

A mesterséges intelligencia és a magyar nyelv

Tudomány-népszerűsítő előadás és bemutató

Időpont:

2022. november 23. 16.00 és 18.00 óra között

Helyszín:

MTA Székház, Díszterem

1051 Budapest, Széchenyi István tér 9.

Részletek:

Az esemény közvetítését [\[itt\]](#) érheti el az MTA YouTube-csatornáján.

Előadók:

Prószéky Gábor, a Nyelvtudományi Kutatóközpont (NYTK) főigazgatója

Váradi Tamás, a Nyelvtechnológiai és Alkalmazott Nyelvészeti Intézet, NYTK igazgatója

Yang Zijian Győző, Ligeti-Nagy Noémi és Laki László (a Nyelvtechnológiai Kutatócsoport tudományos munkatársai)

Névjegyek:

Prószéky Gábor

Prószéky Gábor a Nyelvtudományi Kutatóközpont főigazgatója, valamint a Pázmány Péter Katolikus Egyetem Információs Technológiai és Bionikai Karának egyetemi tanára. A Magyar Tudományos Akadémia doktora. Társalapítója a nyelvtechnológiával foglalkozó MorphoLogic Kft-nek, amely a kutatás-fejlesztés területén olyan partnerekkel dolgozott 1991-es indulása óta, mint például a Microsoft, az IBM, vagy a Xerox, illetve a legnagyobb európai egyetemek és kutatóközpontok. Számos egyetemen tanított a nyelvtechnológia bölcsészettudományokban

való felhasználási módjairól. Több mint 30 nyelvtechnológiához kapcsolódó projektben vett rész. Mintegy 200 tudományos cikk és 3 nyelvtechnológiáról szóló könyv szerzője.

Váradi Tamás

Váradi Tamás a Nyelvtudományi Kutatóintézetben működő Nyelvtechnológiai és Alkalmazott Nyelvészeti Intézet igazgatója. Több, nyelvtechnológiával foglalkozó nemzetközi szervezet vezető tagja (EFNIL, CLARIN, ELRC). Számos európai finanszírozású és hazai nyelvtechnológiai projektet koordinált (MARCELL, CESAR, iTranslate4, Trendminer stb.). Száznál több nyelvtechnológiához kapcsolódó tanulmány szerzője vagy társszerzője, illetve könyvek szerkesztője. Szakmai munkáján túl fontos tudományszervezési, nyelvpolitikai szerepet is betölt annak révén, hogy részt vesz olyan tervezetek megalkotásában, melynek célja a nyelvek digitális támogatása.

Nyelvtechnológiai Kutatócsoport, NYTK

A kutatócsoport elsődleges feladata, hogy felzárkóztassa a magyar nyelvet a világ nagy nyelveihez a digitális technológiai támogatás tekintetében elsősorban a nyelvközpontú mesterséges intelligenciára kidolgozott nemzetközi eljárásoknak a magyarra adaptálásával. Kutatásunk fókuszja így egyrészt az általános célú nyelvmodellek építése, finomhangolása, illetve a nyelvmodellek kiértékelése új *benchmark* adatbázisok tervezésével és létrehozásával.

Összefoglaló

A modern társadalmakban a kommunikáció döntő módon digitális eszközökkel, digitális módon történik. Az elmúlt néhány évben elterjedt egy új technológia, amely látványos fejlődést eredményezett a digitális világ szinte minden területén: a neurális hálók megjelenésével a nyelvtechnológia is megkérdőjelezhetetlenül a digitális kommunikáció mindennapi eszköztárának részévé vált. A rohamosan fejlődő kutatási háttérnek hála a tudományág lehetőségei még messze nincsenek kimerítve. Mivel a fejlesztési trendeket főként a globális technológiai cégek (Google, Microsoft, Amazon stb.) szabják meg, és ezek elsősorban az angol nyelvet célozzák, ennél fogva a digitális szolgáltatások, amelyek egyébként az okostelefon jóvoltából tömegekhez jutnak el, nyelvi támogatás hiányában nem használhatók a magyarra. Másfelől az új kutatási paradigma számos új alkalmazási és kutatási lehetőséget rejt magában. Így a NYTK Nyelvtechnológiai Kutatócsoportjában zajló munkálatok során egyszerre célunk a már meglévő eszközök és módszerek adaptálása a magyarra és az új kutatási lehetőségek feltérképezése is.

Az új paradigmát egyértelműen a neurális nyelvmodellek alapozzák meg. Ezek a nyelvmodellek nagymennyiségű adaton tanított mély neurális hálók, amelyek számos nyelvi

jelenséget illetően rendelkeznek általános tudással. Ezek továbbtanításával pontosabb nyelvtechnológiai alkalmazásokat kaphatunk. A NYTK szuperszámítógépének segítségével sikerült létrehozni az első magyar nyelvű szöveggeneráló GPT-3 modellt, ami képes emberekhez hasonló módon mondatokat generálni.

A humán nyelvtudás mérésének a mintájára a nyelvmodellek nyelvtudását is tudni kell értékelnünk és összehasonlítani. Erre szolgálnak az úgynevezett benchmark adatbázisok, amelyekben változatos tematikájú és nehézségű nyelvi feladatokon mérjük a nyelvmodellek teljesítményét. A Nyelvtudományi Kutatóközpontban folyamatosan dolgozunk a modellek magyar nyelvtudását tesztelő korpuszok létrehozásán. Előadásunkban röviden ismertetjük ezeket.

Egy másik kiemelt kutatási irány a gépi fordítás területe amely az utóbbi években szintén nagy fejlődésen esett át. A neurális hálózatok minőségben a statisztikai modellek felé kerekedtek, és elérték, hogy az emberi fordítóknak is hatékony segítséggé váljanak. A Google 2017-ben publikálta Transformer nevű neurális architektúráját, amit alapvetően gépi fordítás céljából hozott létre. Ez mára mind a nyelvtechnológia, mind a beszédtechnológia piacvezető technológiájává vált, és a transzformer- alapú alkalmazások beépültek szinte az összes okosnak nevezett eszközbe. Munkánk során a magyar nyelv fordítási támogatására összpontosítottunk, és több magyar-centrikus fordítórendszert is létrehoztunk, amelyeknek minősége meghaladja a Google fordító motorjait - elsősorban az angoltól eltérő nyelvekről történő fordítás esetén.

Program

- 16:00 Prószéky Gábor: Bevezetés
- 16:10 Váradi Tamás: Paradigmaváltás a nyelvtechnológiában és az NYTK-ban
- 16:30 Yang Zijian Győző: Magyarul tanuló gépeink
- 16:50 Ligeti-Nagy Noémi: Teszt a lelke mindennek
- 17:10 Laki László: Magyarra fordítva a szót
- 17:30 Yang Zijian Győző: Demó

Szervező: MTA Titkárság Kommunikációs Főosztály