Szószerkezet felismerése mélytanulással – projekt záró beszámoló

Kornai András

Hátér A félidős (2018) részjelentésben összefoglaltuk az agglutináló (pl. magyar) és az izoláló (pl. angol) nyelvek közti különbségeket, a szószerkezet három rétegét (gyökök, képzők, ragok), és a felügyelt ill. felügyelet nélküli gépi tanulásos módszerek közti különbségeket, melyek az eredeti (2016) pályázat struktúráját meghatározták. Ebben a részjelentésben már tárgyaltunk a projekt a magyar morfológia szabályalapú kezelésében is hasznosnak bizonyult eredményeit, elsősorban a GLF alakokon alapuló vizsgálatokat, így ezeket itt csak tömören foglaljuk majd össze.

A statikus szóbeágyazások témakör-szerkezetének vészhelyzeti szókincsen (emergency vocabulary) való vizsgálatának eredményeit sem részletezzük újra, nemcsak azért nem, mert ezek azóta publikálásra kerültek, hanem mert a statikus rendszerek vizsgálatát némiképp időszerűtlenné tették a 2018-2019-ben bevezetett dinamikus rendszerek. Mint azt 2019-es részjelentésünkben írtuk:

> Komoly változást hozott a környezetfüggő beágyazások (contextual embeddings), elsősorban az ELMO (Peters et al 2018), és a BERT (Devlin et al 2019) megjelenése, és célunknak tekintjük, hogy a morfológiai mélytanuláshoz ezeket is használatba vegyük, különösen az eléggé alulkutatott agglutináló nyelvekre, elsősorban a magyarra. Erről természetesen az eredeti pályázatban (2016) még nem volt szó.

Kutatásunk fő célja változatlanul a magyarban és más agglutináló nyelvekben (finn, török, bantu) rendkivül fontos szóalaktan (morfológia) vizsgálata a mesterséges intelligencia-kutatás korszerű, mélytanulásos eszközözeivel. De miután a statikus beágyazásokat felváltották a dinamikusak, mi is átstrukturáltuk a projektet annak érdekében, hogy ne maradjunk le a nemzetközi élvonaltól.

A statikus szakasz A 2019-es átstruktúrálás előtt felügyelt rendszereket alkottunk elsősorban az e-Magyar (Váradi et al 2015) segítségével előállított szegmentációval mint felügyelettel. A Webkorpusz (Halácsi és mtsai 2004) ilyetén szegmentálásával előállított un. gluten-free (GLF) szövegeken kimértük, hogy a modern, neurális nyelvmodellek teljesítménye nem romlik GLF korpuszokon (szemben a hagyományos n-gram modellekkel), hogy GLF esetén a szótár mérete 1/3-a a nyers korpuszból épitett szótárénak, és hogy GLF korpuszon feltanított szóbeágyazások képesek kapcsolatot vonni a hasonló szerepű névutók, határozószók és ragok között (pl. bele legközelebbi szomszédja -ba/-be). Ennek az időszaknak jelentős eredménye, hogy tisztán geometriai eszközökkel is megragadhatóak a szóvektorok és a szófajok közti alapvető összefüggések (Lévai 2019), miközben a szóvektorok közt téma szerinti kapcsolat nem nagyon látható, a statikus vektorok tehát inkább szintaktikai mint a szemantikai modellek (Nemeskey és Kornai 2018).

A dinamikus szakasz Az átstruktúrtálás utáni szakasz eredményei közül az alábbiakat emeljük ki:

- mélytanulásos morfológiai rendszerünk ezüstérmes lett a nemzetközi szempontból is kiemelkedő CoNNL-SIGMORPHON osztott feladatsoron (Ács 2018)
- megalkottuk a Webkorpusz 2.0-t, ami az eddigi legnagyobb, nyilvánosan elérhető és dinamikusan bővülö magyar korpusz
- elkészült és nyilvánossá vált (open source) az első monolinguális magyar BERT modell (HuBERT)
- az eredményeket integráltuk az e-Magyar eszközláncba (Nemeskey 2020)
- sikerült morféma-jellegű alakok felismerése (BlackBox AI) az un. SoPa modellekben (Ács 2019)
- a szószintűnél alacsonyabb tokenizáció hatásait vizsgálja dinamikus beágyazásokban Ács (bírálat alatt)

A projekt közepén bekövetkezett irányváltás miatt a jelen összefoglaló elsősorban az utolsó két futamév eredményeivel foglalkozik. Ács (2018) rendszere annyiban már anticipálta a későbbi a dinamikus modelleket, hogy a tanításban már a BME-HAS rendszerben is központi szerepet tölt be a

figyelmi (attention) mechanizmus. Ezzel a munkával párhuzamosan 2019-re Nemeskey felépítette az eddigi legnagyobb, nyilvánosan elérhető és dinamikusan bővülő magyar korpuszt, a WebKorpusz 2.0-t (a 2.1 változat már előkészületben). A projekt utolsó évében feltanította az első, és tudtunkkal máig egyetlen nyilvánosan elérhető monolinguális magyar BERT modellt (HuBERT), mely ezek nemzetközi repozitóriumából, a HuggingFace website-ról letölthető. Eredményeit PhD disszertációjában foglalta össze, ennek munkahelyi vitája már lezajlott, a hivatalos védés még novemberben várható. A modellt sikerrel tanította tovább névszói csoport- és névelemfelismerésre, felülmúlva az eddig nyilvánosságra hozott modelleket. Az eszközöket integrálta az e-magyar elemzőláncba; az erről szóló publikációja különdíjat kapott a XVI. Magyar Számítógépes Nyelvészeti konferencián.

A morfológiai tanulás kérdéseinek vizsgálatát folytatta Ács Judit az alatt az idő alatt amíg a University of Washingtonon (UW) Fulbright ösztöndíjjal volt vendéghallgató, illetve a projekt záróéveben is (2020). Miután 2018-ban ezüstérmes lett a SIGMORPHON shared taskon, erről szóló beszámolója az MSZNY 2020 best paper díját nyerte el. A 19th International Morphology Meeting-en sikerrel adta elő újabb kutatásait, melyben az OTKA projekt eredeti céljainak megfelelően morfémákat talál a (Schwartz et al 2018) által az UW-n bevezetett SoPa (soft patterns) módszerrel. PhD disszertációjának gerincét is morfológiai eredmények fogják alkotni, bár a projekt támogatásával készült el a szintén gépi tanulásos módszeren alapuló, az uráli nyelvek (köztük a magyar) digitális vitalitását taglaló munka is, mely szintén impakt faktoros folyóiratban került publikálásra.

Kovács et al (2019, 2020) az EMNLP konferencián adták elő a felszini alakokat elemzésből rekonstruáló rendszerüket, melynek morfológiai (szintézis) komponense seq2seq modellt használ biLSTM enkóderrel és figyelem- (attention) alapú LSTM dekóderrel, melyet szintén Ács dolgozott ki. A projekt előző fázisának eredményei is részben itt kerülnek felhasználásra illetve tovább fejlesztésre: Döbrössy et al (2019) rész-szó (subword) beágyazási stratégiákat vizsgálnak magyarra, Indig et al (2019) pedig az e-magyar eszközláncot fejlesztik tovább. Lévai és Kornai (2019) a szóbeágyazások és a ragozás kapcsolatát vizsgálja ugyancsak a statikus esetben, Borbély és Kornai (2019) pedig a mondathosszra ad olyan matematikai modellt, amely konzisztens eredményeket ad mind szó-szám mind morféma-szám szerinti méréseknél. Kovács et al (2020) decemberben a COLING konferencián fogják előadni a "SemEval 2020 Task 2: Predicting Multilingual and Cross-Lingual (Graded) Lexical Entailment" feladaton elért eredményeiket, Ács pedig dinamikus be-

ágyazásokra viszi tovább a rész-szó beágyazási vizsgálatokat amiket Döbrössy et al még statikus vektorokon kezdett el.

Publikációk a 2018-as időközi beszámoló óta – az OTKA listára fel nem tölthető elemek itt félkövérrel szedve

- Acs, Judit (2018). "BME-HAS system for CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection." In: Proceedings of the CoNLL SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection, pp. 121–126
- Ács, Judit and András Kornai (2020). "The Role of Interpretable Patterns in Deep Learning for Morphology." In: XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2020). Szeged, pp. 171–179
- Ács, Judit, Dávid Márk Nemeskey, and Gábor Recski (2019). "Building word embeddings from dictionary definitions." In: K + K = 120: Papers dedicated to László Kálmán and András Kornai on the occasion of their 60th birthdays. Ed. by Katalin Mády Beáta Gyuris and Gábor Recski. Research Institute for Linguistics, Hungarian Academy of Sciences (RIL HAS)
- **Ács, Judit and Géza Velkey (2017). "Comparing word segmentation algorithms." In: Proceedings of the Automation and Applied Computer Science Workshop 2017 : AACS'17. Budapest University of Technology and Economics**
- Borbély, Gábor and András Kornai (June 2019). "Sentence Length." In: Proceedings of the 16th Meeting on the Mathematics of Language. Toronto, Canada: Association for Computational Linguistics, pp. 114 –125. url: https://www.aclweb.org/anthology/W19-5710
- Döbrössy, Bálint, Márton Makrai, Balázs Tarján, and György Szaszák (Aug. 2019). "Investigating Sub-Word Embedding Strategies for the Morphologically Rich and Free Phrase-Order Hungarian." In: Proc. 4th Workshop on Representation Learning for NLP (RepL4NLP-2019). Florence, Italy: Association for Computational Linguistics, pp. 187–193. doi: 10.18653/v1/W19-4321.

- Gémes, Kinga, Ádám Kovács, and Gábor Recski (2019). "Machine comprehension using semantic graphs." In: Proc. Automation and Applied Computer Science Workshop 2019 : AACS'19. Ed. by Dmitriy Dunaev and István Vajk. Budapest University of Technology and Economics, pp. 90–98

- Gyenis, Zalán and András Kornai (2019). "Naive probability." In: ArXiv, p. 1905.10924

- Indig, Balázs, Bálint Sass, Eszter Simon, Iván Mittelholcz, Noémi Vadász, and Márton Makrai (Aug. 2019). "One format to rule them all – The emtsv pipeline for Hungarian." In: The 13th Linguistic Annotation Workshop.

- Kornai, András (2019). Semantics. Springer Verlag. isbn: 978-3-319-65644-1. url: http://kornai.com/Drafts/sem.pdf

- Kovács, Ádám, Evelin Ács, Judit Ács, Andras Kornai, and Gábor Recski (2019). "BME-UW at SRST-2019: Surface realization with Interpreted Regular Tree Grammars." In: Proceedings of the 2nd Workshop on Multilingual Surface Realisation (MSR 2019). Hong Kong, China: Association for Computational Linguistics, pp. 35–40. doi: 10.18653/v1/D19-6304

- Kovács, Ádám, Judit Ács, András Kornai, and Gábor Recski (2020). "Better Together: Modern methods plus traditional thinking in NP alignment." In: Proc. LREC 2020, to appear

- **Kovács, Ádám, Gémes Kinga, András Kornai, and Gábor Recski (2020). "BMEAUT at SemEval-2020 Task 2: Lexical entailment with semantic graphs." In: Proceedings of the 14th International Workshop on Semantic Evaluation.**

- Lévai, Dániel and András Kornai (Jan. 2019). "The impact of inflection on word vectors." In: XV. Magyar Számítógépes Nyelvészeti Konferencia

- **Makrai, Márton (2020). "Tárgyas szerkezetek elemzése tenzorfelbontással– áttekintő cikk." In: XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2020). Szeged, pp. 273–287**

- **Nemeskey, Dávid Márk (2020a). "Egy emBERT próbáló feladat." In: XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020). Szeged, pp. 409–418**
- **Nemeskey, Dávid Márk (2020b). "Natural Language Processing Methods for Language Modeling." PhD thesis. Eötvös Loránd University**