Budapesti Műszaki és Gazdaságtudományi Egyetem
Villamosmérnöki és Informatikai Kar
Automatizálási és Alkalmazott Informatikai Tanszék

Morfológia a kontextuális nyelvi modellek korában

Tézisfüzet

Ács Judit

Témavezető:
Kornai András, D.Sc.

Budapest, 2025.

Bevezetés

A természetes nyelvfeldolgozás (NLP) területe az elmúlt 15 évben átvette a mélytanulási módszereket. Manapság a legtöbb NLP feladatban a legkorszerűbb megoldás valamilyen neurális modell, gyakran egy előre betanított nyelvi modell finomhangolt változata. E modellek hatékonyságát különböző angol nyelvű benchmark adatbázisokon, valamint egyre inkább más egynyelvű és többnyelvű adatbázisokon is bizonyítják.

Ebben a disszertációban a mélytanulási modellek alkalmazását vizsgálom alacsony szintű feladatokra, különösen morfoszintaktikai feladatokra több nyelven.

A disszertáció első része (1. és 2. tézis) a mélytanulási modellek alkalmazását tárgyalja klasszikus morfoszintaktikai feladatokra, mint például a morfológiai inflexió és elemzés több tucat nyelven, különös tekintettel a magyar nyelvre.

A disszertáció második része (3–5. tézis) az előre betanított nyelvi modellekkel, főként a BERT (Devlin et al., 2018) család modelljeivel, foglalkozik. Néhány kísérletet bemutatok a GPT-4o és a GPT-4o-mini modellekkel is. Ezek a modellek kiváló teljesítményt mutatnak különböző angol nyelvű és néhány sok erőforrással rendelkező nyelvű feladatokon. Ugyanakkor közepes és kevés erőforrással rendelkező nyelvek esetében a kiértékelésük hiányos. Bemutatok egy módszertanot a morfoszintaktikai benchmarkok generálására tetszőleges nyelveken, és részletesen elemzek több BERT modellt. Az elemzés fő eszköze a szondázó módszertan (Belinkov, 2021).

1. tézis

Az enkóder-dekóder (más néven sequence-to-sequence vagy seq2seq) modellek jól alkalmazhatók morfológiai inflexióra és generálásra. Ez igaz mind a szó-, mind a mondatszintű feladatokra több nyelven.

Hozzájárulások:

- Készítettem egy ezüst standard magyar nyelvű adathalmazt morfológiai inflexió és elemzés céljából, jó minőségű szabályalapú elemző segítségével.

- Implementáltam két típusú seq2seq vagy enkóder-dekóder modellt figyelemmechanizmussal: LSTM puha figyelemmel és LSTM szigorú figyelemmel.
- Betanítottam és kiértékeltem a modelleket a magyar adathalmazon.
- Egyfős csapatként részt vettem a CoNLL-SIGMORPHON 2018 versenyen. Az 1. feladat a szótípus-szintű ragozás volt több mint 100 nyelven, a 2. feladat a kontextusban történő ragozás (mondatszintű) 7 nyelven. A két feladatban 3. és 2. helyezést értem el több enkóderből és egy dekóderből álló seq2seq neurális hálózatokkal.

A tézishez kapcsolódó eredményeket a Ács (2018)-ban publikáltam. A tézis minden eredménye a saját munkám.

2. tézis

A differenciálható neurális mintázatfelismerés több nyelven képes morfoszintaktikai minták kinyerésére, ha morfológiai inflexió és elemzés céljából enkóderként használják.

Hozzájárulások:

- Újraimplementáltam a Soft Patterns vagy SoPa neurális modellt (Schwartz et al., 2018), amely egy korlátozott és differenciálható véges automatákon alapuló modell, eredetileg szövegosztályozásra. Hozzáadtam egy LSTM dekódert, és SoPa-t enkóder-dekóder modellként használtam. A modell az enkóder oldalon mintázatfelismerést alkalmaz.
- A modellt szótípus-szintű morfológiai elemzésre és ragozásra alkalmaztam 12 tipológiailag vegyes nyelven.
- Kinyertem mintázatokat (karaktersorozatokat) a betanított SoPa modellekből, és manuálisan vizsgáltam azok nyelvészeti relevanciáját.
- Bevezettem egy modellhasonlósági mérőszámot két SoPa modell összehasonlítására. Ez lehetővé teszi olyan feladatok összehasonlítását, ame-

lyek ugyanazt a bemeneti adatot használják. Minél nagyobb a hasonlóság két feladat között, annál valószínűbb, hogy ugyanazokat a mintázatokat használják.

Ezeket az eredményeket Ács and Kornai (2020) publikálta. A tanulmány elnyerte a legjobb cikk díját a 2020-as Magyar Számítógépes Nyelvészeti Konferencián. Az implementáció teljes mértékben az én munkám.

3. tézis

A nyers szövegen betanított nyelvi modellek (Pre-trained Language Model vagy PLM) megtanulják a morfológiát. A PLM-ek reprezentációi megtartják a morfoszintaktikai információkat sokféle tipológiailag eltérő nyelv és különböző feladatok esetén is. Ez az információ szondázó (probing) vagy diagnosztikai osztályozók segítségével kinyerhető.

Hozzájárulások:

- Létrehoztam a legnagyobb többnyelvű morfoszintaktikai probing adathalmazt, amely 247 feladatot tartalmaz 42 nyelven 10 nyelvcsaládból. Egy probing mintát egy mondat-token-morfoszintaktikai címke hármasként definiáltam. A minták generálásához a Universal Dependencies Treebank-et (Nivre et al., 2018) használtam.

- Kiértékeltem 6 többnyelvű PLM-et, részletesen elemeztem az mBERTet és az XLM-RoBERTát, valamint összehasonlítottam őket különböző baseline modellekkel.

- Kiértékeltem a GPT-4o és GPT-4o-mini modelleket a probing adathalmaz tesztkészletén, promptolás segítségével.

- Megvizsgáltam a PLM-ek tokenizáló algoritmusát, és megmutattam, hogy bár a többnyelvű modellek több mint 100 nyelvet támogatnak, a tokenizálás jobban működik a latin betűs nyelveken, különösen az angol nyelven (Ács, 2019).

- A probing mint fekete doboz modellek elemző eszközzel kapcsolatban sok kritika merült fel (Belinkov, 2021). Több mint 10 ablációs módszert vezettem be, és kimutattam, hogy a morfoszintaktikai probing következtetései megbízhatóak.

- A PLM-ek token szintű használata megköveteli az összetett tokenek kezelési módját. Valamilyen aggregáló függvényre van szükség ahhoz, hogy egyetlen vektort kapjunk minden egyes tokenhez. Megmutattam, hogy az aggregáló függvény megválasztása számít, különösen akkor, ha nem tanítjuk tovább a modelleket. 9 aggregáló függvényt hasonlítottam össze 7 nyelven és 3 feladaton (Ács et al., 2021a).

Ezeket az eredményeket Ács (2019); Ács et al. (2021a); Ács et al. (2023) cikkekben publikáltuk. A kísérlet tervezését társszerzőimmel együtt végeztem, az implementáció az én hozzájárulásom.

4. tézis

Az egynyelvű PLM-ek jobbak saját nyelvükön, mint a többnyelvű PLM-ek, de a különbség kicsi és gyakran nem statisztikailag szignifikáns. Mind az egynyelvű, mind a többnyelvű PLM-ek sikeresen transzferálhatóak új nyelvekre, amennyiben azok ugyanazt az írásrendszert használják.

Hozzájárulások:

- 5 magyar nyelvet támogató PLM-et elemeztem, és kimutattam, hogy a HuBERT (Nemeskey, 2021) – egy kizárólag magyar nyelvű modell – jobb morfológiai, POS- és NER-címkézésben, mint a többnyelvű modellek, különösen a distilBERT, de a különbség kicsi.

- Kiterjesztettem ezt a vizsgálatot az uráli nyelvcsaládra. A magyarhoz hasonló eredményekre jutottam észt és finn modellek eseten (más uráli nyelvekre nincs dedikált modell).

- POS- és NER-címkéző modelleket tanítottam be kisebbségi uráli nyelveken, többnyelvű és egynyelvű modellek transzferálásával. A transzferált

modellek rendkívül kis tanulóadat ellenére is sikeresek voltak és state-of-the-art eredményeket értek el nyelvspecifikus erőfeszítések nélkül.

Ezeket az eredményeket a Ács et al. (2021); Ács et al. (2021b) cikkekben publikáltuk. A kísérlet tervezését társszerzőimmel végeztem, az implementáció az én hozzájárulásom.

5. tézis

A morfoszintaktikai információ forrása gyakran lokalizált egy mondatban. Bizonyos információk szisztematikus eltávolítása (perturbáció) feltárja, hogy hol tárolódik az információ. A kontextus szerepe morfoszintaxisban Shapley-értékekkel számszerűsíthető.

Hozzájárulások:

- Bevezettem perturbációs módszereket, amelyek eltávolítanak bizonyos információkat a mondatból. A morfoszintaktikai probing modelleket újratanítottam a 247 probing feladaton. Az eredmények betekintést adnak abba, hogy a mondaton belül hol található bizonyos információ. Gyakran találunk ezekre nyelvészeti magyarázatot.
- A perturbációs kísérletek eredményei alapján a nyelvek klaszterezhetők, és az egy nyelvcsaládból származó nyelvek jellemzően egy csoportba kerülnek néhány kivétellel. A nem rokon, de tipológiailag hasonló nyelvek is gyakran egy klaszterbe kerülnek.
- A probing mondatokat egy 9 szereplős játékként definiáltam, ahol a játékosok a tokenek a probing céltokenhez vett relatív pozíciójuk alapján. Erre alkalmaztam a Shapley keretrendszert.
- Elemeztem a Shapley-értékeket az mBERT, az XLM-RoBERTa és egy LSTM baseline modelleke és azt találtam, hogy a modellek közti hasonlóság magas, ami arra utal, hogy a Shapley-értékek inkább a nyelvről adnak információt, mintsem a modellről. Azonosítottam az átlagtól jelentősen eltérő Shapley-értékeket adó probing feladatokat és azt találtam, hogy a Shapley-értékek gyakran esnek egybe a nyelvészeti tulajdonságaival az egyes feladatoknak.

Ezeket az eredményeket (Ács et al., 2023) cikkben publikáltuk. Az implementáció az én munkám.

Hivatkozások

Yonatan Belinkov. 2021. Probing Classifiers: Promises, Shortcomings, and Advances. *arXiv:2102.12452 [cs]*. ArXiv: 2102.12452.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dávid Márk Nemeskey. 2021. Introducing huBERT. In *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2021)*, pages 3–14, Szeged.

Joakim Nivre, Mitchell Abrams, Željko Agić, et al. 2018. Universal Dependencies 2.3. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Roy Schwartz, Sam Thomson, and Noah A. Smith. 2018. SoPa: Bridging CNNs, RNNs, and Weighted Finite-State Machines. In *Proc. 56th ACL Annual Meeting*, pages 295–305, Melbourne, Australia.

Kapcsolódó publikációk

Judit Ács. 2018. BME-HAS system for CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. *Proceedings of the CoNLL SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 121–126.

Judit Ács. 2019. Exploring BERT's vocabulary. http://juditacs.github.io/2019/02/19/bert-tokenization-stats.html. Accessed: 2021-05-14.

Judit Ács, Endre Hamerlik, Roy Schwartz, Noah A. Smith, and Andras Kornai. 2023. Morphosyntactic probing of multilingual BERT models. *Natural Language Engineering*, page 1–40.

Judit Ács, Ákos Kádár, and Andras Kornai. 2021a. Subword pooling makes a difference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2284–2295, Online. Association for Computational Linguistics.

Judit Ács and András Kornai. 2020. The role of interpretable patterns in deep learning for morphology.

Judit Ács, Dániel Lévai, and Andras Kornai. 2021b. Evaluating transferability of BERT models on Uralic languages. In *Proceedings of the Seventh International Workshop on Computational Linguistics of Uralic Languages*, pages 8–17, Syktyvkar, Russia (Online). Association for Computational Linguistics.

Judit Ács, Dániel Lévai, Dávid Márk Nemeskey, and András Kornai. 2021. Evaluating contextualized language models for Hungarian. In *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2020)*, Szeged.

Ádám Kovács, Judit Ács, András Kornai, and Gábor Recski. 2020. Better together: Modern methods plus traditional thinking in NP alignment. In *Proc. LREC 2020*, pages 3635—-3639.

Ádám Kovács, Evelin Ács, Judit Ács, András 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)*, pages 35–40, Hong Kong, China. Association for Computational Linguistics.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021. SIGMORPHON 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, Online. Association for Computational Linguistics.

Gábor Recski, Ádám Kovács, Kinga Gémes, Judit Ács, and Andras Kornai. 2020. BME-TUW at SR'20: Lexical grammar induction for surface realization. In *Proceedings of the Third Workshop on Multilingual Surface Realisation*, pages 21–29, Barcelona, Spain (Online). Association for Computational Linguistics.

Gábor Szolnok, Botond Barta, Dorina Lakatos, and Judit Ács. 2021. BME submission for SIGMORPHON 2021 shared task 0. A three step training approach with data augmentation for morphological inflection. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 268–273, Online. Association for Computational Linguistics.

Egyéb publikációk

Judit Ács. 2014. Pivot-based multilingual dictionary building using Wiktionary. In *The 9th edition of the Language Resources and Evaluation Conference*.

Judit Ács. 2015. Synonym acquisition from translation graph. In *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015)*, pages 14–21. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Judit Ács, Gábor Borbély, Márton Makrai, Dávid Nemeskey, Gábor Recski, and András Kornai. 2018. Hibrid nyelvtechnológiák. *Magyar Tudomány*, 6.

Judit Ács, László Grad-Gyenge, and Thiago Bruno Rodrigues de Rezende Oliveira. 2015. A two-level classifier for discriminating similar languages. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 73–77, Hissar, Bulgaria. Association for Computational Linguistics.

Judit Ács and József Halmi. 2016a. Comparing diacritic restoration methods for Hungarian. In *Proceedings of the Automation and Applied Computer Science Workshop 2016 : AACS'16*. Budapest University of Technology and Economics.

Judit Ács and József Halmi. 2016b. Hunaccent: Small footprint diacritic restoration for social media. In *Proceedings of the First Workshop on Normalisation and Analysis of Social Media Texts (NormSoMe)*.

Judit Ács and Ágota Illyés. 2015. Language detection and generation. In *Proceedings of the Automation and Applied Computer Science Workshop 2015 : AACS'15*. Budapest University of Technology and Economics.

Judit Ács and András Kornai. 2016. Evaluating embeddings on dictionary-based similarity. In *Proceedings of the First Workshop on Evaluating Vector-Space Representations for NLP (RepEval)*, pages 78–82.

Judit Ács, Dávid Nemeskey, and András Kornai. 2017. Identification of disaster-implicated named entities. In *Proceedings of the First International*

Workshop on Exploitation of Social Media for Emergency Relief and Preparedness.

Judit Ács, Dávid Márk Nemeskey, and Gábor Recski. 2017. Building word embeddings from dictionary definitions. In Katalin Mády Beáta Gyuris and Gábor Recski, editors, *K + K = 120: Papers dedicated to László Kálmán and András Kornai on the occasion of their 60th birthdays*. Research Institute for Linguistics, Hungarian Academy of Sciences (RIL HAS).

Judit Ács, Katalin Pajkossy, and András Kornai. 2013. Building basic vocabulary across 40 languages. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 52–58, Sofia, Bulgaria. Association for Computational Linguistics.

Judit Ács 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.

Botond Barta, Endre Hamerlik, Milán Konor Nyist, and Judit Ács. 2025. HuAMR: A Hungarian AMR parser and dataset. In *XXI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2025)*, pages 185–196. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Botond Barta, Dorina Lakatos, Attila Nagy, Milán Konor Nyist, and Judit Ács. 2023. HunSum-1: an abstractive summarization dataset for Hungarian. In *XIX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2023)*, pages 231–243. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Botond Barta, Dorina Lakatos, Attila Nagy, Milán Konor Nyist, and Judit Ács. 2024. From news to summaries: Building a Hungarian corpus for extractive and abstractive summarization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7503–7509, Torino, Italia. ELRA and ICCL.

Dániel Huszti and Judit Ács. 2017. Entitásorientált véleménykinyerés magyar nyelven. In *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY*

2017), pages 240–250. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

András Kornai, Judit Ács, Márton Makrai, Dávid Márk Nemeskey, Katalin Pajkossy, and Gábor Recski. 2015. Competence in lexical semantics. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 165–175, Denver, Colorado. Association for Computational Linguistics.

Attila Nagy, Bence Bial, and Judit Ács. 2021. Automatic punctuation restoration with BERT models. In *XVIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2021)*.

Attila Nagy, Dorina Lakatos, Botond Barta, and Judit Ács. 2023a. TreeSwap: Data augmentation for machine translation via dependency subtree swapping. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 759–768, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Attila Nagy, Dorina Petra Lakatos, Botond Barta, Patrick Nanys, and Judit Ács. 2023b. Data augmentation for machine translation via dependency subtree swapping. In *XIX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2023)*. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Attila Nagy, Patrick Nanys, Konrád Balázs Frey, Bence Bial, and Judit Ács. 2022. Syntax-based data augmentation for Hungarian-English machine translation. In *XVIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2022)*. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Gergely Dániel Németh and Judit Ács. 2018. Hyphenation using deep neural networks. In *XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018)*. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Gábor Recski and Judit Ács. 2015. MathLingBudapest: Concept networks for semantic similarity. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 138–142, Denver, Colorado. Association for Computational Linguistics.