Байесийн ангилагч ашиглан Монгол текстийг агуулгаар нь ангилах туршилт

Д.Золбоо, А.Хүдэр
Шинжлэх Ухаан Технологийн Их Сургууль, Компьютерийн Техник Менежментийн Сургууль,
Компьютерийн Ухааны Салбар
d.zolboo@csms.edu.mn, khuder@csms.edu.mn

***Хураангуй*— Текст ангилах үндсэн зорилго нь тухайн мэдээг тодорхой ангилалд хамааруулан ангилах, тодорхой баримт бичигт харгалзуулна. Энэхүү судалгааны ажилд Байесийн ангилагч ашиглан Монгол текстийг агуулгаар нь ангилах туршилт хийсэн. Бид 10 ангилал тус бүрээр сургалтын корпус үүсгэсэн. Өнөөг хүртэл Монгол текст дээр энэхүү аргыг туршсан судалгааны ажил хийгдэж байгаагүй тул шинэлэг сэдэв юм.**

Түлхүүр үгс — *Текстийн ангилал, Байесийн ангилагч, корпус*

I. УДИРТГАЛ

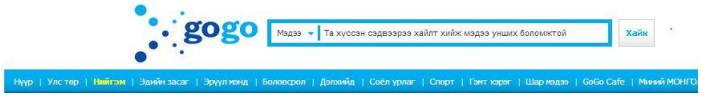
Мэдээллийн олон хэрэгслүүдийн нэг болох интернэтийн цахим хаягнаас өөрт хэрэгтэй мэдээ, мэдээллийг хайж олох нь сүүлийн үед хүндрэлтэй асуудлуудын нэг болсон.

Одоо ихэнх байгууллага, компаниуд болон хувь хүн өөрийн гэсэн цахим хаяг, цахим хуудастай болсон бөгөөд хүмүүсийн хэрэглээг интернэтгүйгээр төсөөлөхийн аргагүй. Түүнийхээ хэрээр мэдээллийн сан хязгаргүй ихээр өсөн нэмэгдэж байна. Эдгээр цахим хаягт оруулж буй мэдээллүүдийг ижил төрлөөр нь ангилан, нэгтгэхэд цаг хугацаа болон санхүү, хүн хүчний хувьд их ажил шаардах нь тодорхой.

Энэхүү их ажлыг хөнгөвчлөхийн тулд цахим хуудсанд шинээр оруулж буй мэдээ мэдээллийг ангилал тус бүрт нь корпус үүсгэн, сургаж, түүнийг тохирох ангилалд ангилах нь дэвшилтэт арга юм.

Мэдээний ангилалыг сонгохдоо Монгол улсын хэмжээнд ашиглагдаж буй мэдээний портал сайтуудыг харьцуулсан үзэж сонгосон. Тухайлбал:

http://news.gogo.mn/



http://www.news.mn

http://www.shuud.mn/

Эдгээр сайтуудад нийтлэг ашиглагдаж буй мэдээний дараах 10 ангилалыг сонгон авлаа. Үүнд:

1. Боловсрол
2. Гэмт хэрэг
3. Дэлхийд
4. Нийгэм
5. Соёл урлаг
6. Спорт
7. Улс төр
8. Шар мэдээ
9. Эдийн засаг
10. Эрүүл мэнд зэрэг болно.

Мэдээний ангилалыг сургах корпус үүсгэхэд материал хангалттай олдсон. Сургалтын корпусын параметрүүдээ http://news.gogo.mn/-ний мэдээн дээр тулгуурлан сонголоо.

Мэдээг ангиласнаар дараах асуудлуудыг шийдвэрлэх боломжтой. Үүнд:

- Тухайн мэдээ ямар ангилалд хамаарч байгааг мэдэх
- Тухайн мэдээнд юуны талаар өгүүлснийг мэдэх

Энэхүү сэдвээр Монголд өнөөг хүртэл хийсэн судалгааны ажил байхгүй байна.

Гадаадад дараах нилээн хэдэн судалгааны ажил хийгдсэн байна. Тухайлбал:

Текстийг ангилагч (Jelinek et al., 1994, Magerman, 1995)-ийн SDT-д үндэслэсэн дүрмийн задлан шинжлэгчийн нэг хэсэг юм. Wall Street сэтгүүл дээр 96.5%-ийн амжилттайгаар ажилласан нь (Magerman, 1995) уг өгүүлэлд ашигласан аргынхтай адил юм. Гэвч дээр өгүүлснээр SDT арга нь өгөгдлийн хуваагдлын эсрэг үгийн олонлогийг (Brown et al., 1992) байгуулахыг шаарддаг ба өгөгдөл хуваагдах үзэгдлийг багасгахын тулд өргөтгөсөн тэгшлэх алгоритмыг хэрэглэдэг. Хамгийн их энтропийн арга нь SDT-гийн адил өгөгдлийг рекурсээр хуваадаггүй тул өгөгдөл хуваагдлаас үүсэх найдваргүй тоон үзүүлэлтээс зайлсхийж чаддаг. Иймд үгийн олонлогийг

хэрэглэх шаардлагагүй байдаг ба энгийн хязгаарлалт тавих аргаар тэгшлэхэд ижил амжилттайгаар ажилладаг.

TBL нь статистик биш арга бөгөөд мөн л олон тооны хувьсагчийг ашиглан нийт үгийг 96.5%-ийн амжилттайгаар, шинэ үгийг 85%-ийн амжилттайгаар ангилдаг (Brill 1994). TBL дэх ойр орших үгсийн дүрслэл нь энэ өгүүллийн загварынхтай адил ба TBL-ийн шинэ үгийн дүрслэл нь манай загварын дүрслэлийн өргөтгөсөн олонлог юм. TBL нь магадлалын загвар биш учраас ямар нэг тархалт ашиглахгүй тул MaxEnt-ийн адил өөр том загварын модуль болон ажиллах боломжгүй юм. Хамгийн их энтропийн арга нь ангилах ямар ч шийдвэрийн хувьд магадлалыг тооцоолж чадах ба энэ мэдээллийг ашиглан нэр үгийн бүлгийг болон дүрмийн модыг байгуулах зэрэгт ашиглах боломжтой (Jelinek et al., 1994, Magerman, 1995).

Гадаад орнуудад хийгдээд нилээдгүй хугацаа өнгөрч, түүнийгээ даган амжилтын хувь болон боломжууд нь өссөөр байгаа юм. Иймээс Байесийн ангилагч ашиглан Монгол текстийг агуулгаар нь ангилах чиглэлийн судалгааны ажлын анхны туршилт юм.

II. БАЙЕСИЙН АНГИЛАГЧ АШИГЛАН ТЕКСТИЙГ АНГИЛАХ НЬ

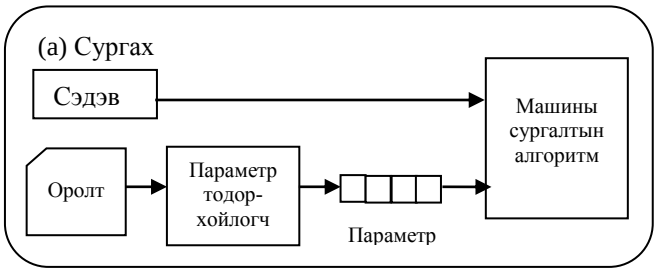
Ангилагч нь өгөгдсөн текстийг зөв ангилал руу сонгон оруулах ёстой. Харин олон мэдээг ангилахдаа тодорхой дэс дарааллын дагуу ангилдаг.

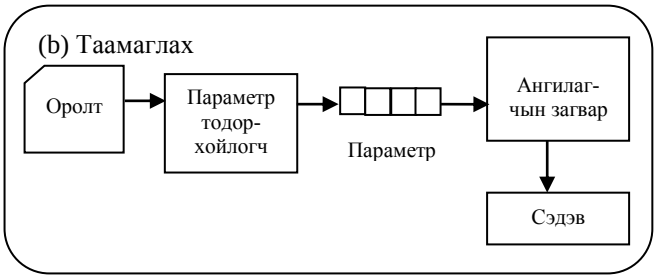
Байесийн ангилагчын дүрэм:
D гэсэн өгөгдлийн олонлог болох $D = (x_1, x_2, \dots, x_n)$ –г $c_j \in C$ ангилал руу ангилахдаа:

$$c_{MAP} = \underset{c_j \in C}{argmax}\, P\ (c_j | x_1, x_2, \dots, x_n)$$

$$= \underset{c_j \in C}{argmax}\, \frac{P\ (x_1, x_2, \dots, x_n | c_j) P(c_j)}{P(x_1, x_2, \dots, x_n)}$$

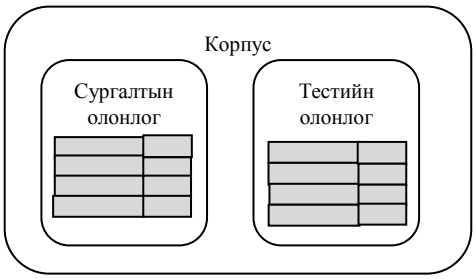
$$= \underset{c_j \in C}{argmax}\, P\ (x_1, x_2, \dots, x_n | c_j) P(c_j)\ (1)$$



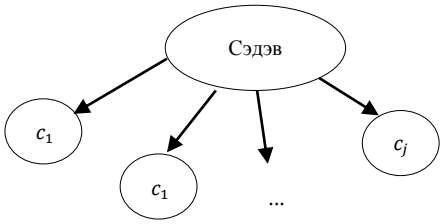


Зураг 1. Автомат ангилагчын мэдээ таних сургалтын машины ерөнхий дараалал

Корпус үүсгэн сургалтын олонлог болон туршилтын олонлогуудыг оруулна.



Зураг 2. Ангилагчийн сургалтын корпусын зохион байгуулалт



Зураг 3. Ангилагчын сүлжээний граф

График 3-т үзүүлсэний дагуу орсон текстийн өгөгдлөөс таних магадлалыг томъёогоор дүрсэлвэл:

$$P(\text{СЭД}|\text{ПАРА} = P(\text{ПАРА}, \text{СЭД}) / P(\text{ПАРА})\ (2)$$

$$\text{P}(\text{ПАРА}) = \sum_{|\text{ОРОЛТ}|\text{СЭДЭВ}} P(\text{ПАРА}, \text{СЭД})\ (3)$$

$$P(\text{СЭД}|\text{ПАРА} = P(\text{ПАРА}, \text{СЭД}) / P(\text{ПАРА})\ (4)$$

$$P(\text{ПАРА}, \text{СЭД}) = P(\text{СЭД}) \times P(\text{СЭД}|\text{ПАРА})\ (5)$$

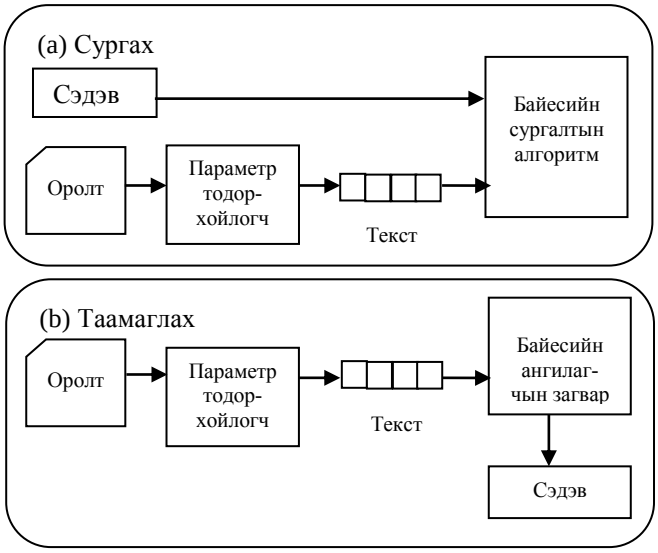
$$P(\text{ПАРА}, \text{СЭД}) = P(\text{СЭД}) \times P_{f\ \text{ОРОЛТЫН ПАРА УТГА}}\, P(\text{ПАРА.}|\text{СЭДЭВ})\ (6)$$

$$P(\text{ПАРА}, \text{СЭД}) = \text{ТООЛ}(f, \text{СЭД}) / \text{ТООЛ}(\text{СЭД})\ (7)$$

ПАРА - Параметр
СЭД – Сэдэв

III. Монгол мэдээнд текстийн ангилал ашигласан туршилтын үр дүн

Мэдээний 10 ангилал тус бүрээр 20, нийт 200 мэдээ бүхий сургалтын корпус үүсгэлээ. Энэхүү корпус дээр үндэслэн дараах График 4-т дүрсэлсэн дарааллын дагуу туршилтыг хийлээ.



Зураг 4. Байесийн ангилагчын мэдээ таних сургалтын машины ерөнхий дараалал

Параметр тодорхойлогчыг **def document_features** гэсэн функцээр илэрхийлсэн. Тодруулбал:

```
def document_features(document):
        document_words = set(document)
        features = {}
        for word in word_features:
                features['contains(%s)'
% word] = (word in document_words)
        return features
```

Жишээгээр Байесийн ангилагчын мэдээ таних сургалтын машины ерөнхий дарааллыг харуулвал:

Оролтын жишээ мэдээ:

<< Сонгинохайрханчууд ням гарагт сонголтоо хийнэ УИХ-ын сонгуулийн 26-р тойргийн дахин санал хураалт энэ сарын 14-ний ням гарагийн 07.00-20.00 цагийн хооронд явагдана. Энэ тойрогт АН-аас Л.Эрхэмбаяр, МАН-аас Д.Сумъяабазар нар нэр дэвшиж буй. Дахин санал хураалтад бэлтгэх ажлын хүрээнд өчигдөр бүх хэсэгт туршилтын санал хураалт явуулжээ. Үүнд 1600 хүн оролцож, бүх автоматжуулсан машин хэвийн ажилласан байна. Тус тойрогт өнгөрсөн сонгуулийн, тодруулбал 2012 зургадугаар сарын 28-ны санал хураалтын нэрсийн хавтгайгаар хүний тоог авч буй бөгөөд нийт 179824 сонгогчийн нэрсийн жагсаалтыг хэсгүүдэд тараасан аж. Дахин сонгуулийн санал хураалтын ажлын хүрээнд сонгуулийн 80 хэсэгт 1500 гаруй хүн ажиллаж байгаа юм байна. Мөн сонгуульд зориулан Монгол улсын УИХ-аас 59 сая төгрөгийн төсвийг батласан тухай тойргийн хорооны дарга М.Баасандорж хэлж байлаа. СХД-ийн Тамгын газар, 26-р тойргийн хорооноос нийт Сонгинохайрханчуудад сонгуульдаа идэвхтэй оролцохыг уриаллаа. Санал хураалтад хувь заахгүй бөгөөд нийт санал өгсөн ирцээс хамгийн их санал авсан нь УИХ-ын гишүүн болох юм. >>

Байесийн ангилагчын мэдээ таних програмыг ажиллуулвал дараах үр дүнд хүрч байна:
Програмын таних магадлал: 0,53
А-Агуулагдаж буй байдал

Текст	Үнэн Худал	Сэдэв1	Сэдэв2	А 1	А2
аж	Худал	Гэмт хэрэг	Улс төр	4.5	1.0
байна	Худал	Дэлхий	Улс төр	4.5	1.0
Монгол	Үнэн	Улс төр	Нийгэм	4.4	1.0
УИХ	Үнэн	Улс төр	Эрүүл мэнд	3.4	1.0
...	...	...	...	...	...

Хүснэгт 1. Өгөгдсөн мэдээнд хамгийн их тохиолдож буй текст

Корпусын нийт 200 мэдээнд орсон хамгийн их агуулагдаж буй эхний 20 текстийг ***Хүснэгт 2***-т харууллаа.
А-Агуулагдаж буй байдал

№	Текст	Сэдэв1	Сэдэв2	А 1	А 2
1	УИХ	Улс төр	Эдийн засаг	5.7	1.0
2	энэ	Эдийн засаг	Дэлхий	4.9	1.0
3	эрүүл	Эрүүл мэнд	Нийгэм	4.4	1.0
4	газар	Боловсрол	Нийгэм	4.4	1.0
5	хоёр	Улс төр	Эрүүл мэнд	3.4	1.0
6	харин	Нийгэм	Шар мэдээ	4.0	1.0
7	бөгөөд	Шар мэдээ	Нийгэм	4.0	1.0
8	болно	Соёл урлаг	Эдийн засаг	4.0	1.0
9	жил	Соёл урлаг	Эдийн засаг	4.0	1.0
10	уу	Соёл урлаг	Эдийн засаг	4.0	1.0
11	гэдэг	Соёл урлаг	Нийгэм	3.6	1.0
12	мэндийн	Эрүүл мэнд	Нийгэм	3.6	1.0
13	хүн	Эрүүл мэнд	Нийгэм	3.6	1.0
14	эх	Боловсрол	Улс төр	3.5	1.0
15	нь	Эдийн засаг	Дэлхий	3.5	1.0
16	юм	Эдийн засаг	Дэлхий	3.5	1.0
17	мянган	Дэлхий	Эдийн засаг	3.5	1.0
18	хувьд	Эдийн засаг	Нийгэм	3.3	1.0
19	эмнэлгийн	Эрүүл мэнд	Улс төр	3.3	1.0
20	дарга	Улс төр	Соёл урлаг	3.1	1.0
	...	...	...	...	...

Хүснэгт 2. Хамгийн их агуулагдаж буй текст

Судалгааны сургалтын корпусыг алхам алхамаар нэмэгдүүлэх бүрт Байесийн ангилагчын мэдээ танилт өсөлт болон бууралттай байсан. Харин корпусын хэмжээ 180 болсоны дараагаар өсөх хандлагатай болсон.

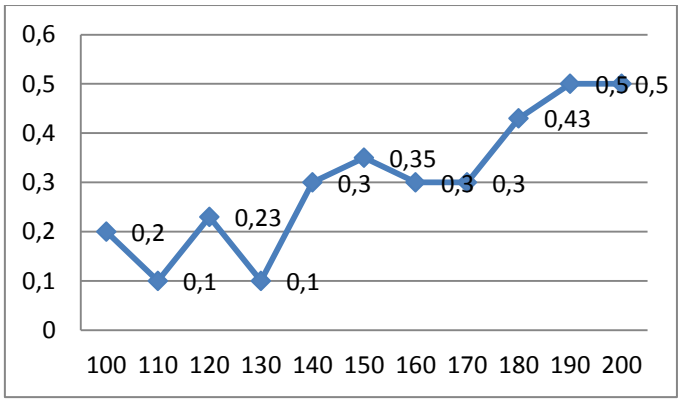


График 1. Байесийн ангилагчын мэдээ танилт корпусын хэмжээнээс хамаарч буйг магадлалын харьцуулалт

IV. Дүгнэлт

Судалгааны ажлыг товч дүгнэвэл:

- Монголын мэдээний портал сайтууд мэдээг ангилахдаа оператор буюу хүний тусламжтайгаар ангилдаг. Энэхүү текст ангилагчийг бодитоор ашиглавал цаг хугацааг хэмнээд зогсохгүй тодорхой ажилтнуудын орон тоог багасгах болон байгууллагын хүний нөөцийн зардлыг бууруулах боломжтой юм.
- Судалгааны ажлыг явуулахад мэдээний сайтууд дээрхи мэдээлэлд үг үсгийн алдаа, ялангуяа нэг үгийг дунд нь зайтай бичсэнээс тухайн үгийг програм хоёр үг гэж таних, ижил үгийг ялгаатай бичсэнээс ялгаатай үг гэж таних зэрэг асуудлууд нь хүндрэл учруулсан. Үүнээс гадна монгол кирилл үсэг програмчлалын хэл дээр танигдахгүй байсан тул галиглах асуудалд ч мөн цаг хугацаа ихээхэн зарцуулсан билээ.
- Корпусын хэмжээг нэмэгдүүлсэнээр тесктийн ангилал таних магадлалыг өндөрсгөх боломжтой.
- Цаашид текстийн ангилагч ашиглан мэдээг таних асуудал дээр тулгуурлан ойролцоо мэдээллийг санал болгох дараагийн судалгааг үргэлжлүүлэн судална.
- Дараагийн судалгаанд хэлний загварын паттерн дээр суурилсан ангилагчийн аргыг ашиглана.

Ном зүй

[1] G. Eason, B. Noble, and I.N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (*references*)

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[3] I.S. Jacobs and C.P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[4] K. Elissa, “Title of paper if known,” unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[8] Christopher D. Manning, Hinrich Schutze 1999. Foundations of Statistical Natural Language Processing. Second Edition. Chapter 16 Text Categorization. pp 575-610

[9] Jian mei, Wu zhang and Suge wang 2012. Grid Enabled problem solving environments for Text Categorization

[10] Processing and Machine Translation, Chapter 1.2.1 Large Scale Multilingual Broadcast Data Collection to Support Machine Translation and Distillation Technology Development

[11] Machine Learning in Automated Text Categorization. Fabrizio Sebastiani Consiglio Nazionale delle Ricerche, Italy

[12] Natural Language Processing and Knowledge Representation Language for Knowledge and Knowledge for Language , Lucja M. Iwanska , Stuart C. Shapiro

[13] Apte, C., Damerau, F. and Weiss, S. Automated learning of decision rules for text categorization. ACM Transactions on Information Systems, 12(3), 233-251,1994.

[14] Chakrabarti, S., Dom, B., Agrawal, R. and Raghavan, P.Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. The VLDB Journal 7, 163-178, 1998.

[15] Chen, H. and Dumais, S. Bringing order to the web: Automatically categorizing search results. Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI’00), 145-152, 2000.

[16] Dumais, S. T., Platt, J., Heckerman, D. and Sahami, M.Inductive learning algorithms and representations for text categorization. Proceedings of the Seventh International Conference on Information and Knowledge Management (CIKM’98), 148-155, 1998.

[17] Hearst, M., and Karadi, C. Searching and browsing text collections with large category hierarchies. Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI’97), Conference Companion, 1997.

[18] Mladenic, D. and Grobelnik, M. Feature selection for classification based on text hierarchy. Proceedings of the Workshop on Learning from Text and the Web, 1998

[19] http://nlp.stanford.edu/ Стэнфордын Их Сургуулийн Эх Хэлний боловсруулалтын судалгааны вэб сайт

[20] http://www.phyton.org Програмчлалын хэл