MATNDAGI TINISH BELGILARINI TIKLASH MUAMMOLARI VA YECHIMLARI

Ochilov Mannon Musinovich
ochilov.mannon@mail.ru
Texnika fanlari bo‘yicha falsafa doktori PhD
TATU dotsenti
Jurayev Dilshod Boymuradovich
dilsamtuit@gmail.com
TATU tayanch doktoranti
Narzullayev Oybek Otabek o‘g‘li
oybeknarzullaev99@gmail.com
TATU magistranti

Annotatsiya. Ushbu maqolada matndagi tinish belgilarini tiklash muammolari va yechimlari haqida so‘z yuritiladi. Maqolada shuningdek, matnlarda tinish belgilarini tiklashning turli usullari va yondashuvlarining afzallik va kamchiliklari ko‘rib o‘tiladi.

Abstract. This article discusses problems and solutions for restoring punctuation marks in text. The article also reviews the advantages and disadvantages of various methods and approaches to punctuation in texts.

Аннотация. В этой статье рассматриваются проблемы и решения по восстановлению знаков препинания в тексте. Также в статье рассматриваются преимущества и недостатки различных методов и подходов к пунктуации в текстах.

Kalit so‘zlar: *Tinish belgilar, statistik usullar, mashinali o‘qitish algoritmlari, neyron tarmoqlar, RNN, BiLSTM, Transformer modellari.*

Bugunga kelib o‘zbek tili bo‘yicha qator ilmiy ishlar olib borilgan bo‘lib, ularning aksariyati Tabiiy tilni qayta ishlash(NLP)ga qaratilgan bo‘lsa-da, ularning orasida o‘zbek tili matnlaridagi mavjud imlo va grammatik xatoliklarni tuzatishga qaratilgani juda kam. Biz ushbu maqolada matndagi turli xildagi xatoliklarni tuzatish bo‘yicha yuzaga keladigan muammolar, mavjud usullar tahlili va yechimlari to‘g‘risida to‘xtalib o‘tamiz. Matndagi xatoliklarni tuzatish muammolari orasida tinish belgilarni qo‘yish muhim ahamiyatga ega. Tinish belgilar bo‘yicha ikki turdagi masala ko‘riladi. Birinchisi, matnda mavjud tinish belgilarni to‘g‘ri qo‘yilganligini tekshirish. Bu turdagi masalalar juda ko‘p uchraydi. Ikkinchisi esa, matnda tinish belgilari umuman mavjud emas vaziyatlarda ularni to‘g‘ri qo‘yish. Keltirilgan ikkinchi masala odatda nutqni tanish tizimlari tomonidan hosil qilingan matnlarda uchraydi. Matndagi tinish belgilarini tiklashning qator muammo va vazifalarini keltirishimiz mumkin. Bu vazifalar murakkab bo‘lib, tinish belgilarining to‘g‘ri joylashishini aniqlash uchun tilning konteksti, semantikasi va grammatik

qoidalarini tushunishi kerak. Bu esa bir qator muammolarni keltirib chiqaradi. Tinish belgilarini tiklash quyidagi vazifalarni o‘z ichiga oladi:

1. Nuqta va vergul qo‘yish: jumlalar yoki alohida bandlarni ko‘rsatish uchun nuqta va vergullarni to‘g‘ri joylashtirishni ta’minlash.
2. So‘roq va undov belgilari qo‘yish: mo‘ljallangan ohangni bildirish uchun savollarga so‘roq va buyruq va his-hayajonlar uchun undov belgilarini qo‘shish.
3. Qo‘shtirnoq belgilari: so‘zlovchining to‘g‘ridan-to‘g‘ri gapiga nisbatan yoki ba’zi terminlarni matnda ajratish uchun qo‘shtirnoqlarni joylash.
4. Ikki nuqta va nuqta-verguldan foydalanish: tegishli jumla tuzilmalari uchun ikki nuqta va nuqta-vergul o‘rtasidagi farqni aniqlashtirish.
5. Qavslar: qo‘shimcha ma’lumotlarni qamrab olish yoki tarkibni aniqlashtirish uchun qavslardan foydalanishni tuzatish.
6. Chiziqchalarni tuzatish: urg‘u yoki uzilish hamda qo‘shma so‘zlar uchun chiziqchalardan to‘g‘ri foydalanish.
7. Tutuq belgisini to‘g‘ri joylashtirish: ba’zi so‘zlarga tutuq belgilarini qo‘yib ularni to‘g‘irlash.

Ushbu vazifalarni bajarishda o‘z navbatida quyidagi muammolar kelib chiqadi:

- **Matn mazmunidagi noaniqlik**: Ba’zi jumlalar to‘g‘ri tinish belgilarini aniqlash uchun kontekstni talab qilishi mumkin. Ayrim iboralar tushunarsiz bo‘lishi mumkin, bu esa to‘g‘ri tinish belgilarini qo‘yishni qiyinlashtiradi.
- **Matn tarkibining qisman yo‘qolganligi**: Tinish belgilarini tiklashga yetishmayotgan ma’lumotlar yoki to‘liq bo‘lmagan jumlalar to‘sqinlik qilishi mumkin. Yetarli kontekstsiz, mo‘ljallangan tinish belgilarini aniqlash qiyin kechadi.
- **Turli xil yozish uslublari**: Turli yozuvchilar tinish belgilaridan boshqacha foydalanishi mumkin va bitta yozish uslubiga o‘rgatilgan model boshqa yozish uslubi bilan yozilgan matnlarga moslashtirishni talab etadi. Bu esa o‘z navbatida tinish belgilarini qo‘yishda katta qiyinchilik tug‘diradi.
- **Murakkab jumlalar**: Bir nechta gap yoki qavs ichidagi iboralar bilan murakkab jumlalar tinish belgilarini to‘g‘ri belgilash qiyin bo‘lishi mumkin.
- **Domenga xos muammolar**: Ixtisoslashgan domenlarda (masalan, yuridik, ilmiy) tinish belgilarini tiklash noyob tinish belgilari bilan ishlash uchun domenga xos bilimlarni talab qilishi mumkin.
- **Ko‘p tillilik muammosi**: Tinish belgilari turli tillarda farqlanadi va bir tilda o‘rgatilgan model boshqa tilda turli tinish belgilari bilan qo‘llanilganda unchalik yaxshi ishlamasligi mumkin.

Matnlarda tinish belgilarini tiklashning turlicha usullari mavjud bo‘lib ularni 4 ta umumiy guruhga ajratib olishimiz mumkin:

1. Qoidalarga asoslangan tinish belgilarini tiklash algoritmlari;
2. Statistik modellar asoslangan yondashuv;
3. Mashinali o‘qitishga asoslangan yondashuv;
4. Chuqur o‘qitish neyron tarmog‘iga asoslangan yondashuv.

Quyida ushbu usullarning har biriga batafsil to‘xtalamiz.

Qoidalarga asoslangan tinish belgilarini tiklash algoritmlari. Ushbu usul haqida [Păiş, Tufiş, 2022] ishda batafsil ma’lumot va turdosh ilmiy izlanishlarni keltirib o‘tishgan. Bu usulning boshqa usullardan farqi shundagi undagi algoritmlar an’anaviy dasturlash tamoyillariga muvofiq amalga oshiriladi. Bunda algoritmni tuzib chiqish to‘laqonli dasturchining ishi hisoblanadi. Keling, shunday algoritmlardan bittasini sodda misol bilan tuzib ko‘raylik. Aytaylik bizga sodda, faqat birinchi shaxsda yozilgan gaplardan iborat, birorta tinish belgisi yo‘q, hamda faqat kichik harflardan iborat matn berilsin:

“men talabaman bugun uyga boramanmi men topshiriqlarning hammasini bajardim”

Matn mazmunini bir qarashda sezish mumkin. Unda tinish belgilarini tiklashni amalga oshirish uchun sodda algoritm yozish yetarli. Ushbu algorimda bor yo‘g‘i ikkita qoida bor:

1. Agar so‘z oxiri “man” bilan tugasa shu so‘zning oxiriga “nuqta” qo‘y va shu so‘zdan keyingi so‘zni bosh harf bilan boshla;
2. Agar so‘z oxiri “mi” bilan tugasa shu so‘zning oxiriga “so‘roq belgisi” qo‘y va shu so‘zdan keyingi so‘zni bosh harf bilan boshla;

Algoritmga yana boshqa qoidalarni kiritish mumkin. Tuzilgan algoritm faqat yuqoridagi misolgagina mutlaq to‘g‘ri ishlaydi. Ammo boshqa tuzilmadagi matnlari uchun bu algoritm to‘liq ishlamaydi. Ushbu algoritmlar bugungi kunda tinish belgilarini tiklashda keng qo‘llanilmaydi. Chunki ushbu usulda bir qator kamchiliklar bor:

1. Cheklangan moslashuvchanlik: Qoidalarga asoslangan algoritmlar oldindan belgilangan qoidalarga tayanadi. Bu ularni turli xil yozish uslublari yoki kutilmagan o‘zgarishlarga kamroq moslashtiradi. Ular o‘rnatilgan qoidalardan chetga chiqadigan nostandart til yoki ijodiy yozuv bilan ishlashda qiyinchliklar keltirib chiqarishi mumkin.
2. Noaniqlik bilan bog‘liq qiyinchilik: Qoidalar tilning barcha xususiyatlarini qamrab olmasligi mumkin, bu esa noto‘g‘ri tinish belgilarini tiklashga olib keladi.
3. Qoidalarning murakkabligi: Tinish belgilarini tiklash uchun keng qamrovli qoidalarni ishlab chiqish murakkab va ko‘p mehnat talab qilishi mumkin. Tabiiy tilning barcha nozik tomonlarini, ayniqsa, yozish uslublari va kontekstlarining xilma-xilligini hisobga olgan holda, qoidalar to‘plamida butun tilni qamrab olish qiyin bo‘lishi mumkin.
4. Tilga bog‘liqlik: Qoidalarga asoslangan algoritmlar ko‘pincha tilga bog‘liq, ya’ni turli tillar uchun alohida qoidalar to‘plami ishlab chiqilishi kerak. Bu resurs talab qilishi va ko‘p tilli ssenariylarni boshqarish uchun yaxshi miqyosda bo‘lmasligi mumkin.
5. Kengayuvchanlik muammolari: Tilning murakkabligi oshgani sayin, barcha mumkin bo‘lgan holatlarni qamrab olish uchun talab qilinadigan qoidalar

soni ham oshadi. Bu masshtablilik muammolariga olib kelishi mumkin, chunki ma’lumotlar to‘plamining hajmi yoki tilning murakkabligi kengayganligi sababli qoidalarga asoslangan tizimni saqlashni qiyinlashtiradi.

Qoidalarga asoslangan tinish belgilarini tiklash algoritmlaridan farqli o‘laroq bugungi kunda mashinali o‘qitish modellari hamda neyron tarmoq modellari o‘zlarining kengayuvchanligi hamda samarali ekanligi bilan ancha ilg‘or hisoblanadi.

Statistik modellarga asoslangan yondashuv. Statistik modellar - bu ma’lumotlar ichidagi munosabatlarni tavsiflash va ifodalash uchun statistik usullardan foydalanadigan matematik modellardir. Statistik modellar yordamida tinish belgilarini tiklash qoidalarga asoslangan algoritmlarga nisbadan samaraliroq hisoblanadi. Statistik modellar ma’lumotlardan o‘rganishi va turli lingvistik kontekstlarga moslashishi, so‘zlar va tinish belgilari o‘rtasidagi bog‘liqlikni ko‘rsatib bera olishi mumkin. Ular yozish uslublari va noan’anaviy tinish belgilaridan foydalanishdagi o‘zgarishlarni yaxshiroq talqin etadi. Chunki ular qoidalarga tayanmasdan, ma’lum bir oldindan mavjud bo‘lgan tinish belgili xatosiz ma’lumotlardan o‘rganadilar. Statistik modellar turli sohalarda, jumladan, fan, muhandislik, iqtisod va ijtimoiy fanlarda keng qo‘llaniladi.

Ko‘pchilik ishlarda tinish belgilarini tiklashda Yashirin Markov modellari (HMMs – Hidden Markov models) va Shartli tasodifiy maydonlar (CRFs – Conditional Random Fields) kabi statistik usullardan keng foydalanilgan [Boros va b., 2017].

Ammo statistik usullar ham qoidalarga asoslangan algoritmlardan samarali bo‘lishiga qaramay bir qator kamchiliklarga ega.

1. Matn kontekstni tushunishda cheklanganlik: Yashirin Markov modellari va shartli tasodifiy maydonlarni o‘z ichiga olgan statistik modellar uzoq muddatli bog‘liqliklarni olish va murakkab kontekstual jumlalarni tushunishda cheklovlarga ega bo‘lishi mumkin. Tinish belgilarini tiklash ko‘pincha kontekstni chuqur tushunishni talab qiladi, ayniqsa murakkab jumla tuzilmalarida.
2. O‘zgaruvchanlikni boshqarish qiyinligi: Statistik modellar yozish uslublaridagi o‘zgarishlar, norasmiy til yoki noan’anaviy tinish belgilaridan foydalanishda qiyinchiliklarga duch kelishi mumkin. Ular ko‘pincha oldindan belgilangan uslublarga tayanadilar va kamroq tuzilgan yoki xilma-xil ma’lumotlarga duch kelganda kutilgan natijani bermasligi mumkin.
3. Katta o‘lchamli ma’lumotlar bilan ishlashda qiyinchilik: Statistik modellarni o‘rgatish, ayniqsa ko‘p sonli parametrlarga ega bo‘lgan modellarni o‘rgatishni amalga oshirish uchun katta resurs kerak bo‘ladi. Katta o‘lchamli ma’lumotlar to‘plamlari bilan ishlash xotira va qayta ishlash talablari nuqtayi-nazaridan qiyinchiliklarga olib kelishi mumkin.
4. Semantik kontekstga kamroq moslashish: Tinish belgilarini tiklash ko‘pincha matnning semantik kontekstini hisobga olishni o‘z ichiga oladi. Statistik

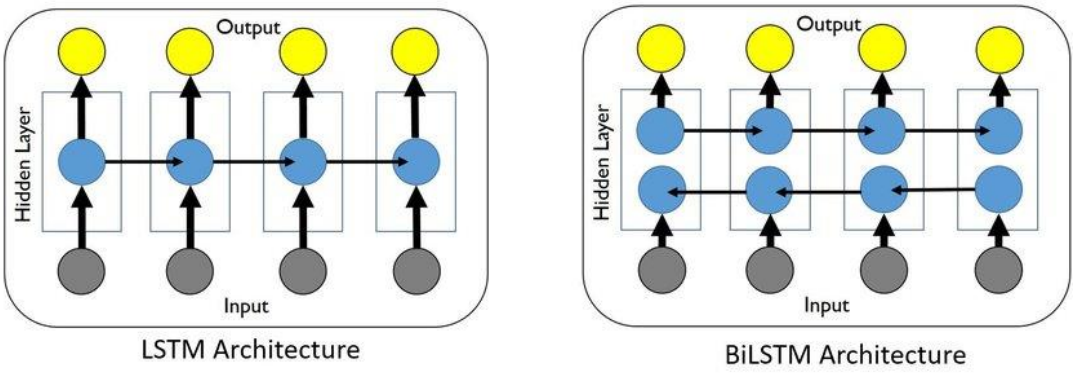
modellar ilg‘or mashinali o‘qitish modellari yoki neyron tarmoqlar bilan solishtirganda semantik ma’lumotni olishda unchalik yaxshi bo‘lmasligi mumkin.

5. Chiziqli bo‘lmagan munosabatlarni olishda qiyinchilik: An’anaviy statistik modellar o‘zgaruvchilar orasidagi chiziqli munosabatlarni nazarda tutadi. Biroq, tinish belgilarini tiklash kabi til vazifalarida munosabatlar chiziqli bo‘lmasligi mumkin va statistik modellar bu murakkab ssenariylarda samarali ishlamasligi mumkin.

Mashinali o‘qitishga asoslangan yondashuv. Mashinali o‘qitish modellari tabiiy tilni qayta ishlash jarayonlarida, jumladan, tinish belgilarini tiklashda ham ishlatilishi mumkin. Qaror daraxtlari (Decision Trees), tasodifiy o‘rmonlar (Random Forecast) va qo‘llab-quvvatlovchi vektor mashinalari (Support Vector Machines) kabi ko‘plab mashina o‘rganish algoritmlari statistik asoslarga egadir. Shuning uchun ham mashinali o‘qitish modellarini statistik modellashtirishning kengaytmasi sifatida ko‘rish mumkin, ammo ular ko‘pincha murakkabroq algoritmlarni va kattaroq ma’lumotlar to‘plamini o‘z ichiga oladi. Biroq bugungi kunda tinish belgilarini tiklash va shu kabi masalalar uchun an’anaviy mashinali o‘qitish modellaridan ko‘ra chuqur o‘qitish neyron tarmoq modellari nisbatan muammoni hal etishda ancha ko‘proq samara beradi. Boshqacha qilib aytganda, an’anaviy mashinali o‘qitish modellari tinish belgilarini tiklash uchun ishlatilgan bo‘lsa-da, ular ilg‘or chuqur o‘rganish modellari bilan solishtirganda til va kontekstual bog‘liqliklarning nozik tomonlarini qamrab olishda sezilarli kamchiliklarga ega.

Chuqur o‘qitish neyron tarmog‘iga asoslangan yondashuv. Neyron tarmoq modellarining kashf etilishi mashinali o‘qitishda yana bir sohaning paydo bo‘lishiga olib keldi. Chuqur o‘qitish neyron tarmoqlarining kengayishi bilan tabiiy tilni qayta ishlash masalalari, jumladan tinish belgilarini qayta ishlash masalalari o‘zining cho‘qqisiga chiqdi. Sohada natijalar sezilarli ravishda ortdi. Tabiiy tilni qayta ishlashda chuqur o‘qitish nayron tarmoqlaridan RNN (Recurrent neural network), CNN (Convolutional neural network) va Transformer larga asoslangan neyron tarmoq modellaridan keng foydalaniladi.

Takroriy neyron tarmoqlari (RNN) ma’lumotlardagi ketma-ket bog‘liqliklarni modellashtirish qobiliyati tufayli tinish belgilarini tiklash vazifalarida ishlatilgan [Tarwani va b., 2017]. Tinish belgilarini tiklash vazifasi berilgan matnning so‘zlari atrofidagi boshqa so‘zlardan kelib chiqib, to‘g‘ri tinish belgilarini bashorat qilishni o‘z ichiga oladi. Ushbu neyron tarmoqlariga mansub uzoq-qisqa muddatli xotira (LSTM - Long Short-Term Memory) hamda ikki yo‘nalishli uzoq-qisqa muddatli xotira (BiLSTM - Bidirectional LSTM) kabi neyron tarmoq arxitekturalarini tinish belgilarini tiklash uchun keng ko‘lamda ishlatilganligini misol keltirishimiz mumkin. LSTMlar uzoq muddatli bog‘liqliklarni olishda ustunlik qiladi, ikki yo‘nalishli LSTMlar esa bu imkoniyatni har ikki yo‘nalishdagi kontekstni hisobga olgan holda kengaytiradi va bu bilan tabiiy tilni qayta ishlashda tinish belgilarini tiklash kabi ikki tomonlama kontekstda muhim bo‘lgan vazifalarni bajarish mumkin bo‘ladi.



1-rasm. LSTM va BiLSTM neyron tarmoq arxitekturalari

Konvolyutsion neyron tarmoqlari (CNN) ko‘proq kompyuter vizualizatsiyasi (Computer vision) vazifalari bilan bog‘liq bo‘lsa-da, ular tabiiy tilni qayta ishlash vazifalarida, shu jumladan tinish belgilarini tiklashda ham qo‘llanilgan [Tundik, Szaszak, 2018]. CNN ma’lum NLP (Natural Language Processing) vazifalarini hal etishda foydalanib kelinadi. Biroq ular matnlarda uzoq muddatli bog‘liqliklarda LSTM yoki Transformer modellari va takrorlanuvchi neyron tarmoq arxitekturalari kabi samarali bo‘lmasligini ta’kidlash kerak.

Transformerlarga asoslangan oldidan o‘qitilgan (pre-trained) til modellari. Ushbu modellari turli xil tabiiy tillarni qayta ishlash vazifalariga, shu jumladan tinish belgilarini tiklashga bugungi kunda keng qo‘llanildi. Transformerlarga asoslangan til modellari tinish belgilarini tiklash va tabiiy tilni qayta ishlash kabi vazifalarda sezilarli darajada o‘sishni amalga oshirdi. Bularga misol tariqasida BERT [Devlin va b., 2019] va RoBERTa [Liu va b., 2019] keltirish mumkin. Bundan tashqari hozirda shunga o‘xshash bir qancha oldindan o‘qitilgan Transformerlarga asoslangan til modellari ishlab chiqarilgan. Ushbu modellar yordamida hozirda ko‘plab matnlardagi tinish belgilarini tiklash masalalari amalga oshirilmoqda. Bularga misol qilib Tanvirul Alam, Akib Khan, Firoj Alamlar Bangla tili uchun o‘zlarining ilmiy ishida dataset yig‘ib bangla tili uchun tinish belgilarini tiklash modelini ishlab chiqishgan hamda uning natijalarini ingliz tilidagi model natijalari bilan solishtirishgan. Eng yaxshi natijalarni (yangiliklar transkripsiyalari dataseti uchun: Precision:87.4%, Recall:86.6%, F1 score:87.0%) XLM-RoBERTa modelida olishganlar [Alam va b., 2020]. Uygar Kurt, Aykut Chayırlar esa o‘zbek tiliga nisbatan yaqin bo‘lgan til turk tili uchun tinish belgilarni tiklashda 1.8 mln token, 146 mingta gapdan iborat korpus shakllantirib, ushbu dataset yordamida Transformerlarga asoslangan BERT, ELECTRA, ConvBERT kabi oldindan o‘qitilingan modellar yordamida yangi til modelini olishdi hamda eng yaxshi natijani (Precision: 0.823%, Recall: 0.859%, F1-score: 0.839%) ELECTRA modelida qayd etishdi [Kurt, Çayır, 2023].

Bundan ko‘rinib turibdiki Transformerlarga asoslangan oldindan o‘qitilgan til modellari hozirgi kunda tinish belgilarini tiklash kabi tabiiy tilni tanib olish

masalalarida eng ilg‘or yondashuv deb hisoblash mumkin. Kelgusida o‘zbek tili uchun ham aynan ushbu yondashuvdan foydalanish maqsadga muvofiqdir.

Foydalanilgan adabiyotlar:

1. Păiş, V., Tufiş, D. Capitalization and punctuation restoration: a survey. Artif Intell Rev 55, 1681–1722 (2022).
2. Liberman, M.Y. (1991). The Trend towards Statistical Models in Natural Language Processing. In: Klein, E., Veltman, F. (eds) Natural Language and Speech. ESPRIT Basic Research Series. Springer, Berlin, Heidelberg.
3. Tiberiu Boros, Stefan Daniel Dumitrescu and Sonia Pipa. Fast and Accurate Decision Trees for Natural Language Processing Tasks. 2017.
4. Kanchan M. Tarwani, Swathi Edem. Survey on Recurrent Neural Network in Natural Language Processing. 2017
5. Arvind T. Mohan, Datta V. Gaitonde. A Deep Learning based Approach to Reduced Order Modeling for Turbulent Flow Control using LSTM Neural Networks.
6. Mate Akos Tundik and Gyorgy Szaszak. Joint Word- and Character-level Embedding CNN-RNN Models for Punctuation Restoration. 2018.
7. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019.
8. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019.
9. Tanvirul Alam, Akib Khan, Firoj Alam. Punctuation Restoration using Transformer Models for High-and Low-Resource Languages. 2020.
10. Uygar Kurt, Aykut Çayır. Transformer Based Punctuation Restoration for Turkish. 2023.