Fordítástudomány 24. (2022) 1. szám 127–132.

A többnyelvű Európa újraformálása: A nyelvközpontú mesterséges intelligencia

(3. magyar ELRC workshop, online esemény, 2022. február 7.)

Jelencsik-Mátyus Kinga, Vadász Noémi
E-mail: matyus.kinga@nytud.hu, E-mail: vadasz.noemi@nytud.hu

A Nyelvtudomány Kutatóközpont és a European Language Resource Coordination közösen rendezte meg a 3. magyar ELRC Workshopot 2022. február 7-én. A két intézmény harmadik alkalommal hívta össze a magyar kutatás, ipar és közigazgatás szakértőit, hogy megvitassák, miként alakíthatja át a magyar nyelvre elérhető nyelvtechnológia a kommunikációnkat a digitális térben. A járványhelyzet miatt az eseményt online formátumában rendezték meg. Az középpontban három téma állt: a nyelvközpontú mesterséges intelligencia (MI), a nyelvi adat és a gépi fordítás. Éppen ezért az esemény mottója, ahogy jelen beszámoló címében is olvasható: A többnyelvű Európa újraformálása: a nyelvközpontú mesterséges intelligencia.

A Európai Hálózatfinanszírozási Eszközt (Connecting Europe Facility – CEF) az Európai Bizottság indította útjára 2014-ben. Célja az Európán átívelő hálózatok és infrastruktúrák támogatása, a 28 tagállamra, valamint Norvégiára és Izlandra kiterjedő Digitális Közös Piac létrehozása. Váradi Tamás rövid bevezetőjében kiemelte, hogy a kulturális, turisztikai, egészségügyi és jogi szolgáltatásoknak mindenki számára elérhetőnek kellene lenniük a polgárok anyanyelvére való tekintet nélkül. Mindennek elengedhetetlen feltétele az információáramlás biztosítása, amelyhez a CEF gépi fordítás szolgáltatása (CEF AT) nyújt segítséget.

A mai gépi fordítóknak hatalmas mennyiségű adatra van szükségük, az European Language Resource Coordination (ELRC) pedig ezen a ponton kerül a képbe. Az ELRC koordinálja a nyelvi erőforrásokat és eszközöket az EU-országokban. Manapság az ELRC külön figyelmet szentel a többnyelvű NLP-nek, valamint az egynyelvű korpuszoknak is, amelyek tanítóanyagként szolgálnak a nyelvi modellek építéséhez. Az NLP világán belül paradigmaváltás ment végbe az elmúlt években. A neurális hálók segítségével nagyobb és hatékonyabb egy- és többnyelvű nyelvmodellek építhetők. Utóbbiak lehetővé teszik a transzfer learninget, ami lehetőséget nyújt a többnyelvűség okozta kihívások kezelésére.

Hivatkozás: Jelencsik-Mátyus K., Vadász N. 2022. A többnyelvű Európa újraformálása: A nyelvközpontú mesterséges intelligencia (3. magyar ELRC workshop, online esemény, 2022. február 7.) *Fordítástudomány* 24. évf. 1. szám. 127–132.
DOI: https://doi.org/10.35924/fordtud.24.9

A gépi fordítás mellett számos egyéb program segíti a kommunikációt – és ezáltal munkánkat, vásárlási, utazási és ügyintézési szokásainkat – mindennapjainkban: a helyesírást segítő programok, a beszédünket szöveges üzenetekké alakító digitális asszisztensek, az ügyfélszolgálatokhoz intézett hívásainkat megválaszoló robotok ma már szinte észrevétlenül életünk részei. A workshop arra koncentrált, hogy léteznek-e már a magyar nyelvre megfelelő nyelvtechnológiai megoldások ahhoz, hogy lépést tartsunk a technológiai fejlődéssel a mesterséges intelligencia korában.

A meghívott előadók prezentációi mellett három panelbeszélgetést is tartottak a fő témakörökben, amelyeken a magyar nyelvtechnológia kiemelt képviselői vettek részt. A egész eseményt és a panelbeszélgetéseket az ELRC magyar képviselői, Váradi Tamás és Bódi Zoltán moderálták. Összesen 150 résztvevő csatlakozott az eseményhez. Jelen beszámolóban az előadásokat és az első demó bemutatót foglaljuk össze.

Az új Digital Europe Programme és a Language Data Space

Az első meghívott előadó Philippe Gelin, a DG-CONNECT Többnyelvűségi Központjának igazgatója volt, aki bemutatta az új Digitális Európa programot és a Language Data Space-t. Ő is azzal kezdte előadását, hogy a nyelvi modellek megjelenése paradigmaváltást, és ezzel soha nem látott gyorsaságú fejlődést hozott a nyelvtechnológiába. Ennek eredményeként számos új alkalmazás jött létre a nyelvi akadályok legyőzésére, és így az új kultúrák megismerése már pusztán a mobiltelefonunk használatával lehetségessé vált. Az angol nyelv digitális piacot uraló dominanciájának ellensúlyozására kiemelten fontos, hogy az EU többi, kisebb hivatalos nyelvére is szülessenek nyelvtechnológiai alkalmazások. Ezzel az egyes nyelvek gazdasági versenyképességét is jelentősen növelhetjük.

Ezt támogatandó az új Digital Europe Programme és a Language Data Space biztonságos és jó minőségű megoldásokat nyújt nemcsak minden európai nyelv, hanem más nagyobb nyelvek támogatására is, mint a kínai és a japán. Az Európai Bizottságban számos kezdeményezés jött létre az ezzel kapcsolatos fejlesztések finanszírozására az 5G hálózat kiépítésétől a közlekedés fejlesztésén át a fordítás támogatásáig.

Az Európai Bizottság 2020-ban mutatta be adatstratégiáját[1], amelyben fontos helyet kaptak a nyelvi adatok, elsősorban azok elérhetővé tétele ipari felhasználásra a KKV-k számára. A 20 meghatározott témacsoport közül a Language Data Space fogja segíteni az információáramlást a nyelvtechnológiai szektor szereplői között, ezzel lehetőséget biztosítva hatékony nyelvi eszközök és szolgáltatások kidolgozására. Ez pedig a mesterséges intelligenciához kapcsolódó megoldások

[1] https://digital-strategy.ec.europa.eu/en/policies/strategy-data

gyors térnyerésével egyre fontosabbá válik. Előadása végén Philippe Gelin felhívta a hallgatóság figyelmét három olyan weblapra[2], amelyek több száz magyar NLP-erőforrás bemutatásával hasznos segítséget nyújtanak például a fordításban.

A nyelvtechnológia és az MI lehetőségei - hol tartunk most és mik a célok

Prószéky Gábor, a Nyelvtudományi Kutatóközpont főigazgatója ezután a nyelvtechnológia és a mesterséges intelligencia lehetőségeiről beszélt. Amikor a mesterséges intelligencia megjelent az 1950-60-as években, elsődleges fókuszában az állt, hogy a lehető legjobban tudja modellezni az emberi intelligenciát. Az 1980-as években, a gépi tanulás megjelenésével már olyan összetett algoritmusokat hoztak létre, amelyek segítségével egyes jól meghatározott területeken már képesek tapasztalati tanulásra, amennyiben elegendő adatot biztosítanak a számukra. Ez az adatmennyiség a korábbi gyakorlatnál sokkal, akár százszor nagyobb. Összegyűjtése idő- és energiaigényes a fejlesztésben részt vevő szakemberek számára. A gépi tanulás középpontjában a mély tanulás (deep learning) áll (ezt a fogalmat nevezi a média gyakran mesterséges intelligenciának). Azért nevezik mély tanulásnak, mert a neurális hálózatok rendszere több rétegből áll a be- és kimeneti rétegek mellett. Az egyes rétegek egymásra épülnek felhasználva a létrejött információt, és ennek köszönhetően saját adatfeldolgozással tanulhatnak.

Napjainkban a mesterséges intelligencia átszövi mindennapjainkat a mezőgazdaságtól az önvezető autókig. Nyelvészeti szempontból már most számos olyan alkalmazással találkozhatunk, amelyek elérik, sőt akár meg is haladják az emberi munka eredményességét. Az automatikus beszédfelismerés, a szövegleíró szolgáltatások, a chatbotok, a szentimentelemzés, a névelem-felismerés és a gépi fordítás mind jó példák erre. Fontos kiemelni, hogy míg nagy nyelvekre már eredményes eszközök jöttek létre komplex feladatokra is, kisebb nyelvek esetében, mint például a magyar, ez még várat magára. Erre erős hatással van nemcsak az, hogy milyen mennyiségű, hanem az is, hogy milyen minőségű nyelvi adat áll rendelkezésünkre a fejlesztéshez.

Zárásként Prószéky Gábor kiemelte, hogy a Nyelvtudományi Kutatóközpont nagy szerepet vállalt a nyelvközpontú mesterséges intelligencia fejlesztésben. Ennek legjobb példája a HILBERT, egy magyar nyelvre kifejlesztett BERT-large modell, valamint számos kísérleti nyelvmodell a HILANCO projekt[3] keretében. És a java még csak ezután jön!

[2] https://digital-strategy.ec.europa.eu/en/policies/strategy-data, https://language-tools.ec.europa.eu/, https://www.european-language-grid.eu/

[3] https://hilanco.github.io/

A CEF gépi fordítás platform

Farkas Ágnes a DG-Translation képviseletében bemutatta a CEF gépi fordítás platformját. A jelenleg futó rendszer, az eTranslation már 2017 óta segíti a fordítók munkáját. Számos domént lefed és neurális gépi tanulási (NMT) módszereket használ. Ezen szolgáltatások fejlesztéséhez sok pénz szükséges, amelyet a CEF-platform, illetve az annak nyomába lépő DIGITAL[4] biztosít. Az eTranslation egy rendszer, de két közönséget céloz: elsősorban a fordítók, és egyéb EU-szervezetek dolgozóinak a támogatására hozták létre, valamint a rendszer integrálható az EU-szervezetek weboldalaiba és digitális szolgáltatásaiba is. A másik célközönség a CEF-programok által finanszírozott közösségek: páneurópai digitális közszolgáltatások, az EU-tagországok (valamint Norvégia és Izland) digitális közigazgatása, egyetemek, az EU-nak dolgozó szabadúszó fordítók és KKV-k. Tehát az eTranslation mind emberi felhasználásra alkalmas szolgáltatásokat, mind pedig digitális szolgáltatásokba építhető API-kat is biztosít (ez utóbbi jelenleg a csak a közigazgatás számára elérhető). Minden szolgáltatás ingyenesen hozzáférhető regisztráció után. Az oldal rendkívül biztonságos, mivel az EB tűzfala védi, és minden adatot törölnek 24 óra után (a beállításokban választhatjuk az adatok azonnali törlését is). Az eTranslation számos domént lefed, például EU hivatalos szövegek, általános szövegek, EU Bíróság dokumentumai, pénzügyi (EKB) és közegészségügyi írások. Mindez 30 nyelvre elérhető: az EU hivatalos nyelvein kívül hat nagy nyelvre. Minden feltöltött szöveget lefordíthatunk bármely, vagy akár az összes felsorolt nyelvre. A fordításokat emailben, vagy a szolgáltatáson létrehozott fiókunkba kérhetjük. Az alkalmazás számos bemeneti formátumot támogat.

A legfontosabb tényező természetesen a minőség. Az eTranslation az EU szövegek esetében nyújtja a legjobb eredményeket. Ez annak köszönhető, hogy rendszert a hivatalos EU-fordításokon tanították. Sajnos a kisebb nyelvek esetében, mint amilyen a magyar, kevesebb adat áll rendelkezésre, ez befolyásolja az fordítás minőségét is. A program kevésbé hatékony új szavak, kontextus nélküli szavak valamint kreatív szövegek esetén. A honlap további NLP szolgáltatásokat is nyújt, például beszédfelismerés és többnyelvű tweet. Előadásában Farkas Ágnes kiemelte, hogy annak ellenére, hogy a gépi fordítók már egyre jobb eredményt nyújtanak, nem helyettesíthetik az emberi fordítást, utószerkesztésre mindenképpen szükség van. Hangsúlyozta, hogy az Európai Bizottságnál sem géppel fordítják az uniós jogszabályokat, az alkalmazások csak támogatják a fordítók munkáját.

[4] https://digital-strategy.ec.europa.eu/en/activities/digital-programme

Az adat szerepe a csúcsminőségű nyelvtechnológiai fejlesztésekben

Feldmann Ádám a Pécsi Tudományegyetem Alkalmazott Adattudományi és Mesterséges Intelligencia csoportjának vezetője előadását azzal kezdte, hogy a mesterséges intelligencia fejlődésében a nyelvtechnológia fejlődése a húzóerő, főleg a transzformer-alapú modellek megjelenése óta. A transzformer-alapú modellek jelenleg a legjobb teljesítményű technológiák, ugyanakkor rengeteg adatra van szükség hozzájuk. Kiemelte, hogy a transzformer-alapú nyelvmodellek megjelenése óta a nyelvtechnológia fejlődése belépett a nagy nyelvmodellek korszakába. A nagy nyelvmodellek a nyelv reprezentációjának vagy tömörített formában való megjelenésének tekinthetők.

A PTE és a Nyelvtudományi Kutatóközpont által a Microsoft Magyarország támogatásával kifejlesztett magyar BERT-large modell, a HILBERT elkészítéséhez például 3-4 milliós szövegkorpuszt kellett létrehozni, ugyanakkor a kínai GPT-3 modellhez 410 milliárdos korpuszt használtak. Ezek az óriási nyelvmodellek lehetővé teszik, hogy egy egészen kisméretű tanítókorpusszal, néhány tanítópéldával (few-shot learning), sőt, akár tanítópéldák bemutatása nélkül (zero-shot learning) alkalmazzuk őket valamilyen specifikusabb feladatra (pl. osztályozási feladatokra, névelemfelismerésre stb.). Ez azt jelenti, hogy megéri nagy mennyiségű adatot gyűjteni óriási általános nyelvmodellek építéséhez, mert így gyorsabban és könnyebben megoldhatók a specifikusabb nyelvtechnológiai feladatok.

A magyar nyelv esetében azonban nem áll a rendelkezésünkre korlátlan mennyiségű szöveg. Erre nyújt megoldást az ún. transfer-learning, amely lehetővé teszi az angol nyelvre kifejlesztett megoldások használatát más (kisebb) nyelvekre plusz tanítóanyag előállítása nélkül. Ez az eljárás hasonlít a humán tanulás folyamatára: amit az egyik nyelven megtanulunk, azt alkalmazni tudjuk majd egy újabb idegen nyelvre is. Az előadó szerint a közeli jövőben a multimodális adatok gyűjtése kerül majd előtérbe (pl. kép-narratíva párok, videó-szöveg párok), hiszen ezek képezhetik majd az alapját az újabb modelleknek. Érdemes lenne a magyar kutatások esetében is követni ezeket a nemzetközi trendeket.

Demók bemutatása

A demó szekcióból a European Language Equality projektet szeretnénk bemutatni, mivel ennek eredményei befolyással lesznek a többnyelvű Európa jövőjére, például az NLP kutatási és fejlesztési irányainak meghatározásakor, beleértve a finanszírozást is. Ebben a kezdeményezésben 52 partner, köztük minden EU-s tagországból kutatóhelyek, egyetemek, csakúgy mint ipari szereplők és páneurópai szervezetek (mint az EFNIL és a CLARIN) vettek részt. A projekt célja egy stratégiai akcióterv kidolgozása, és ennek részeként az EU hivatalos nyelveire NLP-erőforrás térkép létrehozása azért, hogy támogassuk a nyelvi egyenlőség létrejöttét Európában 2030-ig.

A Nyelvtudományi Kutatóközpont volt a magyar koordinátora a nagyszabású adatgyűjtésnek, amelynek célja a lehető legtöbb magyar nyelvre elérhető nyelvi erőforrás feltérképezése és katalogizálása volt. Vagyis sorra vettek minden elérhető korpuszt, lexikai adatbázist, eszközt, nyelvtant, nyelvi modellt és szolgáltatást, és ezzel egy több mint 500 elemű listát hoztak létre. Fontos hangsúlyozni, ez a szám csupán egy 2021 végi pillanatkép egy villámgyorsan fejlődő tudományterületről, és az adatbázis megjelenése óta már biztosan számos új nyelvi erőforrás jelent meg. Az adatbázis elérhető a European Language Grid honlapján[5]. A gyűjteményből megtudhatjuk, hogy melyek azok a területek, amelyeken sok, jó minőségi nyelvi erőforrást találunk: számos adatbázis elérhető például a két- és többnyelvű korpuszoknál, ami elengedhetetlen a gépi fordítás fejlesztéséhez. Ugyancsak több és jó minőségű szövegelemző eszköz és elemzőlánc elérhető magyar nyelvre, amelyek közt van olyan, amely ipari célokra is használható. Azonban a projekt eredményeként az is kiderült, hogy melyek azok a területek, amelyeken az erőforrások hiánya hátráltatja a fejlődést. Általánosan elmondható, hogy a neurális hálók térnyerésével a nyelvtechnológia számos területén szükség van nagy, részletesen annotált korpuszok létrehozására.

Összefoglalva, a 3. magyar ELRC workshop a magyar kutatás, ipar és közigazgatás szakértőit hívta össze, hogy megtárgyalják, milyen lehetőségei vannak a magyar nyelvtechnológiának a nyelvközpontú mesterséges intelligenciával kapcsolatban. A különböző szektorok képviselői egyetértettek abban, hogy bár már léteznek kiemelkedő minőségű nyelvtechnológiai szolgáltatások a magyar nyelvre, a nyelvi modellek térnyerésével kiemelten fontos szerepet kaptak a nyelvi adatok. Így ahhoz, hogy technológiai támogatottság terén a magyar nyelv is lépést tudjon tartani a fejlődéssel, nagyszabású összefogás szükséges az adatgyűjtésben.

[5] https://live.european-language-grid.eu/catalogue/?&language__term=Hungarian