Tuba Márta

Hatodik európai nyelvtechnológiai konferencia

Kivonat

A beszámoló áttekinti a konferencián bemutatott európai uniós nyelvtechnológiai kutatásokat. Kiemelt fejlesztések folynak a gépi fordítás, az adatok anonimizálása és az álhírek kiszűrése terén. A 2020-as digitális évtized végére kiépülnek a stratégiai fontosságú, energiahatékony és európai értékek mentén működő adatbázisok. A nyelvmodellek a gazdasági igények kielégítése mellett egyre inkább oktatási és közigazgatási célokat fognak szolgálni.

Kulcsszavak: Európai Unió, nyelvtechnológia, mesterséges intelligencia, kutatás

2022. március 31-én hatodik alkalommal rendezték meg az uniós nyelvtechnológiai és mesterségesintelligencia-kutatási konferenciát (European Language Research Coordination, ELRC). Az online rendezvény kiemelt témái közé tartozott az európai nyelvek digitális adatbáziskezelése, a nagy nyelvmodellek létrehozása, fejlesztése és a multimodális információ elemzése.

A résztvevőket Philippe Ghelin, az Európai Unió Multilingvális Kutatóközpontjának vezetője és Andrea Lösch, a Német Mesterséges Intelligencia Kutatóközpont projektmenedzsere üdvözölte.

A délelőtti első téma – digitális Európa – első előadását Andrea Lösch tartotta, aki *ELRC Update* címmel az alkalmazott nyelvészeti kutatás korpuszairól, modelljeiről és forrásairól beszélt. Kiemelte, hogy a nyelvtechnológiai kutatások kulcsfontosságúak az európai piac számára. A nagy nyelvi adatbázisok száma rohamosan nő, ezek 80%-a nyilvános, bárki számára hozzáférhető. A kutatás jelenleg a gépi fordítás fejlesztésére, az adatok anonimizálásának kiterjesztésére és az álhírek szűrésére fókuszál. Utóbbival kapcsolatban felmerül a kérdés, hogy mely nyelvekre kellene koncentrálni ezt a tevékenységet. Andrea Lösch azzal zárta előadását, hogy az Európa Tanács kezdeményezte az adatdonációt ukrán nyelvre, mivel márciusban korlátozták a közösségi média működését az országban.

Eileen Marra (Német Mesterséges Intelligencia Kutatóközpont), az ELRC kommunikációs igazgatója *ELRC White Paper: Language Data Sharing & Language-centric Artificial Intelligence (AI)* címmel tartott előadást. Hangsúlyozta, hogy a nyelvi adatok megosztását a soknyelvű Európában is fenn kell tartani, mert azok ma már nemcsak a közszféra, hanem a kis- és közepes vállalkozások számára is értékesek. Az egyszerű gépi fordításon túl az adatok anonimizálása és a nyelvtechnológiai eszközök szerepének elemzése vezethet valóban egységes digitális nyelvi piachoz. 2022 őszére várható az új White Paper felmérés eredményének publikálása, amely frissített tagországi profilokat tartalmaz, valamint áttekinti a nemzeti nyelvtechnológia működő gyakorlatait és szabályait. Eileen Marra előadásához tartozott a résztvevők kérdőíves felmérése, amely visszajelzés szerint a konferencia nagyjá-

ból 200 fős közönsége megosztottan nyilatkozott a saját anyanyelvén folyó nyelvtechnológiai kutatásfejlesztés közismertségéről és eredményességéről.

A digitalis Európa témát Philippe Ghelin *Quo vadis? The European Language Data Space* című előadása zárta. Philippe Ghelin arról beszélt, hogy a 2020-as évek „digitális évtized" lesz, ugyanis 2030-ra kiépülnek a stratégiai fontosságú adatbázisok, valamint megvalósulnak az energiahatékony, megbízhatóan működő, egyenlően hozzáférhető szolgáltatások. A kutatás és a piaci hasznosulás közötti távolság egyre csökken. 2030-ra a társadalmi-gazdasági szektorokra specializálódott adatbázisok (környezetvédelem, mezőgazdaság, egészségügy, turizmus, kulturális örökség stb.) ökoszisztematikusan fognak működni. A cél a közös európai nyelvi adatbázisok fenntartható, ugyanakkor egyre hatékonyabb működtetése. A folyamat két éven belül az Európai Digitális Infrastruktúra Konzorcium projekt keretében indul.

A konferencia második kiemelt témájának – nagy európai nyelvmodellek – előadásait Jörge Bienert, a Német Szövetségi Mesterségesintelligencia-kutató Társaság elnöke kezdte (*Why Europe Needs Large Language Models. An Economic Perspective*). Kijelentette, hogy a vezető technológiai pozíciót elfoglaló országok gazdasági fölénye vitathatatlan, a mesterséges intelligencia hatalmas gazdasági potenciált jelent. A becslések szerint 2030-ra Kína 26%-os, az USA 14,5%-os GDP-növekedést vár a kutatásfejlesztéstől. A versenyképesség megőrzéséhez szükséges nagy európai nyelvmodelleket létrehozni és európai értékek mentén működtetni. A soknyelvű digitális források legyenek hozzáférhetők, átláthatók, CO_2-semlegesek, valamint nyújtsanak biztonságos adatkezelést a felhasználók számára. Jörge Bienert beszámolt arról, hogy Németországban széles körű gazdasági támogatással építik a Gaia-X projekt első kísérleti, számítógépes beszédfelismerő modelljét.

Ezután Igor Carron, a LightOn vezérigazgatója *Possibilities and Limitations of Large Language Models: PAGnol, VLM-4 and Muse* című előadása következett. 2020–2022 között létrejöttek a nyelvtechnológiai kutatás új modelljei (GPT-3, Codex, AlphaCode, AlphaFold), amelyeknek megdöbbentő tulajdonsága, hogy új feladatokat természetes nyelven kapott utasításokra is végre tudnak hajtani. Igor Carron demonstrálta a Muse és VLM-4 modell szöveggenerálási, szövegosztályozási, kulcsszófelismerési, oktatási tevékenységét. Prezentációját azzal zárta, hogy van elég adat ahhoz, hogy a modellek ne csak angol, hanem pl. francia, spanyol, német, arab nyelven és meghatározott piaci igényeket kielégítve működjenek, de a gazdasági versenyképesség megőrzéséhez, az oktatásban való felhasználás kiterjesztéséhez további minőségi adatok, gondozott adatbázisok kellenek.

Magnus Sahlgren, a National Language Understanding kutatásvezetője a kis nyelveken működtetett nyelvmodellekről tartott előadást (*Language Models for Swedish Authorities*). Bár a legtöbb svéd vállalat nemzetközi, a közszférában szükség van a svéd nyelvtechnológiai eszközök működtetésére. Jelenleg az angol SuperGLUE-nak megfelelő svéd SuperLim értékelő rendszer létrehozásán és a GPT-SW3 modell 2-es verzióján dolgoznak. A svéd mesterségesintelligencia-kutatás egyesíteni szeretné a morfológiailag hasonló, északi germán nyelvek adatbázisait a hozzáférhetőbb, kiterjesztett hatókörű feladatellátáshoz.

A konferencia vitarésze Nikos Sarris (CERTH) és Maria Bieliková (KInIT) előadásán alapult (*Language Technologies for Fighting Disinformation*). Nikos Sarris a Görög Tudományos és Technikai Kutatóközpont munkatársaival végzett tanulmány eredményeit ismertette az etikus újságírásra vonatkozóan. Hangsúlyozta a dezinformáció jelenségének összetettségét, ezért a helyzet nem rendezhető kizárólag szabályozással. Az ukrán háborúra tekintettel ismertette az Európai Digitális Médiafigyelő kezdeményezés (EDMO) működését. Maria Bieliková, az Intelligens Technológiák Kempelen Intézetének általános igazgatója az álhírek kiszűrését a tényellenőrzéssel hasonlította össze. Az álhírek szűrése virális mémek lenyomozásán alapul, amely repetitív tevékenységet (állításfelismerés, álláspont azonosítás, szöveges megvalósulás) a számítógépek megbízhatóan el tudják végezni.

A további előadások a multimodális nyelvi adatok és a nagy nyelvmodellek témái köré szerveződtek. Suzanne Little-től, a Dublini Egyetem számítástechnikai karának docensétől az Insight szolgáltatás keretében végzett nyelvtechnológiai kutatásokról hallottunk (*Behind the Scenes: Multimodal Data Analysis*). A trollmémek gépi felismerése annak megfigyelésére épül, hogy mely verbális és képi összetevők milyen gyakorisággal, milyen kontextusokban torzítják el a szövegek jelentését. A 19. századi napilapok karikatúráinak elemzése segíti a mai mémeket humorosnak, szarkasztikusnak vagy esetleg már sértőnek, gyűlöletkeltőnek minősíteni. A kutatásban az jelenthet továbblépést, ha a kontextuális jelentést nem binárisnak, hanem kontinuumnak fogjuk fel. A Crowd4Access projekt keretében a városlakók okostelefonjáról feltöltött képeket közlekedési információkkal társított vizuális modellé alakítják, elsősorban a gyalogosok kiszolgálására.

Alexandra Konig neuropszichológus a francia Nemzeti Digitális Tudományos és Technológiai Kutatóintézet (INRIA) képviseletében a nyelvtechnológiai eszközök növekvő orvosdiagnosztikai szerepéről tartott prezentációt (*Multimodal Data in the Medical Domain*). A mentális betegségek diagnosztizálásában a bevett módszereket – megfigyelés, felmérés, skála – egyre inkább átveszi a digitális fenotipizálás. Az adatgyűjtés történhet a páciens okostelefonjáról és testén viselt szenzorairól, illetve a beszéd- és videóanalízis akár telefonbeszélgetéshez, akár személyes interjúhoz társulhat. A klinikai szempontból releváns paralingvisztikai és mimikai tényezők alapján egyre korábban és pontosabban diagnosztizálható pl. az Alzheimer-kór vagy a demencia. A visszajelzések szerint bevált a számítógépes konzultáció, monitorozás a depresszió kezelésében olyan kieső területeken, ahová az egészségügyi személyzet a járvány miatt nem jutott el személyesen. A Mephesto-programmal hosszú távon nyomon követhetők a multimodális orvos-beteg interakciók.

Anabela Barreiro (Inesc-id, Lisszabon), az Európai Tudományos és Technológiai Együttműködés (COST) elnöke előadásában az interdiszciplináris kezdeményezés létrejöttéről és működéséről hallottunk (*Multi3Generation: Multimodal Data for Natural Language Generation*).

A nagy nyelvmodellek projektjei közül Senia Pollack (Szlovénia) az Embeddia, Sebastian Andersson (Helsinki) a Microservice, Dimitra Anastasiou (Luxemburg) az Enrich4all működéséről és szolgáltatásairól számolt be a konferencia közönségének.

A tudományos rendezvény Andrea Lösch összegzésével zárult. A prezentációk anyaga elérhető a https://lr-coordination.eu/6thELRC és a https://youtu.be/ebAbv5KgvrQ felületeken.

Tuba Márta
középiskolai tanár, PhD
Gárdonyi Géza Általános és Középiskola, Érd
E-mail: dr.tuba.marta@gmail.com
https://orcid.org/ 0000-0002-8264-2108

Abstract

Tuba, Márta

Sixth European Conference on Language Technology

The report reviews the European Union language technology research presented at the conference. Major developments are underway in the field of machine translation, data anonymization and filtering out fake news. By the end of the 2020–2030 digital decade, strategically important, energy-efficient and European-valued databases will be built. In addition to meeting economic needs, language models will increasingly serve educational and administrative purposes.

Keywords: European Union, language technology, artificial intelligence, research