



QISQA MATNLARDA PARAFRAZ ANIQLASH SAMARADORLIGINI OSHIRISH: GRAF-NAZARIY MA'LUMOT KENGAYTIRISH VA KO'P-KASKADLI CHUQUR MODEL YONDASHUVI

Subxonqulov Umidjon

Buxoro Davlat universiteti

subxonqulovumidjon@gmail.com

Dolzarblik

Parafraz aniqlash bir xil mazmunni turli sintaktik shaklda ifodalovchi matn juftliklarini avtomatik tarzda tanib olish vazifasi bo'lib, plagiat aniqlash, takrorlanuvchi savollarni bir-biriga bog'lash va mijozlarga xizmat ko'rsatish tizimlarini yaxshilash kabi ko'plab amaliy sohalarda muhim ahamiyat kasb etadi. Chuqur o'rganish modellari parafraz klassifikatsiyasida yuqori natijalar bergani bilan, ularning samaradorligi katta hajmdagi qo'lda yorliqlangan ma'lumotlar mavjudligiga bevosita bog'liq, biroq bunday annotatsiya jarayoni vaqt va moliyaviy resurs talab qiladigan murakkab jarayondir. Ayniqsa qisqa matnlar (yangilik sarlavhalari, savollar, tvitlar) uchun mavjud yorliqlangan to'plamlar hajman cheklangan bo'lib, bu esa chuqur modellarning umumlashtirish qobiliyatini pasaytiradi. Shu bilan birga, qisqa va norasmiy (shovqinli) matnlar uchun mo'ljallangan yechimlar odatda toza matn (masalan, yangilik sarlavhalari) uchun samarasiz bo'lib chiqadi, bu esa bir vaqtning o'zida ham toza, ham shovqinli matnlarda barqaror ishlaydigan universal yondashuvlar ishlab chiqish zaruratini keltirib chiqaradi.

Tadqiqot maqsadi

Tadqiqotning maqsadi mavjud parafraz va parafraz-emas annotatsiyalaridan qo'shimcha, ishonchli yangi annotatsiyalarni inson aralashuvisiz, sof matematik-mantiqiy asosda generatsiya qiluvchi ma'lumot kengaytirish strategiyasini ishlab





chiqish, shuningdek turli xususiyat o'rganish istiqbollarini (CNN, LSTM, diqqat mexanizmi bilan va bilan sizsiz) birlashtiruvchi, ham toza, ham shovqinli qisqa matnlarda barqaror ishlaydigan chuqur klassifikatsiya arxitekturasini taklif etishdan iborat.

Metodologiya

Taklif etilgan yondashuvning ilmiy yangiligi shundaki, parafraz va parafraz-emaslik tushunchalari matnlar to'plami ustidagi ikkilik munosabat (binary relation) sifatida formal to'plamlar nazariyasi doirasida ta'riflanadi va graf nazariyasi vositalari orqali qayta ishlanadi. Annotatsiyalangan korpus yo'naltirilmagan graf sifatida modellashtiriladi, bunda tugunlar matnlarni, "parafraz" yorlig'iga ega qirralar esa ularning semantik ekvivalentligini ifodalaydi. Ushbu grafda refleksivlik (har bir matn o'zining parafrazi), simmetriya (parafraz munosabatining ikki yo'nalishliligi) va oldindan belgilangan K tartibli tranzitiv kengaytirish (agar A matn B ga, B esa C ga parafraz bo'lsa, ma'lum shartlar ostida A ham C ga parafraz deb hisoblanadi) izchil qo'llaniladi. Tranzitiv yopilish natijasida graf o'zaro bog'langan to'liq kichik graflar (klik) to'plamiga aylanadi, bunda har bir klik yagona semantik tushunchani ifodalaydi. Parafraz-emas munosabatlari esa mavjud parafraz va parafraz-emas annotatsiyalarini o'zaro solishtirish orqali, jumladan turli kliklar orasidagi to'liq ikki bo'lakli graf qurish yo'li bilan kengaytiriladi. Algoritm shuningdek asl korpusdagi ziddiyatli (bir vaqtning o'zida ham parafraz, ham parafraz-emas deb belgilangan) annotatsiyalarni avtomatik aniqlash va bartaraf etish imkonini beradi. Klassifikatsiya bosqichida uchta mustaqil, parallel ishlaydigan xususiyat o'rganuvchi "kaskad" taklif etiladi: yumshoq diqqat mexanizmi bilan mustahkamlangan konvolyutsion tarmoq, oddiy rekurrent (LSTM) tarmoq, va yumshoq diqqatga ega LSTM tarmog'i. Har bir kaskad matn juftligini mustaqil tarzda parafraz yoki parafraz-emas sifatida tasniflashga o'rgatiladi, so'ngra ularning yakuniy qatlamidan olingan vakillashtiruvlar qo'lda ishlab chiqilgan lingvistik xususiyatlar (leksik o'xshashlik, n-gram ustma-ustlik,



nomlangan mavjudotlar mosligi va boshqalar) bilan birga diskriminator tarmoqqa uzatiladi, u esa yakuniy binar qarorni chiqaradi. Bunday "keng va chuqur" arxitektura turli xil semantik signallarni birlashtirish orqali yagona modelga qaraganda barqarorroq natija berishga xizmat qiladi. Yondashuv uchta mezon toza matnli Quora va MSRP hamda shovqinli matnli SemEval-2015 Twitter parafraz to'plamlarida sinovdan o'tkazilgan.

Asosiy natijalar

Ma'lumot kengaytirish strategiyasi barcha uchta to'plamda annotatsiyalar sonini sezilarli darajada oshirdi — masalan, Quora to'plamida parafraz annotatsiyalari 139 mingdan 964 mingdan ziyodgacha, SemEval to'plamida esa taxminan 4 mingdan 108 mingdan ortiqqacha ko'paydi. Muhimi, kengaytirish jarayonining o'zi asl annotatsiyalardagi xatolarni aniqlashda samarali vosita bo'lib chiqdi: masalan, Quora to'plamida tranzitiv kengaytirish orqali 214 ta ziddiyatli holat aniqlanib, ularning aksariyatida generatsiya qilingan yorliq asl (xato) yorliqdan ko'ra to'g'riroq ekanligi tasdiqlandi. Klassifikatsiya natijalariga ko'ra, refleksivlik, simmetriya va tranzitiv kengaytirish birgalikda qo'llanilganda barcha uch to'plamda eng yuqori F1-ko'rsatkichga erishildi, biroq parafraz-emas munosabatlarining tranzitiv kengaytirilishi (ayniqsa asl annotatsiyada xatolar mavjud bo'lganda) natijani yomonlashtirishi mumkinligi aniqlandi, chunki bitta noto'g'ri parafraz-emas belgisi ikki katta semantik klaster orasida ko'plab noto'g'ri bog'lanishlarni keltirib chiqaradi.

Lingvistik xususiyatlarning qo'shimcha ta'siri ma'lumotlar to'plamining hajmi va shovqinlilik darajasiga bog'liq bo'lib chiqdi: katta hajmli Quora to'plamida bu xususiyatlar sezilarli foyda bermagan, aksincha kichik hajmli MSRP va shovqinli SemEval to'plamlarida ularning qo'shilishi klassifikatsiya sifatini sezilarli darajada oshirdi. Shovqinli matnlar uchun to'liq tranzitiv yopilish ($K = \text{cheksiz}$) o'rniga cheklangan tartibli (masalan, $K = 1$) kengaytirish yanada barqaror natija berishi aniqlandi, chunki graf bo'yicha uzoq masofadagi



tranzitiv bog'lanishlar shovqinli, norasmiy matnlarda semantik ma'noning asta-sekin siljib ketishiga olib keladi. Yakuniy model barcha uchta ochiq mezon (benchmark) to'plamida - Quora, MSRP va SemEval - o'sha davrdagi eng yaxshi natijalarni yangilab, mos ravishda 90,3%, 84,8% va 75,4% F1/aniqlik ko'rsatkichlariga erishdi, bunda hech qanday og'ir ansambl yondashuvi yoki hisoblash jihatidan qimmat belgi darajasidagi xususiyatlarga murojaat qilinmagan.

Xulosa

O'tkazilgan tadqiqot parafraz aniqlash muammosini ikki yo'nalishda nazariy jihatdan puxta asoslangan graf-nazariy ma'lumot kengaytirish strategiyasi va bir nechta o'rganish istiqbollarini birlashtiruvchi keng-chuqur klassifikatsiya arxitekturasini orqali samarali hal qiladi. Ilmiy yangilik parafraz munosabatini to'plamlar va graf nazariyasi asosida rasmiylashtirish, undan annotatsiya xatolarini avtomatik aniqlash uchun ham foydalanish, hamda tranzitivlik darajasini boshqaruvchi K parametri orqali kengaytirish jarayonini matn turiga (toza yoki shovqinli) moslashtirish imkoniyatida namoyon bo'ladi. Amaliy ahamiyati shundaki, taklif etilgan strategiya inson ishtirokisiz, hisoblash jihatidan arzon tarzda mavjud kichik hajmdagi annotatsiyalangan to'plamlarni sezilarli kengaytirish imkonini beradi, bu esa plagiat aniqlash, savol-javob tizimlari va mijozlarga xizmat ko'rsatish platformalarida amaliy qo'llanilishi mumkin. Kelgusi tadqiqotlar ma'lumot kengaytirish jarayonida yuzaga keladigan ziddiyat va xatolarni yanada chuqurroq tahlil qilish hamda ularni avtomatik tuzatish mexanizmlarini ishlab chiqishga qaratilishi maqsadga muvofiqdir.

FOYDALANILGAN ADABIYOTLAR

1. Shakeel, M. H., Karim, A., & Khan, I. (2020). A multi-cascaded model with data augmentation for enhanced paraphrase detection in short texts.



Information Processing & Management, 57(3), 102204.
<https://doi.org/10.1016/j.ipm.2020.102204>

2. Agarwal, B., Ramampiaro, H., Langseth, H., & Ruocco, M. (2018). A deep network model for paraphrase detection in short text messages. *Information Processing & Management*, 54(6), 922–937.
3. Wang, Z., Hamza, W., & Florian, R. (2017). Bilateral multi-perspective matching for natural language sentences. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 4144–4150.
4. Tomar, G. S., Duque, T., Täckström, O., Uszkoreit, J., & Das, D. (2017). Neural paraphrase identification of questions with noisy pretraining. *Proceedings of the Workshop on Subword and Character Level Models in NLP (SCLeM)*, 142–147.
5. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2227–2237.
6. Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.
7. Xu, W., Callison-Burch, C., & Dolan, B. (2015). SemEval-2015 Task 1: Paraphrase and semantic similarity in Twitter (PIT). *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, 1–11.
8. Dolan, B., Quirk, C., & Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, 350–357.



9. Lan, W., & Xu, W. (2018). Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering. Proceedings of the 27th International Conference on Computational Linguistics (COLING), 3890–3902.

10. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437.