Криста Шонинг Валтер **004.8:02(430)**
Одељење набавке и индексирања,
Немачка национална библиотека, Франкфурт на Мајни

Елизабет Моден
руководилац Одељења за аутоматске каталошке процесе, Немачка национална библиотека, Франкфурт на Мајни

ВЕШТАЧКА ИНТЕЛИГЕНЦИЈА И ДИГИТАЛНЕ ХУМАНИСТИЧКЕ НАУКЕ У БИБЛИОТЕКАМА

Размена искустава на нивоу радионице

У новембру 2022. године, у Немачкој националној библиотеци (DNB Deutsche Nationalbibliothek) у Франкфурту на Мајни, одржана је радионица о примени вештачке интелигенције и дигиталних хуманистичких наука у библиотекама. Фокус је био на коришћењу иновативних метода и услуга за анализу података, текста и слике. Учесници из Берлинске државне библиотеке (SBB Staatsbibliothek), Баварске државне библиотеке (BSB Bayerische Staatsbibliothek), Лајбнишког информационог центра за економију (ZBW Leibniz-Informationszentrum Wirtschaft), Лајбнишког информационионог центра за технологију и природне науке (TIB Leibniz-Informationszentrum Technik und Naturwissenschaften) и Немачке националне библиотеке (DNB Deutsche Nationalbiblithek) разменили су мишљења о пројектима, циљевима и резултатима рада. Радионица је била о мрежним механичким процесима у индексирању. Немачка национална библиотека је тиме подржала дијалог и пренос информација на тему дигиталних промена.

Динамичан развој дигиталних технологија отвара нове могућности за изградњу и проширење колекција, за њихову доступност и коришћење за истраживачке задатке. Примене компјутерски потпомогнутих процеса и дигиталних ресурса у хуманистичким и културним студијама називају се дигиталним хуманистичким наукама (Digital Humanities – DH). Термин

вештачка интелигенција (Künstliche Intelligenz – KI) изражава употребу алгоритама, који подржавају специфичне задатке. У библиотекама користи се за класификацију садржаја текстова или слика, за проналажење сличних објеката у неретко великим збиркама или за пружање функција семантичких претрага.

Данашња решења вештачке интелигенције углавном су заснована на методама машинског учења. За обуку се користе узорци, да би се израчунали предвиђени модели. тј. на основу структуре и шаблона у примерима се користе генерисана правила за класификацију непознатих података, текстова или слика. Перформансе и разноврсност метода брзо су се развиле последњих година.

Решења вештачке интелигенције у библиотекама

У свим библиотекама које су учествовале на радионици већ се користе или тестирају методе које раде по тим принципима. Библиотеке развијају, процењују или користе технике вештачке интелигенције, на пример, помоћу којих се могу препознати сличности у садржају и семантичке везе у текстовима или сликама. Ово укључује тестирање Yewno Discover[1] у Баварској државној библиотеци, где се тестира Yewno Discover, као актуелни претраживач, и Yewno Unearth,[2] као алат за подршку откривању садржаја који такође користи тренутне технологије машинског учења за претраживање сличности слика. У другом пројекту, Баварска државна библиотека ради заједно са Универзитетом у Пасау (Universität Passau) на изазовима који се јављају у вези са креирањем веб-архива и тражењем релевантних веб-страница. Овај други рад је потпомогла Немачка истраживачка фондација (DFG Deutsche Forschungsgemeinschaft).

Још један софтверски систем, који привлачи велико међународно интересовање, јесте Open Source-Toolkit Annif[3] из Националне библиотеке Финске. Ово је компилација доказаних алгори-

[1] Истраживачки алат нове генерације који корисницима библиотеке омогућава да превазиђу претрагу по кључним речима.

[2] Комплексно мапирање садржаја са разним опцијама филтрирања.

[3] Комплет алата отвореног кода за аутоматско индексирање, интегрише неколико алгоритама заснованих на машинском учењу и вештачкој интелигенцији за класификацију текста.

тама, који су погодни за класификацију и кључне речи текстова на природном језику. Немачка национална библиотека је једна од библиотека која успешно користи Annif. Она има комплет алата у својој машини за индексирање и користи га за обраду великог броја својих колекција. Модуларна системска архитектура система индексирања омогућава флексибилно комбиновање алгоритама и – пратећи технички напредак – њихово постепено проширење или замену. Немачка национална библиотека тренутно користи Omikuji-Bonsai метод у вези са другим методама које су доступне у Annifu. Информациони центар за економију у Лајбницу иде сличним путем и користи Annif као основну компоненту свог AutoSE система од 2020. године. Annif представља оквир за комбинацију различитих најсавременијих модела са интерним развојем Лајбнишког информационог центра за економију, а за индексирање специјализованих публикација са загарантованим квалитетом са дескрипторима из Стандардног тезауруса за економију (STW). Лајбницов информациони центар за технологију и природне науке, такође користи Annif и спрема се да пређе на овај комплет алата за коришћења предмета на порталу Лајбнишког информационог центара за технологију и природне науке.

Учешће у истраживачко-развојним пројектима

Берлинска државна библиотека посебно је укључена у развој нових метода за анализу докумената, слика и садржаја и пружање података. Пројекат под називом Mensch.Maschine.Kulture (Човек, Машина, Култура), који је започео 2022. године, финансира Државно министарство за културу и медије. Библиотека ће наставити развој решења заснованих на вештачкој интелигенцији из пројеката OCR-D (Optical Character Recognition) и пројекта Куратор на циљани начин како би их имплементирала у апликације и услуге. OCR-D се фокусира на препознавање текста, док пројекат Куратор креира решења за курирање дигиталног садржаја. Ово укључује алате за претраживање сличности слика и класификацију слика, али и процену квалитета OCR-а и аутоматску накнадну корекцију OCR резултата или препознавање људи, места и организација у неструктурираним пуним текстовима.

Примери актуелних тема истраживања су анализа распореда и препознавање текста, коришћењем неуронских мрежа. OCR-D финансира Немачка истраживачка фондација, а Куратор Савезно министарство образовања и истраживања.

У свом истраживачком пројекту „Систем аутоматског индексирања – индексирање садржаја публикација помоћу вештачке интелигенције”, Немачка национална библиотека жели систематски да истражи која решења воде ка напретку у аутоматском индексирању текстова на природном језику. Користи се терминологија заједничке стандардне датотеке (GND). Семантички концепти стандардне датотеке представљају термине предмета, особа, корпорација, конференција, географских локација и радова. Кључне речи обезбеђују да су текстови тематски класификовани и повезани са другим публикацијама на исту тему. Заједничка стандардна датотека садржи више од милион термина који су потенцијалне кључне речи. Дакле, потребна су решења за такозвану екстремну класификацију текста са више ознака Extreme Multi-Label Text Classification (XMLC). Пројекат Немачке националне библиотеке такође финансира Државно министарство за културу и медије као део Националне стратегије за вештачку интелигенцију.

Евалуација најсавременијих техника

За своје примене, библиотеке понекад користе технике које су такође основа савремених претраживача. Тренутна динамика развоја вештачке интелигенције је и прилика и изазов. Архитектура за посебне алгоритме Deep-Learning-Architekture модела Трансформер тренутно је препозната као најсавременија технологија за употребу обраде природног језика. Добро познати производи су GPT-3 компаније OpenAI, која тренутно привлачи пажњу са ChatGPT-ом, или Google развојним BERT-ом. Ови језички модели су са изузетно великим количинама података. Да ли су модели трансформатора такође погодни за нашу употребу? И колико је труда потребно за ово? Да бисмо одговорили на таква питања, тренутно у Немачкој националној библиотеци спроводимо експерименте вештачке интелигенције, пројект AI са језичким моделом Luminous из Хајделбершког AI start-up Aleph Alpha.

Не ради се само о технологији

Када су резултати механичких процеса довољно добри? Како се квалитет може мерити? Да ли решења вештачке интелигенције такође могу помоћи у осигурању квалитета? Ова централна питања тичу се свих укључених библиотека и заузимају велики део посла. Један пример је концепт осигурања квалитета ZBW-а за услугу AutoSE, који комбинује интелектуалне и механичке мере: грађевински блокови укључују људске повратне информације – такође познате као затворен систем за контролу Human in the loop (HITL) – и аутоматизоване мере за процену квалитета и оптимизацију коришћењем неуронских мрежа.

Дигитална трансформација, такође, отвара нова правна и друга питања. Библиотеке све више користе софтвер отвореног кода, прилагођавајући га за своје апликације и вршећи евалуације. Наш сопствени развој се заузврат дели са заједницом. Који модели лиценци су погодни за ово? Да ли се модели и подаци такође могу делити? Експерименти захтевају поновљиве процесе. За моделирање и тестирање су потребне велике количине података. Да ли алати као што је Контролна верзија података Data Version Control (DVC, верзија система отвореног кода) може помоћи да се постигне управљање подацима? За машинско учење потребна је огромна рачунарска снага. Шта то значи за скалирање техничке инфраструктуре? Да ли је економично куповати сопствене графичке процесоре – или има више смисла радити са центрима података високих перформанси? Да ли се иновације могу убрзати, ако су екстерне истраживачке групе такође укључене у експерименте? Ово је само неколико примера пратећих питања и дилема које су биле истакнуте на радионици.

Извор:

https://blog.dnb.de/ki-und-digital-humanities-in-bibliotheken-ein-erfahrungsaustausch-auf-werkstattebene/ (преузето 1. 2. 2023).

Превела са немачког Тамара Вујков Ђурановић