

LEKSIK DARAJADAGI PEREFRAZ TAKSONOMIYASINI AMALGA OSHIRISH BOSQICHLARI

Axmedova Xolisxon Ilxomovna,
Texnika fanlari falsafa doktori PhD
a.xolisa@navoiy-uni.uz
ToshDO‘TAU

Ko‘charova Iroda Shavkat qizi,
Kompyuter lingvistikasi yo‘nalishi talabasi
kocharovairoda516@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada, tabiiy tilga ishlov berish (Natural Language Processing - NLP) sohasidagi perefraz qilish (paraphrasing) algoritmlariga oid muhim tushunchalar keltirib o‘tiladi. Perefraz qilishning leksik darajadagi taksonomiyasi, sinonimlari, antonimlari va morfologik o‘zgarishlari kabi turlari orqali gaplarni semantik jihatdan ekvivalent tarzda ifodalash jarayoni o‘rganiladi. Glove algoritmi so‘z vektorlarining o‘rtacha qiymati va Skip-gram usullari yordamida gaplarning semantik o‘xshashligini baholashga yordam beradi.

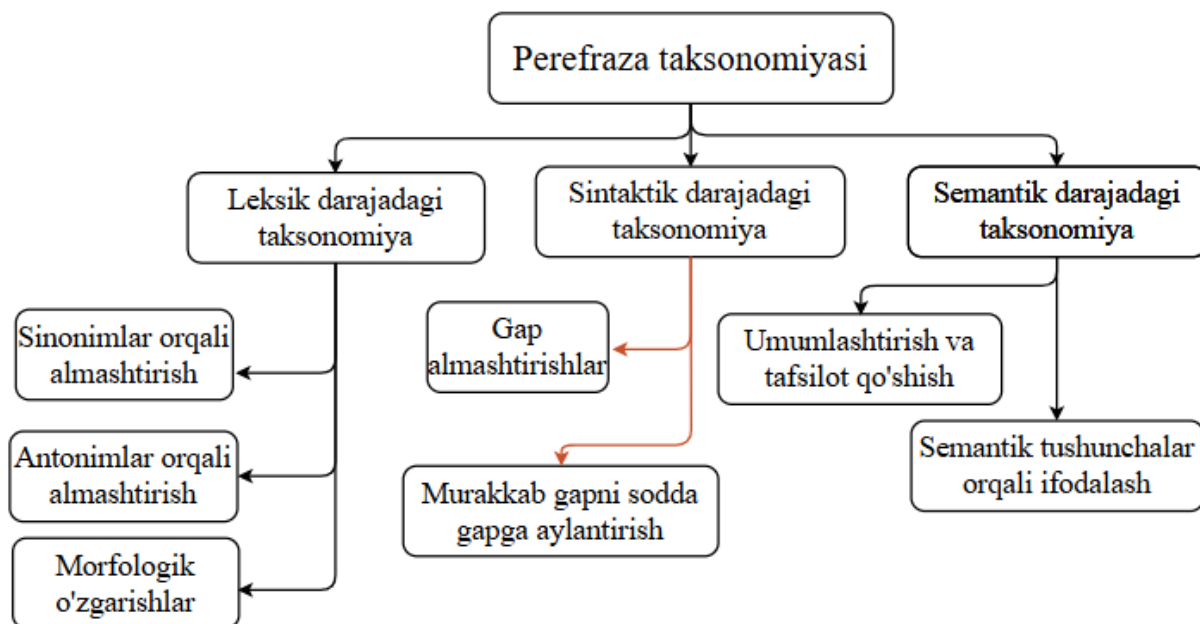
Abstract: This article presents important concepts related to paraphrasing algorithms in the field of Natural Language Processing (NLP). It explores the process of expressing sentences in a semantically equivalent manner through types such as lexical-level taxonomy, synonyms, antonyms, and morphological variations. The GloVe algorithm and Skip-gram methods help evaluate the semantic similarity of sentences based on the average value of word vectors.

Аннотация. В данной статье представлены важные понятия, относящиеся к алгоритмам перефразирования в области обработки естественного языка (Natural Language Processing — NLP). Рассматривается процесс выражения предложений в семантически эквивалентной форме через такие виды, как таксономия на лексическом уровне, синонимы, антонимы и морфологические изменения. Алгоритм GloVe и метод Skip-gram помогают оценить семантическое сходство предложений на основе среднего значения векторов слов.

Kalit so‘zlar: *Perafraz taksonomiyasi, NLP, Skip-gram, leksik darajadagi taksonomiya, sinonim, antonym, morfologik o‘zgarishlar, GloVe.*

Tabiiy tilga ishlov berish texnologiyalari sun‘iy intellektning muhim yo‘nalishlaridan biri bo‘lib, tilni tushunish va ishlov berishda asosiy o‘rinni egallaydi. Ushbu maqola so‘zlarni sonli ifodalashning turli yondashuvlari va chuqur o‘rganish modellarining asoslari, afzallik va qiyinchiliklarini o‘rganish uchun tezkor ma’lumot bo‘lib xizmat qilishi mumkin. Perafraz - bu bir xil ma’noni ifodalovchi turli xil so‘z yoki iboralar bilan ifodalash usuli. Masalan, “o‘qituvchi” va “muallim”

soʻzlari bir xil maʼnoni bildiradi, lekin turli kontekstlarda qoʻllanilishi mumkin. Tabiiy tilga ishlov berish sohasida gaplarni perefraz qilish, gap ekvivalentlarni yaratish, gaplarni semantik generatsiya qilish tushunchalari oʻzaro sinonim tushunchalardir. Gapning perefrazi uning maʼnosini saqlab qoladi, biroq asl shakldan farqli soʻz va iboralarni tanlaydi. Perefraz yaratish (Paraphrase Generation-PG) turli tabiiy tilga ishlov berish ilovalari ishlab chiqishda muhim masala boʻlib, u tabiiy til birliklarini kengaytiradi. Perefraz yaratish *savol-javob tizimlari, mashina tarjimasini va axborot izlash* kabi koʻplab vazifalarni bajarishda uchun muhim rol oʻynaydi [1]. Savol - javob tizimlarida PG semantik jihatdan ekvivalent savollarni hamda ularga turli koʻrinishdagi javoblarni yaratish uchun ishlatilishi mumkin. Bu javoblarni qidirishda soʻzlar mos kelmasligi muammolarini hal qilishga yordam beradi [2]. Perafraza taksonomiyasi esa bunday oʻxshashliklarni turkumlash va tasniflash uchun tuzilgan sxema(1-rasm).



1-rasm: Perefraz qilish taksonomiyasi

Oʻzbek tilidagi gaplarni perefraz qilishni taksonomiya boʻyicha amalga oshirish jarayonini koʻrib chiqamiz.

Leksik darajadagi taksonomiya - bu soʻzlar, ularning maʼnolari va bir-biriga boʻlgan munosabatlari asosida soʻzlarning turkumlarini tasniflash yoki tartibga solish tizimi. Odatda, bu jarayon leksik birikmalarni oʻrganish va tasniflash orqali amalga oshiriladi. Leksik darajadagi taksonomiyaning asosiy yoʻnalishlari 1-rasmda ham aks etgan boʻlib, bulardan birinchi bosqich sinonimlar orqali almashtirishdir. *Sinonimlar orqali almashtirish* - bu soʻzlarni yoki iboralarni maʼno oʻzgarishsiz, lekin turlicha shaklda ifodalash uchun ishlatiladigan usul. Masalan, quyidagi gapni koʻrib chiqamiz.

Asl gap: *“Odobli insonlar nafaqat boshqalarga hurmat ko‘rsatadi, balki o‘zini ham tinch va baxtli his qiladi”*

Sinonimi orqali almashtirilgan gaplar: bunda gapdagi har bir so‘zning sinonimlari ro‘yxatini shakllantirish kerak bo‘ladi. Ya’ni

Odobli -> boadab, axloqli, tarbiyali.

1. Boadab insonlar nafaqat boshqalarga hurmat ko‘rsatadi, balki o‘zini ham tinch va baxtli his qiladi.

2. Axloqli insonlar nafaqat boshqalarga hurmat ko‘rsatadi, balki o‘zini ham tinch va baxtli his qiladi.

3. Tarbiyali insonlar nafaqat boshqalarga hurmat ko‘rsatadi, balki o‘zini ham tinch va baxtli his qiladi.

Bu yerda bir so‘z o‘rnida uning sinonimi ishtirok etgan bo‘lsada, gapning ma’nosiga ta’sir ko‘rsatmayabdi. Sinonimlar orqali almashtirish, tilni yanada boyitishga yordam beradi va gaplarni turli shakllarda ifoda etishda samarali usul hisoblanadi. Bu orqali matnni yangilash va gapning ma’no samaradorligini oshirish mumkin.

Antonimlar orqali almashtirish - bu o‘zaro qarama-qarshi bo‘lgan so‘zlar yoki iboralar yordamida matndagi so‘zlar yoki gaplarni almashtirish usuli. Masalan:

Asl gap: *“Uning so‘zlari aniq va mantiqiy edi, lekin fikrlashi tor”.*

Antonimi orqali almashtirilgan gap: *“Uning so‘zlari aniq va mantiqiy edi, lekin fikrlashi keng emas”*

Asl gap: *“Izohli lug‘at kartotekasi hali ham kambag‘al”*

Antonimi orqali almashtirilgan gap: *“Izohli lug‘at kartotekasi hali ham boy emas”*

Yuqoridagi misolda ko‘rganimizdek, asl gapga nisbatan antonimi orqali almashtirilgan gapda ma’no boyoqdoligi yuqoriroqdir. Ushbu yo‘nalish, ayniqsa, *badiiy asarlarda, insholar va tahliliy matnlarda* juda foydalidir. O‘qishga qiziqarli va boy ma’no kiritish uchun antonimlardan foydalanish matnni yangi nuqtai nazardan ko‘rishga yordam beradi.

Leksik darajadagi taksonomiyaning oxirgi yo‘nalishi bu - morfologik o‘zgarish. Ushbu yo‘nalish so‘zlarga qo‘shiladigan shakllarni o‘zgartirish orqali ifodalanadi. Misol sifatida quyidagilarni keltirish mumkin.

Asl gap: *“U mamlakatda ekonomika fanlari akademiyasini barpo qilish masalasini muhokama qilish kerak edi deb hisoblaydi”.*

Perefraz qilingan gap: “*U mamlakatda ekonomika fanlari akademiyasini barpo qilish masalasini muhokama qilish kerak edi deb hisoblamaydi emas*”.

Asl gap: “*Uning gaplari bolalarcha sodda, samimiy va aniq edi*”.

Perefraz qilingan gap: “*Uning gaplari bolalarcha sodda, samimiy va noaniq emas edi*”.

So‘zlarning morfologik shaklini o‘zgartirish gap yoki matn ma’nosiga ta’sir qilmaydi. Gap uchun to‘g‘ri tanlangan morfologik shakllar so‘zlarning ma’nosini aniqroq ifodalashga yordam beradi, matnni boyitadi, soddalashtiradi yoki yangilaydi.

Leksik avtomatik perefrazni amalga oshirishda bir algoritmlardan foydalanish mumkin. Shunday algoritmlardan biri GloVe algoritmi bo‘lib, uning asosiy vazifasi so‘zning semantik o‘xshashlik koeffitsiyentlari aniqlashdan iborat. O‘xshashlik koeffitsiyentlarini aniqlashda dastlab berilgan dataset asosida matritsa tuziladi. Bu matritsadan gaplarning xususiyatlarini hisoblash uchun ishlatiladi. Ushbu matritsaning qatorlari korpus gaplarini, ustunlari esa xususiyatlarni ifodalaydi. Xususiyatlar quyidagilarni o‘z ichiga olishi mumkin:

Unigrams (bitta so‘z)	“Men”, “yangi”, “kitob”, “o‘qidim”
Bigrams (ketma-ket keluvchi so‘z juftliklari)	“Men yangi”, “yangi kitob”, “kitob o‘qidim”
Dependency pairs (so‘zlar orasidagi munosabat)	(“Men”, “o‘qidim”) - “Men” subyekt, “o‘qidim” fe’l bilan bog‘langan. (“yangi”, “kitob”) - “yangi” sifat, “kitob” ot bilan bog‘langan. (“kitob”, “o‘qidim”) - “kitob” obyekt, “o‘qidim” fe’l bilan bog‘langan.

2-rasm: Xususiyatlarni taqqoslash jadvali

Matritsa tuzilganidan so‘ng, so‘z chastotasining Kullback-Leibler farqlanishi (TF-KLD) xususiyat uchun vazn sxemasi hisoblanadi. Ushbu vazn berishning maqsadi muhimroq xususiyatlarga yuqori vazn berishdir. Yangi vazn berish sxemasi barcha xususiyatlarning vaznini hisoblash uchun quyidagi faktorlarga asoslanadi:

Xususiyatning parafrazlangan gaplarda mavjud bo‘lish ehtimoli	U juda yaxshi raqsga tushishni biladi.
Xususiyatning parafrazlanmagan gaplarda mavjud bo‘lish ehtimoli	U yaxshi raqsga tushadi.

3-rasm: Perefrazlashda xususiyatlar: Mavjudlik ehtimollari

Matritsa faktorizatsiyasi usuli yordamida gaplar vektorlarini ajratish uchun ishlatiladi. Ushbu vektorlar yordamida gaplar o'rtasidagi o'xshashlikni o'lchash mumkin.

So'z vektorlarining o'rtacha qiymati

Ushbu usulda gapni ifodalovchi vektor so'zlarning o'rtacha vektorlari ushbu gapga tegishli deb hisoblanadi. Oldindan o'qitilgan *qo'lqop* so'z vektorlaridan foydalangan[12]. *Qo'lqop* semantik maydonidan gapdagi har bir so'z uchun mos keladigan vektor tushuniladi va gaplarni ifodalash uchun ushbu vektorlar orasidagi kosinus o'xshashligi gaplarning semantik o'xshashligini baholash uchun ishlatiladi.

O'qitishga asoslangan vektorlar. Bundan tashqari, so'zlarni joylashtirish (word embedding) yordamida yaratilgan so'z vektorlari asosida gap o'xshashligini aniqlash modelini Mueller taklif qilgan[19]. Uzoq va qisqa muddatli xotira (LSTM) modelida oddiy bir o'zgartirish taklif qilinadi va treningda qo'llaniladi. Kirish ma'lumotlarida har bir gap so'z vektorlaridan iborat bo'lib, bu vektorlar bir xil o'lchamga ega. Bundan tashqari, har bir gap juftligi uchun kirish ma'lumotlarida inson tomonidan baholangan o'xshashlikni ifodalovchi belgi beriladi. Taklif qilingan Manhattan LSTM (MaLSTM) modelini kirish gapi vektorlarini o'qish uchun ishlatadi va oxirgi yashirin qatlamni kirish gaplari uchun taqdimot sifatida ishlatadi. Gaplar uchun yangi yaratilgan taqdimot semantik o'xshashlik o'lchovini o'rganish uchun qo'llaniladi.

Skip-Gramm vektorlari. Ushbu usul nazoratsiz o'quv modeliga asoslangan gaplar uchun vektorli tasvirni oladi. Ushbu modelga kiritilgan senentsiyalar vektorlar to'plami sifatida ifodalanadi. Bu vektorlar, gaplarning so'zlariga mos keladigan, oldindan o'qitilgan word2vec fazosidan tanlanadi. Skip-thought modeli, berilgan so'zga asoslanib atrofdagi so'zlarni bashorat qiladigan Skip-gram modeliga o'xshashdir. Biroq, Skip-thought modeli fokuslangan gapga asoslanib atrofdagi gaplarni bashorat qiladi. Gaplar o'xshashligini hisoblashda, avvalo gaplar oldindan o'qitilgan modelga yuborilib, vektor taqdimoti olinadi. So'ngra, o'xshashlik o'rganish vazifasiga asoslanib hisoblanadi.

Xususiyat	Skip-gram	Skip-thought
Maqsad	Soʻzlarni konteksdagi boshqa soʻzlar yordamida bashorat qilish	Gap yoki matnni vektorga aylantirish
Yondashuv	Soʻzlar oʻrtasidagi semantic oʻxshashlikni oʻrganish	Gaplar oʻrtasidagi semantic oʻxshashlikni oʻrganish
Natija	Soʻz vektorlarini yaratadi	Gap yoki matn vektorlarini yaratadi
Turi	Word2Vec modelida ishlatiladi	Recurrent Neural Network (RNN) yordamida ishlaydi
Misol	“Men kitobni oʻqidim” - [oʻqish, kitob]	“Men kitobni oʻqidim” va “Men sport bilan shugʻullandim” orasidagi oʻxshashlik

4-rasm: Skip-gram va Skip-thought model xususiyatlari

Vektor asosidagi oʻxshashlik yondashuvi, ayniqsa chuqur oʻrganish yondashuvini qoʻllaganidan soʻng yaxshi natijalar koʻrsatgan boʻlsa-da, gaplarining asosiy xususiyatlarini aniqlashda hali koʻp yaxshilanishlar uchun imkoniyatlar mavjud.

Umuman olganda, koʻplab yondashuvlar tushuntirilgan texnikalarning ikkita yoki undan ortiq birlashtirilishini qoʻllaydi. 1-jadvalda muhokama qilingan baʼzi yondashuvlar va har bir taklif qilingan tizimda ishlatilgan texnikalar koʻrsatilgan. Bundan tashqari, har bir taklif qilingan tizimni baholash uchun ishlatilgan maʼlumotlar toʻplami va erishilgan natijalar keltirilgan. 1-jadvalda har bir yondashuvning tafsilotlarini koʻrsatadi va zamonaviy texnologiyalarni chizadi. 1-jadvalda koʻrsatilganidek, gap oʻxshashligini oʻlchash uchun aytib oʻtilgan metodlar aralash tarzda qoʻllanilgan.

Bundan tashqari, gaplar oʻrtasidagi oʻxshashlikni oʻlchash uchun turli yondashuvlarni birlashtirish yaxshi natijalar koʻrsatdi, chunki bu muammoni turli nuqtai nazarlardan koʻrib chiqadi. Masalan, soʻzga asoslangan yondashuvni va tuzilishga asoslangan yondashuvni birlashtirish, yaxshi tuzilgan gaplar uchun ham, yaxshi tuzilmagan gaplar uchun ham samarali ishlaydigan tizimni yaratadi. Shuningdek, oʻxshashlikni oʻlchashda strukturaviy maʼlumotlardan foydalanish keng tarqalgan emas. Bundan tashqari, gap oʻxshashligini hisoblashda gap tuzilmasidan foydalanishni oʻrganish uchun koʻplab imkoniyatlar mavjud.

Foydalanilgan adabiyotlar:

1. J Edward Hu, Rachel Rudinger, Matt Post, and Benjamin Van Durme. Parabank: Monolingual bi text generation and sentential paraphrasing via lexically constrained neural machine translation. arXiv preprint, 2019.
2. Lin, D., Pantel, P.: Discovery of inference rules for question-answering. Natural Language Engineering 7(4), 343–360 (2001)
3. Mengqiu Vang, Smit na, Mitamura T. jeop- ardy modeli nima? Qa uchun kvazi-sinxron grammatika, yilda: EMNLP-CoNLL; 2007, 7. p. 22-32.
4. Mamdouh Faruk, Mitsuru Ishizuka, Danushka Bollegala. Grafik taalukli asoslangan semantik qidiruvi. Metadata va semantikani tadqiq qilish bo'yicha 12-xalqaro konferensiya materiallari, Kipr; 2018 yil. https://doi.org/10.1007/978-3-030-14401-2_8.
5. Aliguyev RM. Matnni avtomatik umumlashtirish uchun yangi gap o'xshashlik o'lchovi va gapga asoslangan ekstraktiv texnika, ilovalar bilan ekspert tizimlari, 2009; 36:7764-72. <https://doi.org/10.1016/j.eswa.2008.11.022>.
6. De Boni M, Manandhar S. savolga javob berish uchun gap o'xshashligini semantik dolzarblik metrikasi sifatida ishlatish. Savollarga javob berishda yangi yo'nalishlar bo'yicha AAAI simpoziumining materiallari; 2003.
7. Alzahrani S, Salim N, Abraham A. Plagia - rism lingvistik naqshlarini, matn xususiyatlarini va aniqlash usullarini, tizimlar, odam va kibernetika bo'yicha IEEE operatsiyalarini tushunish. 2012; 42(2): 133-49. <https://doi.org/10.1109/TSMCC.2011.2134847>.
8. Jaro MA. 1985 yilda qo'llanilgan rekord bog'lanish metodologiyasidagi yutuqlar Tampa Florida aholini ro'yxatga olish, Amerika statistika jamiyati jurnali. 1989; 84(406): 414-20. <https://doi.org/10.1080/01621459.1989.10478785>.
9. Vinkler biz. Fellegi-Sunter modelidagi yozuvlarni bog'lash modelidagi taqqoslagich o'lchovlari va Kengaytirilgan qaror qoidalari. Tadqiqotni tadqiq qilish usullari bo'limining materiallari, Amerika statistika assotsiatsiyasi; 1990 yil. p. 354-59.
10. Mikolov t, Chen K, Corrado G, Dean J. vektor kosmosda so'z vakolatxonalari samarali estima - yozildi, arXiv preprint arXiv: 1301.3781. 2013.
11. Miller GA. So'z tarmog'i: ingliz tili uchun leksik ma'lumotlar bazasi, ACM aloqalari. 1995; 38(11):39-41. <https://doi.org/10.1145/219717.219748>.
12. Vu Z, Palmer M. fe'l semantikasi va leksik tanlov. Davomlari 32nd annual meeting of the association Uchun Hisoblash Tilshunosligi, Las Cruces, Nyu-Meksiko; 1994. <https://doi.org/10.3115/981732.981751>.
13. Vijaymeena MK, Kavitha K. matn qazib olish, mashinani o'rganish va ilovalardagi o'xshashlik mea - sures bo'yicha so'rov: xalqaro jurnal. 2016; 3(1):19-28. <https://doi.org/10.5121/mlaij.2016.3103>.



14. Guessoum D, Miraoui M, Tadj C. keng tarqalgan hisoblashda semantik o'xshashlik choralarini o'rganish, aqlli sezgi va aqlli tizimlar bo'yicha xalqaro jurnal. 2015; 8(1): 125-58. <https://doi.org/10.21307/ijssis-2017-752>.
15. Gomaa, fahmi AA. Matn o'xshashligi yondashuvlarini o'rganish, kompyuter ilovalarining InternationalJournal. 2013; 68(13):13-18. <https://doi.org/10.5120/11638-7118>.
16. Pradhan N, Gyanchandani M, Vadhvani R. irda ishlatiladigan matn o'xshashligi texnikasi va uni qo'llash bo'yicha sharh, xalqaro kompyuter dasturlari jurnali. 2015; 120(9). <https://doi.org/10.5120/21257-4109>.
17. Xorxe Martines-Gil, Mario Pich. Semantic yozishmalar kashfiyotini qo'llab-quvvatlash uchun inson adabiyotida so'zlarning birgalikda paydo bo'lishini tahlil qilish. Bilim texnologiyalari va ma'lumotlarga asoslangan biznes bo'yicha 14-xalqaro konferensiya materiallari, Graz, Avstriya; 2014 yil. <https://doi.org/10.1145/2637748.2638422>.
18. Terra E, Clarke CLA. Statistik so'z o'xshashligi o'lchovlari uchun chastotani baholash, NAACL '03 inson Tili texnologiyasi bo'yicha hisoblash tilshunosligi Assotsiatsiyasining Shimoliy Amerika bobining 2003-yilgi konferensiyasining materiallari-1-jild, Edmonton, Kanada; 2003 yil. p. 165-72. <https://doi.org/10.3115/1073445.1073477>.
19. Jonas Myuller, Thyagarajan A. Siam takroriy arxi - tectures gap o'xshashligini o'rganish uchun. Sun'iy intellekt bo'yicha o'ttizinchi AAAI konferensiyasi materiallari, Feniks, Arizona; 2016. p. 2786-92. <https://dl.acm.org/iqtibos.cfm?id=3016291>.