QORAQALPOQ TILI UChUN SUBWORD TOKENIZATSIYA ALGORITMLARINING QIYOSIY TAHLILI

Uteuliev Nietbay Uteulievich – *fizika-matematika fanlari doktori, professor*
utewlievn@mail.ru
Abdalieva Gulistan Rafikovna – *filologiya fanlari doktori (DSc)*
Asenbaev Nurbek Adilbaevich – *assistent o'qituvchi*
Orazbaev Jaqsilik Abdimuratovich – *assistent o'qituvchi*
Nukus davlat texnika universiteti
Asenbaeva Dalloris Adilbay qizi – *tayanch doktorant*
dallorisasenbaeva@gmail.com
Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti

СРАВНИТЕЛЬНЫЙ АНАЛИЗ АЛГОРИТМОВ СУБСЛОВНОЙ ТОКЕНИЗАЦИИ ДЛЯ КАРАКАЛПАКСКОГО ЯЗЫКА

Утеулиев Ниетбай Утеулиевич – *доктор физико-математических наук, профессор*
Абдалиева Гулистан Рафиковна – *доктор филологических наук (DSc)*
Асенбаев Нурбек Адилбаевич – *ассистент преподаватель*
Оразбаев Жаксылык Абдимуратович – *ассистент преподаватель*
Нукусский государственный технический университет
Асенбаева Даллорис Адилбаевна – *базовый докторант*
Ташкентский университет информационных технологий имени Мухаммада аль-Хоразмий

COMPARATIVE ANALYSIS OF SUBWORD TOKENIZATION ALGORITHMS FOR THE KARAKALPAK LANGUAGE

Uteuliev Nietbay Uteulievich – *Doctor of Physical and Mathematical Sciences, Professor*
Abdalieva Gulistan Rafikovna – *Doctor of Philological Sciences (DSc)*
Asenbaev Nurbek Adilbaevich – *Assistant Lecturer*
Orazbaev Jaqsilik Abdimuratovich – *Assistant Lecturer*
Nukus State Technical University
Asenbaeva Dalloris Adilbayena – *PhD Student*
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi

Tayanch so'zlar: qoraqalpoq tili, tokenizatsiya, subword, BPE, WordPiece, Unigram, Byte-Level BPE.

Rezyume. Maqolada qoraqalpoq tili uchun BPE + Metaspace, WordPiece, Unigram + Metaspace va Byte-Level BPE algoritmlarining 12 konfiguratsiyasi taqqoslandi. Tokenizatorlar 22 913 ta train yozuvida o'qitilib, 234 validation hujjatida baholandi. 8K lug'atda Unigram 1,443 token/so'z, 16K va 24K lug'atlarda BPE 1,279 va 1,205 natija berdi. Ketma-ketlik uzunligi, lug'atdan foydalanish va embedding hajmi bo'yicha BPE + Metaspace 16K muvozanatli nomzod sifatida tanlandi.

Ключевые слова: каракалпакский язык, токенизация, субсловные единицы, BPE, WordPiece, Unigram, Byte-Level BPE.

Резюме. Сравнены 12 конфигураций BPE + Metaspace, WordPiece, Unigram + Metaspace и Byte-Level BPE для каракалпакского языка. Токенизаторы обучались на 22 913 записях и оценивались на 234 документах. Unigram показал 1,443 токена на слово при 8K, а BPE — 1,279 и 1,205 при 16K и 24K. С учётом длины последовательности, использования словаря и размера embedding-слоя BPE + Metaspace 16K выбран как сбалансированный кандидат.

Key words: Karakalpak language, tokenization, subword, BPE, WordPiece, Unigram, Byte-Level BPE.

Summary. Twelve BPE, WordPiece, Unigram, and Byte-Level BPE configurations were compared for Karakalpak. The tokenizers were trained on 22,913 records and evaluated on 234 documents. Unigram achieved 1.443 tokens per word at 8K; BPE achieved 1.279 and 1.205 at 16K and 24K. Considering sequence length, vocabulary utilization, and embedding size, BPE + Metaspace 16K was selected as the balanced candidate.

Kirish. Transformer modellarida matn token identifikatorlari ketma-ketligi sifatida qayta ishlanadi. Tokenizator tanlovi ketma-ketlik uzunligi, lug'at va embedding hajmi, noma'lum belgilarni qamrab olish hamda hisoblash xarajatlariga bevosita ta'sir qiladi. Subword yondashuv so'z darajasidagi ochiq lug'at muammosi bilan belgi darajasidagi uzun ketma-ketlik o'rtasida muvozanat yaratadi. Sennrich, Haddow va Birch BPE yordamida kam uchraydigan so'zlarni kichik birliklar orqali ifodalash samaradorligini ko'rsatgan [1]. Kudo Unigram modelini [2], Kudo va Richardson esa xom matndan o'qitiladigan SentencePiece tizimini taklif qilgan [3]. WordPiece BERT modellarida [4], Byte-Level BPE esa 31enerative modellar oilasida keng qo'llanadi [5].

Qoraqalpoq tili 31enerative31ve bo'lgani sababli negizga qo'shimchalar ketma-ket qo'shilib, ko'plab so'z shakllarini hosil qiladi. Cheklangan lug'atda ularning barchasini alohida token sifatida saqlash samarasiz, haddan tashqari mayda segmentatsiya esa kontekst oynasidan foydalanishni pasaytiradi. Lotin yozuvidagi Á, Ǵ, Í, Ń, Ó, Ú va I kabi belgilar matnni yo'qotishsiz qayta tiklash talabini ham kuchaytiradi. Kam resursli tillarning NLPda yetarli namoyon bo'lmasligi [6] hamda maxsus monolingual tokenizatorning model natijalariga ta'siri [7] qoraqalpoq tili uchun alohida tajriba zarurligini ko'rsatadi.

Ushbu tadqiqotning maqsadi qoraqalpoq tilidagi yangiliklar korpusi asosida BPE + Metaspace, WordPiece, Unigram + Metaspace va Byte-Level BPE algoritmlarini 8 000, 16 000 va 24 000 birlikli lug'atlarda bir xil sharoitda taqqoslash hamda 31enera 31enerative til modeli uchun dastlabki nomzodlarni aniqlashdan iborat. O'rganilgan ilmiy manbalarda qoraqalpoq tili uchun aynan shu to'rtta

algoritm va uchta lug‘at hajmini yagona protokolda baholagan ish aniqlanmadi. Shu sababli tadqiqot qoraqalpoq tili tokenizatorlari uchun qayta tiklanadigan boshlang‘ich benchmarkni taqdim etadi.

Metodologiya. Tajribalarda Dalloris/karakalpak-news-corpus-private-v1 korpusining 1aed9f88432139e66f46f470948f7ab0ea60f70e commit versiyasi ishlatildi. Lug‘at faqat 22 913 ta train yozuvida o‘rganildi; 234 hujjat va 70 179 so‘zdan iborat validation qismi baholash uchun qo‘llandi. Test qismi ishlatilmadi. Matnlar NFC shakliga normallashtirildi, katta-kichik harflar saqlandi. Minimal chastota 2, maksimal token uzunligi 32, maxsus tokenlar <pad>, <unk>, <bos> va <eos> etib belgilandi.

Baholashda bir so‘zga to‘g‘ri keladigan tokenlar soni, noma’lum tokenlar, exact round-trip, lug‘atdan foydalanish, hujjat tokenlarining 95-foizli chegarasi, kodlash tezligi va 512 o‘lchamli embedding parametrlari hisoblandi. Generativ model uchun UNK = 0 va exact round-trip kamida 99,9 foiz qat’iy filtr sifatida olindi. Keyin lug‘at hajmi va token/so‘z bo‘yicha Pareto chegarasi aniqlandi. Yashirin vaznlangan umumiy ball qo‘llanmadi, chunki uning og‘irliklari subyektiv bo‘lishi mumkin.

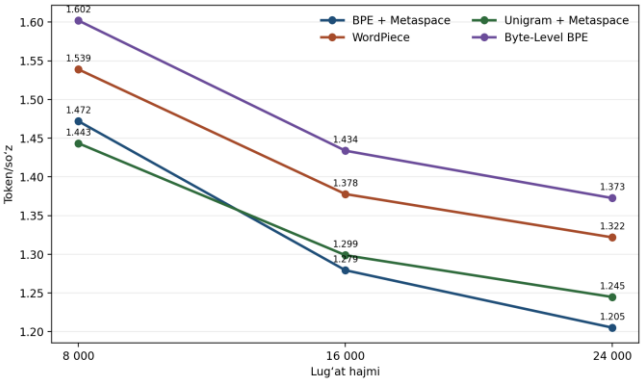
Natijalar va muhokama. To‘rtta algoritm va uchta lug‘at hajmi bo‘yicha jami 12 ta tajriba bajarildi. Barcha konfiguratsiyalarda noma’lum tokenlar soni nol bo‘ldi. BPE + Metaspace, Unigram + Metaspace va Byte-Level BPE matnni 100 foiz aynan tikladi. WordPiece uchun round-trip 0,427 foiz bo‘lib, bu noma’lum token emas, BertPreTokenizer va decoderning bo‘shliq hamda tinish belgilarini normallashtirish xususiyati bilan izohlanadi. Shu bois WordPiece masked model uchun rad etilmaydi, lekin qo‘shimcha sozlashsiz causal modelning asosiy nomzodi sifatida olinmadi.

1-jadval. Tokenizatorlarni validation qismida taqqoslash natijalari

Usul	Vocab	O‘qitish, s	Token/so‘z	Round-trip, %	Vocab, %	p95 token	Matn/s
BPE + Metaspace	8K	39.85	1.472	100.000	89.61	1643.3	1122.5
WordPiece	8K	23.75	1.539	0.427	83.77	1695.7	959.1
Unigram + Metaspace	8K	250.56	1.443	100.000	87.91	1591.0	905.8
Byte-Level BPE	8K	27.04	1.602	100.000	88.41	1720.3	807.3
BPE + Metaspace	16K	22.76	1.279	100.000	75.08	1428.7	1106.4
WordPiece	16K	20.89	1.378	0.427	67.70	1510.0	1037.5
Unigram + Metaspace	16K	204.28	1.299	100.000	67.97	1451.4	854.6
Byte-Level BPE	16K	25.98	1.434	100.000	71.02	1544.6	797.2
BPE + Metaspace	24K	22.46	1.205	100.000	60.51	1340.7	1130.6
WordPiece	24K	22.22	1.322	0.427	53.08	1449.1	1025.3
Unigram + Metaspace	24K	199.21	1.245	100.000	53.04	1387.1	832.7
Byte-Level BPE	24K	25.97	1.373	100.000	55.80	1474.3	774.9

8K lug‘atda eng kam fragmentatsiya Unigram + Metaspace’da kuzatildi: 1,443 token/so‘z. BPE + Metaspace 1,472, WordPiece 1,539 va Byte-Level BPE 1,602 natija berdi. Unigramning o‘qitish vaqti 250,56 soniya bo‘lib, BPEdan 6,3 marta uzoq davom etdi. 16K lug‘atda BPE + Metaspace 1,279 token/so‘z bilan birinchi o‘rinni oldi; Unigram 1,299, WordPiece 1,378 va Byte-Level BPE 1,434 natija ko‘rsatdi. BPE 16K uchun lug‘atdan foydalanish 75,08 foiz, p95 hujjat uzunligi 1 428,7 token va embedding hajmi 8,192 million parametrni tashkil etdi.

24K lug‘atda BPE + Metaspace 1,205 token/so‘z bilan eng qisqa ketma-ketlikni yaratdi. Biroq 16K dan 24K ga o‘tish fragmentatsiyani faqat 5,80 foiz kamaytirdi, embedding parametrlarini 50 foiz oshirdi va lug‘atdan foydalanishni 75,08 foizdan 60,51 foizga tushirdi. Demak, eng past token/so‘z qiymati umumiy jihatdan eng maqbul konfiguratsiyani avtomatik belgilamaydi.



1-rasm. Lug‘at hajmi va bir so‘zga to‘g‘ri keladigan tokenlar soni

Lug‘at kattalashgan sari barcha algoritmlarda fragmentatsiya kamaydi. 8K dan 24K ga o‘tishda token/so‘z BPE uchun 18,12 foiz, Byte-Level BPE uchun 14,32 foiz, WordPiece uchun 14,11 foiz va Unigram uchun 13,76 foiz qisqardi. Shu bilan birga 24K konfiguratsiyalarida validationda lug‘atdan foydalanish 53,04–60,51 foizgacha pasaydi. BPE mazkur korpusda ketma-ketlikni qisqartirish bo‘yicha eng katta foydani berdi.

Generativ filtr va Pareto tahlili uchta nomzodni ajratdi: Unigram + Metaspace 8K, BPE + Metaspace 16K va BPE + Metaspace 24K. Unigram 8K 4,096 million embedding parametri bilan eng tejamkor; BPE 24K eng qisqa ketma-ketlikni berdi; BPE 16K esa ular orasidagi muvozanatni ta’minladi. BPE 16K 24K variantga qaraganda 4,096 million kam embedding parametrini talab qiladi, token/so‘z farqi esa 0,074 ni tashkil etadi. Shu sababli BPE + Metaspace 16K keyingi causal Transformer tajribasi uchun asosiy nomzod sifatida tavsiya qilindi.

2-jadval. Pareto chegarasidagi dastlabki nomzodlar

Konfigu-ratsiya	Token/so‘z	Vocab, %	p95 token	Embedding, mln	Asosiy afzallik
Unigram + Metaspace 8K	1,443	87,91	1 591,0	4,096	Kichik embedding
BPE + Metaspace 16K	1,279	75,08	1 428,7	8,192	Muvozanatli
BPE + Metaspace 24K	1,205	60,51	1 340,7	12,288	Eng qisqa ketma-ketlik

Olingan natijalar maxsus til tokenizatori va lug‘at qamrovi muhimligi haqidagi Rust va hammualliflar xulosasiga mos keladi [7]. BPEning 16K va 24K lug‘atlarda kichikroq fragmentatsiya berishi uning kam uchraydigan birliklarni korpus chastotasiga ko‘ra bosqichma-bosqich birlashtirish xususiyati bilan bog‘liq. Unigram 8Kning tejamkor natijasi esa kichik hisoblash budjetida muqobil konfiguratsiya sifatida ahamiyatli. Biroq bu taqqoslash algoritmlarning til modeli sifatidagi ustunligini emas, faqat keyingi nazoratli tajriba uchun asoslangan nomzodlarni ko‘rsatadi.

Tadqiqot faqat ichki tokenizer metrikalariga va yangiliklar domeniga tayangan. Morfema chegaralari, kirill yozuvi, badiiy matnlar hamda til modeli sifati tekshirilmadi. Ichki fragmentatsiya downstream sifatni har doim to‘liq bashorat qilmasligi sababli [8], yakuniy tanlov uchun Pareto nomzodlari bilan bir xil arxitekturali til modelini o‘qitish zarur.

Xulosa. Qoraqalpoq tili uchun to‘rtta subword algoritmining 12 konfiguratsiyasi bir xil protokolda baholandi. Barcha tajribalarda UNK nol bo‘ldi. 8K lug‘atda Unigram + Metaspace, 16K va 24K lug‘atlarda BPE + Metaspace eng kam fragmentatsiyani ko‘rsatdi. 24K BPE eng qisqa ketma-ketlikni bergan bo‘lsa-da, embedding hajmi va lug‘atdan foydalanish hisobga olinganda BPE + Metaspace 16K muvozanatli tanlov bo‘ldi. Keyingi bosqichda Unigram 8K, BPE 16K va BPE 24K bir xil kichik Transformer modelida loss, bits per character va generatsiya sifati bo‘yicha tekshirilishi lozim.

Adabiyotlar

1. Sennrich R., Haddow B., Birch A. Neural Machine Translation of Rare Words with Subword Units. // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – Berlin, Germany: Association for Computational Linguistics, 2016. – P. 1715-1725.

2. Kudo T. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – Melbourne, Australia: Association for Computational Linguistics, 2018. – P. 66-75.

3. Kudo T., Richardson, J. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. – Brussels, Belgium: Association for Computational Linguistics, 2018. – P. 66-71.

4. Devlin J., Chang, M.-W., Lee, K., Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. – Minneapolis, Minnesota: Association for Computational Linguistics, 2019. – P. 4171-4186.

5. Radford A., Wu J., Child R., Luan D., Amodei D., Sutskever I. Language Models are Unsupervised Multitask Learners. – OpenAI Technical Report. – OpenAI, 2019.

6. Joshi P., Santy S., Budhiraja A., Bali K., Choudhury M. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. – Association for Computational Linguistics, 2020. – P. 6282-6293.

7. Rust P., Pfeiffer J., Vulić I., Ruder S., Gurevych I. How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models. // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). – Association for Computational Linguistics, 2021. – P. 3118-3135.

8. Ali M., Fromm M., Thellmann K., et al. Tokenizer Choice For LLM Training: Negligible or Crucial? // Findings of the Association for Computational Linguistics: NAACL 2024. – Mexico City, Mexico: Association for Computational Linguistics, 2024. – P. 3907-3924.