NLPDA MATNLARNI UMUMIYLASHTIRISH VA LEKSIK TAHLIL

Fayzullayeva Zarnigor Inatillayevna
PhD, v.b. dotent Alxorazmiy nomidagi
Toshkent axborot texnologiyalar Universiteti.
E-mail: zarnigor18z02@gmail.com

Karimov Nodirbek Nosirjon o‘g‘li
Doktarant Namangan davlat Universiteti.
E-mail: topcoder1600@gmail.com

Annotatsiya: Hozirgi kunda NLP da matnni umumlashtirish yordamida erishilgan natijalarga juda ham ko‘pligi e'tiborga molik, ushbu maqolada, biz **chuqur o'rganishdan foydalanib, matnni** yig'ish uchun zarur bo'lgan barcha tushunchalarni qamrab olgan holda, uni yaratish bo'yicha bosqichma-bosqich jarayonni ko'rib chiqdik va birinchi matnni umumlashtirish modelimizni amalga oshirdik, Umumlashtirish uchun esa xulosa chiqarish turlarini va amalga oshirish qadmlarini hamda jarayonlarini ko‘rib chiqdik.

Kalit so‘zlar: NLP, Chuqur o‘rganish, leksik tahlil, sun’iy intellekt, xulosa chiqarish, model, vector, matrissa

Abstract: It is remarkable that there is so much progress in NLP using text summarization, in this article we will use deep learning to create a step-by-step approach to text summarization, covering all the concepts needed for text summarization. We looked at the -step process and implemented our first text summarization model, and for Summarization we looked at inference types and implementation steps and processes.

Keywords: NLP, Deep learning, lexical analysis, artificial intelligence, summarization, model, vector, matrix

Аннотация: Примечательно, что в НЛП с использованием обобщения текста достигнут такой большой прогресс, что в этой статье мы будем использовать глубокое обучение для создания пошагового подхода к обобщению текста, охватывающего все концепции, необходимые для обобщения текста. Пошаговый процесс и реализовали нашу первую модель суммирования текста, а для суммирования мы рассмотрели типы вывода, а также этапы и процессы реализации.

Ключевые слова: НЛП, глубокое обучение, лексический анализ, искусственный интеллект, вывод, модель, вектор матрица.

Tabiiy tilni qayta ishlash (NLP) - bu mashinalarga inson tilini o'qish, shifrlash, tushunish va qayta ishlash imkonini beruvchi eng mashhur AI texnologiyasi hamda inson tili ma'lumotlarni yozilishi, og‘zaki va tartibga solinishi bilan boshqara oladigan mashinalarni qurishdir. Matnni bashorat qilish va his-tuyg'ularni tahlil qilishdan nutqni aniqlashgacha, NLP mashinalarga inson aqli va qobiliyatlarini ta'sirchan tarzda taqlid qilish imkonini beradi. NLP ilovalarini yaratish uchun til modellari kalit hisoblanadi. Biroq, noldan murakkab NLP tili modellarini yaratish ancha murakkabdir, shuning uchun AI va ML ishlab chiquvchilari va tadqiqotchilari oldindan

tayyorlangan til modellari bilan birgalikda ishlashadi. Tilni modellari nutqni tanib olish, belgilarni optik aniqlash, qo‘l yozuvini aniqlash, mashina tarjimasi va imloni tuzatish kabi NLPning turli masalalarni yechish uchun qo‘llanadi. Ushbu modellar matnni mashina tiliga o‘qitish uchun teglash texnikasidan foydalanadi, bunda model topshiriqni bajarish uchun bitta ma’lumotlar to‘plamida o‘qitiladi hamda oldingi so‘zning xususiyatidan kelib chiqqan holda, gapdagi keyingi so‘zning qanday bo‘lishini taxmin qilish taklifnii beriladi. Bazi modellar yangi ma’lumotlar to‘plamida turli NLP funksiyalarini bajarish uchun o‘zgartiriladi. Ushbu vazifaning maqsadi ma’lum bir tilda so‘zlar ketma-ketligining paydo bo‘lish ehtimolini o‘rganishdir. Oldindan o‘rgatilgan NLP til modellarning bir nechta turlari majud bo‘lib, ularning ishlash funksiyalarini ko‘rib chiqamiz.

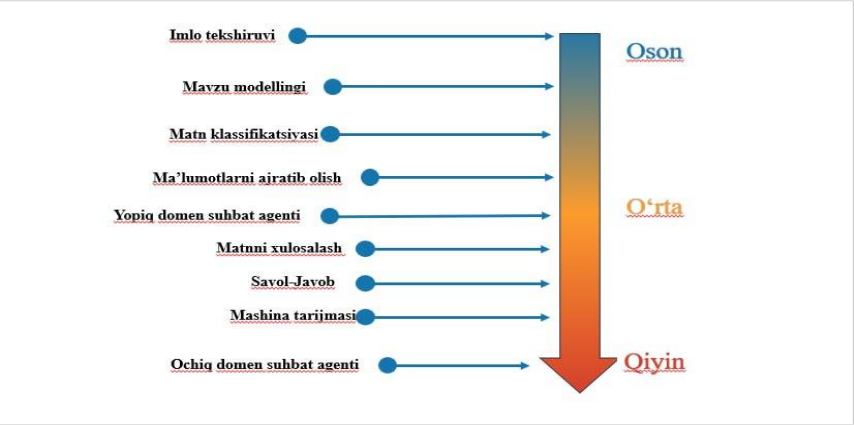
GPT-4 (generativ oldindan o'rgatilgan transformator) - OpenAI tomonidan ishlab chiqilgan katta til modeli (LLM). Bu GPT til modellari seriyasining to‘rtinchi avlodi bo‘lib, 2023-yil 14-martda birinchi marta tadbiq qilingan. GPT-4 multimodal model bo‘lib, u matn hamda tasvirlarni qabul qiladi. Bu faqat matn kiritish sifatida qabul qilinishi mumkin bo‘lgan oldingi GPT modellariga qaraganda ko‘p qirralidir. Ushbu model ChatGPT Plus orqali ommaga ochiq tarzda nomoyon qilingan, uning tijorat API-siga esa kutish ro‘yxati orqali kirish mumkin. Ushbu model rivojlanish davomida keyingi tarkibni oldindan bilishga harakat qiladi. Bu uning insoniy qadriyatlarga mos kelishini va kerakli siyosatlarga mos kelishini ta’minlash uchun qilingan.

U yanada ijodiy va hamkorlikka asoslangan, ijodiy va texnik yozish vazifalarida foydalanuvchilar bilan yaratish, tahrirlash va takrorlash imkoniyatiga ega. U hali ham ishlab chiqilmoqda, lekin u turli xil ilovalar, jumladan, kontent yaratish, tarjima, yozishda yordam, ta’lim, mijozlarga xizmat ko‘rsatish va tadqiqotlar uchun kuchli vosita bo‘lish potentsialiga ega.

BERT (Transformatorlardan ikki tomonlama kodlovchi tasvirlar)- Google tomonidan ishlab chiqilgan NLP-dan oldingi trening texnikasi. U tilni tushunish uchun o‘z-o‘ziga e’tibor mexanizmiga asoslangan yangi neyron tarmoq arxitekturasi transformerdan foydalanadi. U ketma-ketlik transduktsiyasi yoki neyron mashina tarjimasi muammosini hal qilish uchun ishlab chiqilgan. Bu shuni anglatadiki, u kirish ketma-ketligini chiqish ketma-ketligiga aylantiradigan har qanday vazifa uchun mos keladi, masalan nutqni aniqlash, matnni nutqqa o‘zgartirish va hokazo.

Ushbu model ikkita alohida mexanizmni o‘z ichiga oladi: kodlovchi (matn kiritishni o‘qiydi) va dekoder (vazifa uchun bashorat ishlab chiqaradi). BERT mexanizmining maqsadi til modelini yaratishdir. BERT algoritmi 11 ta NLP vazifasini samaralibajaradi.

Matnni mashina tiliga o‘girish uchun ML (machine learning) foydalanamiz bu jarayon ko‘p sonli misollar asosida vazifalarni avtomatik tarzda bajarishni ko‘rishimiz mumkin. Matnni umumiyl ashtirish uchun Chuqur o‘rganish (deep learning, DL) sun’iy neyron tarmoqlari arxitekturalariga asoslangan mashinali o‘rganib chiqamiz. ML, DL va NLP sun’iy intellekt sohadagi barcha kichik maydonlar bo‘lib, ular orasidagi munosabatlar 1-rasmda tasvirlangan.



1-rasm. ML, DL va NLP orasidagi munosabatlar.

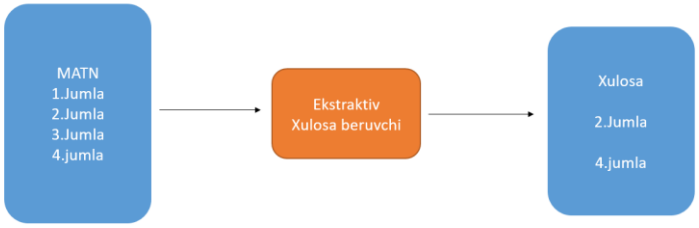
Ma'lumki matnni umumlashtirish bizning hayotimizga katta ta'sir ko'rsatishi mumkin bo'lgan tabiiy tilni qayta ishlash (NLP) ilovalaridan biridir. Raqamli ommaviy axborot vositalari va tobora o'sib borayotgan nashriyotlar bilan - kim foydali yoki yo'qligini aniqlash uchun butun maqolalar / hujjatlar / kitoblarni ko'rib chiqishga hozirgi kunda ko'pgina insonlar vaqtini qizg'onadi shuning uchun ham bu texnologiya allaqachon mavjud lekin o'zbek tili uchun ushbu model hali ishlab chiqilmagan.

Matnni umumlashtirish - bu hujjatning muhim nuqtalarini ajratib ko'rsatish orqali erishilgan asl matnning asosiy ma'lumotlari va asosiy g'oyalarini qamrab oladigan qisqa va yaxshi yozilgan xulosaga uzun matnli hujjatni ixchamlashtirish texnikasini anglatadi. Matnni umumlashtirish uchun ikki xil yondashuv qo'llaniladi:

- Xulosa chiqarish
- Abstraktiv xulosa

Xulosa chiqarish bu butun matn bo'yicha ko'p takrorlanadigan so'zlarni hamda ularning sinonimini izlaydi va **asl matndan muhim jumlalarni yoki iboralarni aniqlaymiz va faqat matndan ajratib olib,** chiqarilgan jumlalar xulosa sifatida taqdim qiladi. Xulosa chiqarishni ikki toifaga bo'lish mumkin – **ekstraktiv xulosa** va **mavhum xulosa**.

Ekstrativ xulosa chiqarish: Bu usullar matndan iboralar va jumlalar kabi bir nechta qismlarni ajratib olishga va xulosa yaratish uchun ularni bir joyga to'plashga tayanadi. Shuning uchun xulosa qilish uchun to'g'ri jumlalarni aniqlash ekstraktiv usulda eng muhim ahamiyatga ega.



2-rasm. Ekstrativ xulosa.

Mavhum xulosa: Bu usullar mutlaqo yangi xulosani yaratish uchun ilg'or NLP usullaridan foydalanadi. Ushbu xulosaning ba'zi qismlari hatto asl matnda ham ko'rinmasligi mumkin.

Abstraktiv xulosa Bu juda qiziq yondashuv. Bu erda biz asl matndan yangi jumlalar yaratamiz. Bu biz ilgari ko'rgan ekstraktiv yondashuvdan farqli o'laroq, biz faqat mavjud

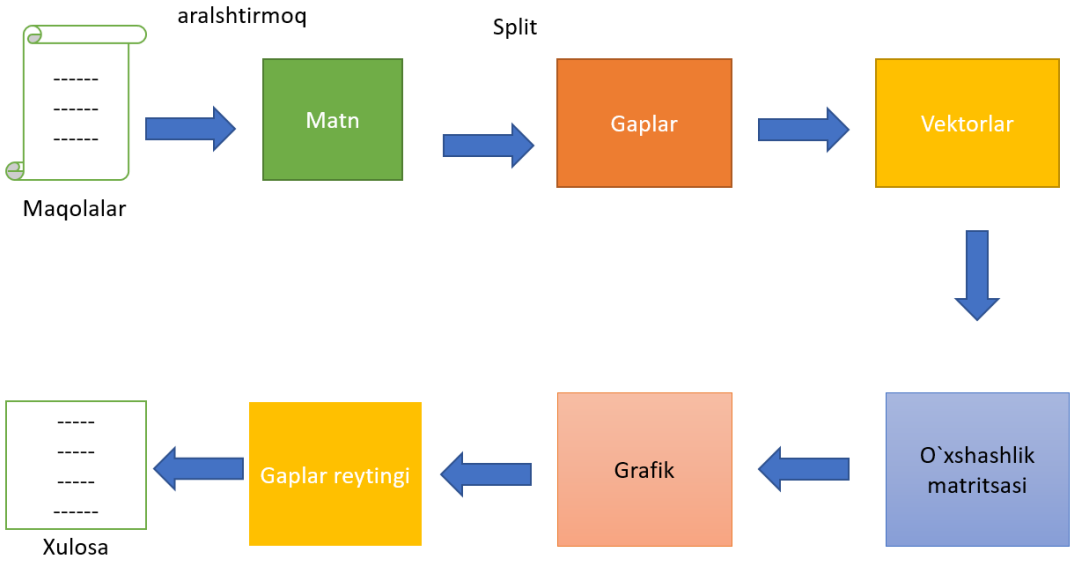
jumlalarni ishlatganmiz. Mavhum umumlashtirish orqali tuzilgan jumlalar asl matnda mavjud bo'lmasligi mumkin va bu jarayon rasmda tasvirlangan:

3-rasm. Abstraktiv xulosa.

Biz matnni umumiylashtirish uchun TextRank algoritmidan foydalandik.TextRank algoritminini amalaga oshirish uchun PageRank foydalandik. Ushbu ikki algoritm o‘rtasidagi o'xshashliklarni quyidagicha:

- jumlalarni ishlatamiz
- har qanday ikkita jumla o‘rtasidagi o‘xshashlik web-sahifaga o‘tish ehtimoliga ekvivalent sifatida ishlatiladi
- o‘xshashlik ballari PageRank uchun ishlatiladigan M matritsasiga o‘xshash kvadrat matritsada saqlanadi.

TextRank - bu ekstraktiv va nazoratsiz matnni umumlashtirish usuli. Keling, biz amal qiladigan TextRank algoritmi oqimini ko'rib chiqaylik:



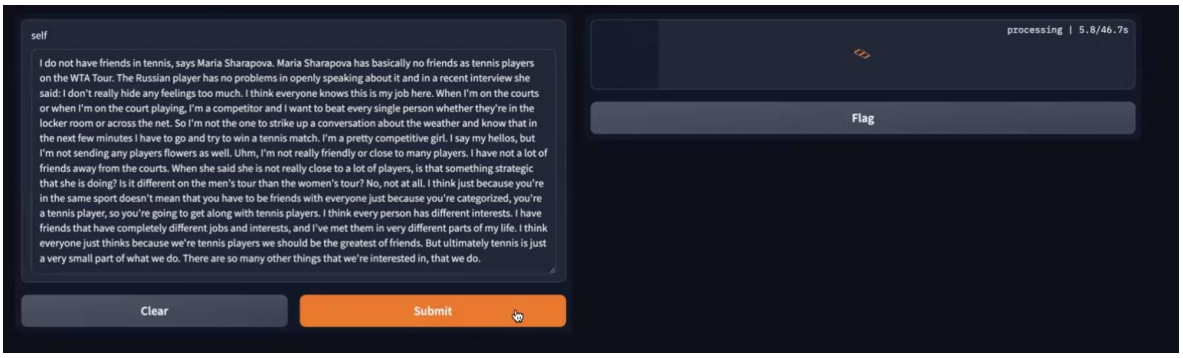
4-rasm. TextRank algoritmini ishlash jarayoni.

- birinchi qadam maqolalardagi barcha matnlarni birlashtiriladi
- matnni alohida jumlalarga ajratiladi
- keyingi bosqichda har bir jumla uchun vektor tasviri (so‘zlarni joylashtirish) topiladi
- jumla vektorlari orasidagi o‘xshashliklar hisoblab chiqiladi va matritsada saqlanadi
- keyin o‘xshashlik matritsasi jumlalar darajasini hisoblash uchun cho‘qqilar sifatida jumlalar va qirralar sifatida o‘xshashlik ballari bilan grafikga aylantiriladi.

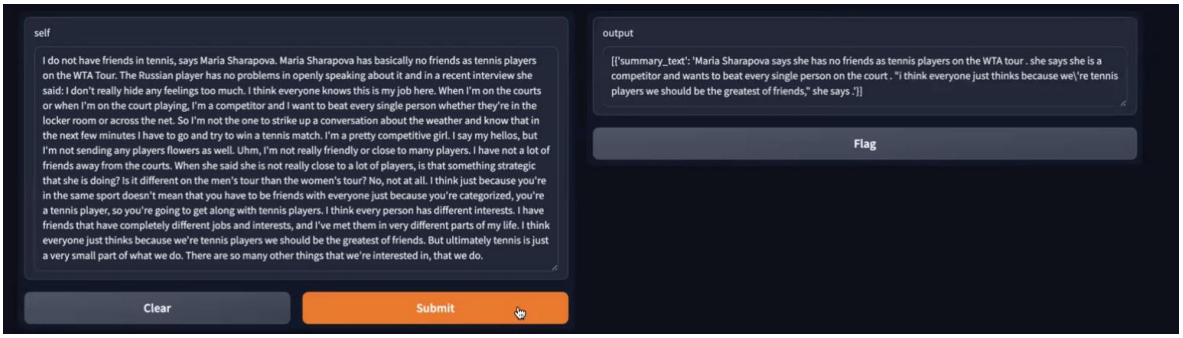
- nihoyat, ma’lum miqdordagi yuqori o‘rinli jumlalar orqali yakuniy xulosa tashkil qilinadi

Matnni Pyhton dasturlash tilida umumiylashtirish jarayonnini quyidagicha

5-rasm. Matnni umumiylashtirih jarayoni interfeysi.



6-rasm. Matnni umumiylashtirih jarayoni.



7-rasm. Matnni umumiylashtirih jarayoni natijasi.

Foydalanilgan adabiyotlar:

1. Abduraxmonova N. Mashina tarjimasining lingvistik ta’minoti(monografiya).-Toshkent:Muharrir, 2018. 66-86 b.
2. Abduraxmonova N. O‘zbek tili elektron korpusining kompyuter modellari(monografiya).-Toshkent:Muharrir, 2021.165 b.
3. Напр., Krovetz R. (2000). Viewing morphology as an inference process. In *Artificial Intelligence*, 118, 277–294.
4. Напр., Sheremetyeva S., Nirenburg S., Nirenburg I. (1996). Generating Patent Claims from Interactive Input. In Proceedings of the 8th International Workshop on Natural Language Generation (Herstmonceux, Sussex, June 1996), 61–70.
5. Goldsmith, J. (2001). Unsupervised Learning of the Morphology of a Natural Language. In *Computational Linguistics*, 27(2), 153–198.
6. Ford A., & Singh R. (1991). Propedeutique Morphologique. *Folia Linguistica*, 25 (3–4), 549–575; Neuvel Sylvain (2002). Whole Word Morphologizer: Expanding the Word-Based Lexicon: A Nonstochastic Computational Approach. In *Brain and Language* 81, 454–463.

7. Neural Machine Translation (seq2seq) Tutorial – Режим доступа: https://www.tensorflow.org/tutorials/seq2seq.
8. Word2vec – Режим доступа: https://ru.wikipedia.org/wiki/Word2vec.
9. Klymenko, N. F. et al. (2014). Morfemno-slovotvirnyj fond ukrajins'koji movy jak doslidnyc'ka ta informacijno-dovidkova systema. In *Klymenko N. F. Vybrani praci*, pages 545–558, Kyiv.
10. Korpus ukrajins'koji movy. Accessible at: http://www.mova.info/corpus.aspx, retrieved 2017-03-15.