O., Och F. J. and Ney H. 2003. Maximum Entropy Models for Named Entity Recognition. In Proceedings of CoNLL-2003, 148-151.

[21] Goldberg D. E. 1989. Genetic Algorithm in Search, Optimization, and Machine Learning. Addison-Wesley Publishing Company.

[22] Florian R., Ittycheriah A., Hongyan Jing and Tong Zhang. 2003. Named entity recognition through Classifier Combination. Proceedings of the seventh conference on Natural Language Learning at HLT-NAACL.

[23] Ekbal A and Saha S. 2010. Maximum entropy classifier ensembling using genetic algorithm for NER in Bengali. Proceedings of the International Conference on Language Resource and Evaluation (LERC).

[24] Desmet B. and Hoste V. 2010. Dutch Named Entity Recognition using Classifier Ensembles. Preceedings of the 20th Meeting of Computational Linguistics in the Netherlands.

[25] Isozaki H. and Kazawa H. 2002, Efficient support vector classifiers for named entity recognition, Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002), Taipei, Taiwan.

[26] Lafferty J., McCallum A. and Pereira F. 2001. Conditional Random Fields: Probabilistic Modelsfor Segmenting and Labeling Sequence Data. Machine Learning International Workshop.