«8D06101 – Информатика, компьютерлік инженерия және бақылау» мамандығының PhD докторанты Ахметов Искандер Рафаилұлының
«Ағылшын тіліндегі ғылыми мәтіндерді ақпараттық экстракциялық рефераттау әдісін әзірлеу» тақырыбындағы диссертациялық жұмысына

АҢДАТПА

Кіріспе

Ақпаратты жедел өңдеу қазіргі уақытта әрбір заманауи адамға қажетті өмірлік маңызы бар функция болып табылады. Бұл саладағы технологияның үнемі дамып келе жатқанына қарамастан, Автоматты мәтінді рефераттау (АРТ) процесі көптеген қиындықтарға тап болады және бұл проблема 1958 жылдан бері зерттеліп келеді. Мысалы, алынған рефераттардың сапасын қалай бағалауға болады және салыстыру үшін эталон қандай қызмет атқаруы тиіс? АРТ процесінде шешілетін екі негізгі міндет бар:

1. Берілген мәтіннен сыни ақпаратты таңдаңыз.
2. Осы ақпаратты қоюландырылған түрде ұсыну.

АРТ - табиғи тілді өңдеу саласындағы күрделі міндет, себебі ол мәтіннің түпнұсқалық деректерінің қысқаша негізделген көрінісін жасау үшін мәтінді мұқият семантикалық және лексикалық талдауды көздейді. Сапалы реферат негізгі ақпаратты қамтуы, фактілерге қатысты нақты болуы, өзекті, оқылатын және артық емес болуы тиіс. Бұл саладағы зерттеулер 1958 жылы басталды және 2003 жылдан бастап жыл сайын көптеген жаңа жұмыстар мен әдістер пайда болды, ол кезде осы мақсатта үлкен көлемде деректер мен қажетті есептеу техникасы қолжетімді болды, бұл зерттеу тақырыбына деген қызығушылықты қайта жаңғыртты.

Мәтінді абстрактілі ету проблемаларын зерттеудің басынан бастап көптеген түрлі әдістер жасалды; Егжей-тегжейлі ақпарат алу үшін 3.2-тарауды қараңыз. әдістер өздері қолданатын құжаттар саны бойынша әр түрлі болады; сөйтіп, бір құжаттық және көп құжаттық автокешендеу бар. мәтінді қорыту әдістерінің екі класын анықтады:

1. **Бастапқы** мәтіннен нақты сөйлемдерді еш өзгеріссіз алу қадамдарын қамтитын экстракциялық реферат.
2. **Реферат** - бастапқы мәтінді еркін түрде келісілген және қысқаша баяндауды көздейді.

Егер осы екі әдісті салыстырсақ, екінші түрі адам ойлау қабілетіне көбірек ұқсас, себебі сөздерді синонимдермен алмастырып, оларды орындарға қайта реттеу қажеттілігі туындайды. Соған қарағанда, экстракциялық әдіс — ең маңызды сөйлемдерді таба отырып, бастапқы мәтіннен түйіндеме жасау.

Осылайша, экстракциялық түйіндемелерді алу оңай және абстрактілі түйіндемелерге қарағанда жақсы нәтижелер береді деп күтілуде.

Осыдан екінші класс күрделірек, себебі ол табиғи тіл генерациясы сияқты күрделі тәсілдерді қамтиды.

Қазіргі кезде дүние жүзі бойынша зерттеулер абстрактілі автоабстракцияға қайта бағдарланған. Дегенмен, соңғы екі жылдағы ғылыми еңбектерден көруге болатындай, өндіруші автоматты реферат әлі де трендте тұр. Түйіндемені қалыптастырудың күрделілігінен басқа, ғылыми қауымдастықта оның бағасы ашық сұрақ туындайды. Мәтіндердің сапалық көрсеткіші табиғи тілдің түсініксіздігін ескеруі тиіс.

Өзектілігі

Ғылыми мәтіндерді автоматты ақпараттық абстрактілеу әдістерін әзірлеу жөніндегі жұмыс қазіргі уақытта жалпы және ғылыми ақпараттың, атап айтқанда, бүгін әлемде экспоненциалды өсуі нәтижесінде бұрынғыдан да өзекті болып отыр.

Қазіргі кезде мәтіндерді автоматты түрде рефераттаудың ең заманауи модельдері миллиардтаған параметрлері бар нейралды желілердің күрделі архитектуралары негізінде салынған және орасан зор көлемдегі деректерге үйретілген және BERT, GPT-3 және т.б. сияқты алдын ала дайындалған тілдік модельдерден эмбадингтерді пайдаланады. Бұл осындай модельдердің шамадан тыс күрделілігі, олардың қорыту қабілеті туралы мәселелерді көтереді, себебі нейралды желінің миллиардтаған параметрлері модельдерге дұрыс жауаптарды жай ғана «есте сақтауға» мүмкіндік береді, ақырында бұл модельдердің экономикалық тиімділігі мен экологиялық қауіпсіздігі туралы сұрақтар туындайды.

Қазіргі кезде әрбір адам және зерттеуші бірінші кезекте ақпаратпен тиімді жұмыс істеу үшін құралдарды шұғыл қажет етеді, оның бірі мәтіндерді автоматты түрде рефераттау жүйесі болуы мүмкін. Бұл тақырып классикалық шығармаларда егжей-тегжейлі зерттелген.

Зерттеу объектісі

Ағылшын тіліндегі ғылыми мәтіндерді автоматты, ақпараттық және экстракциялық рефераттау.

Зерттеу субъектісі

Мәтіндерді автоматты рефераттаудың М этод.

Зерттеу мақсаты

Диссертациялық жұмыстың мақсаты ағылшын тіліндегі ғылыми мәтіндерді ақпараттық экстракциялық рефераттау әдісін әзірлеу, уақытты үнемдеу және өңдеуге арналған ақпарат көлемін азайту болып табылады.

Зерттеу мақсаттары

1. Келісілген алгоритм тәсілі негізінде автоматтық сілтеме әдісін әзірлеу.
2. Әзірленген әдістің бақылау параметрлерін таңдау.
3. Мәтінді абстракциялаудың экстракциялық автоматты әдістерін қолдана отырып, ROUGE көрсеткішін ең жоғары мәнге эксперименттік бағалау жүргізу.
4. Қолданыстағы әдістермен әзірленген әдістің нәтижелеріне салыстырмалы талдау жүргізу.

Зерттеу материалдары

Жұмыста қойылған міндеттерді шешу мәтіндік деректермен тәжірибелер шеңберінде жалпы ғылыми зерттеу әдістерін қолдану және алынған нәтижелерді сандық бағалау негізінде жүзеге асырылды. Бағдарламалау және бастапқы кодтар Python 3.6 (Pandas, Numpy) тілінде жазылған.

Ғылыми жаңалық

Ұсынылып отырған үлгінің жаңалығы мәтіндерді алу, ақпараттық рефераттау әдісінде сараң алгоритмді бірегей қолдануда жатыр.

Сондай-ақ, модель қазіргі заманғы рефераттық модельдер деңгейінде өнімділікті көрсетеді, оларды әзірлеу кезінде нейралды желілер мен оқытуға арналған деректердің орасан зор көлемі пайдаланылды. Бұл ретте ұсынылып отырған модель салыстырмалы түрде қарапайым, және оқыту үшін әлдеқайда аз уақыт пен деректерді қажет етеді.

Біздің зерттеуіміздің ғылыми білімге қосқан үлесі мынадай: 1) экстракциялық жиынтықтау әдістері үшін жоғарғы шекті анықтау (ВНС, сараң алгоритм, генетикалық алгоритм) және келісті алгоритммен инициализацияланған ВНС берілген тапсырма үшін алгоритмдердің кез келгеніне қарағанда одан да жақсы жұмыс істейтінін ашу; 2) салыстырмалы қарапайымдылығына қарамастан, жоғары деңгейде жұмыс істейтін келісті алгоритм негізінде экстракциялық қосынды әдісін ұсыну, 3) жоғары ROUGE және пайдалы мәтін статистикасы бар резюмелердің әр түрі бар тазартылған деректер жиынтығы.

Практикалық маңыздылығы

Біз қарапайым және ескі әдістерді қолданатын, бірақ сонымен бірге күрделі нейралды желі архитектурасын және оқыту үшін орасан зор көлемдегі деректерді пайдаланатын заманауи модельдер деңгейінде жұмыс істейтін экстрактивтік жиынтықтау тәсілін ұсынамыз; Інжір дегенді қараңыз. 1.1 біздің көзқарасымызды қысқаша сипаттау үшін. Біздің тәжірибелерімізде қолданған 17 000 мақаланың arXiv үзіндісі деректер жиынтығы

https://data.mendeley.com/datasets/nvsxfcbzdk/1 қол жетімді. Ұсынылып отырған тәсілдің кейбір басқа артықшылықтарына мыналар жатады:[1]

- Есептеу қарапайымдылығы.
- Machine Learning моделін оқыту талап етілмейді, бірақ статистикалық шығарылым қолданылады.
- Алгоритм жасаған түйіндемелер мәтіннен пайдалы ақпаратқа бай.

Мәтінді автоматты түрде экстракциялық рефераттаудың әзірленген моделі ғылымда, білім беруде және бизнесте практикалық қолданудың кең ауқымына ие.

Ғылым:

- Әдебиетке шолуды автоматтандыру.
- Мақала рефератының ұрпағы.
- Мультимодальды автоорындықтау.
- Ғылымды танымал ету.
- Ғылыми ақпаратты жаңарту.
- Басқа NLP тапсырмаларында autrafering қолданыңыз.
- Тақырыптық модельдеу.
- Көңіл-күйді талдау.

Білімі:

- Жазбаларды автоматты түрде жасау.
- Жадынама.
- Ақыл карталары.
- Көрсетілім слайдтарын жасау.
- Тест сұрақтарын генерациялау.
- Эссе жазуы.

Бизнес:

- Мәтіннің үлкен көлемдерінің қысқаша сипаттамасы (есептер, зерттеулер, бизнес-жоспарлар).
- Отырыс хаттамасын қалыптастыру.
- Сұрауға бағдарланған реферат.
- Контексттік жарнаманы оңтайландыру.

Қорғау жөніндегі әдістемелік нұсқаулар

Қорғау үшін келесілер ұсынылады:

1. Рефераттың сапа деңгейін руж-1 метрикасы бойынша геуристік бағалау әдісі мәтіндерді автоматты түрде рефераттаудың экстракциялық әдістерінің көмегімен қол жеткізуге болады, 0,59 нәтижесін береді, бұл

[1] Бастапқы код мына мекенжай бойынша қолжетімді: ГитХуб https://github.com/iskander-akhmetov/Greedy-Summarization

нейралды желілерді пайдаланудың қазіргі заманғы әдістері үшін 0,46-дан едәуір жоғары.

2. ARXIV деректер жинағында ROUGE-1 метрикасы бойынша 0,42 нәтижесін көрсететін автоматты экстракциялық абстракциялаудың (AER) GreedSum әдісі әзірленді.
3. Құжаттың ең төменгі жиілігінің (min_df) бақылау өлшемін немесе төреші мәтін сөйлемдеріндегі сөздердің туындауының ең аз жиілігін TFIDF матрицасын құруға арналған сөздікті жасағанда ескеру техникасы. ARXiv деректер жиынтығынан 376 мәтіннің үлгісінде қарапайым іздеу арқылы min_df оңтайлы мәні 0,042 деп анықталды (яғни сөз сөйлемдердің кемінде 4,2%-ында пайда болуы тиіс).

Алынған нәтижелерді апробациялау

Диссертациялық зерттеулер нәтижесінде 13 ғылыми мақала, оның ішінде Scopus халықаралық дәйексөз базасына енгізілген шетелдік басылымдарда жарияланған 10 мақала (4 журнал және 6 конференция) жарық көрді, процентильдері 27-ден 80-ге дейін болса, бір авторлық куәлік алынды.

Құрылым

Диссертациялық зерттеулердің құрылымы оның мақсатымен және міндеттерімен анықталады: диссертация кіріспеден, әдебиетке шолу жасаудан, негізгі бөлімнен, қорытындыдан және қолданылатын сілтемелер тізімінен тұрады.

«Негізгі бөлімде» деректер мен әдістерден басқа мәтіндерді автоматты түрде рефераттаудың экстракциялық әдістерін, сондай-ақ келісті алгоритм негізінде автоференцбайлау моделін қолдана отырып, қол жетімді рефераттар сапасының жоғарғы шегін бағалауға арналған эксперименттер жазылады.

Зерттеудің жалпы көлемі 147 бетті құрайды, сілтемелер тізіміне144 тақырып кіреді.