Az EFNILEX és egy fiatal kutató. Hat év magyar szóbeágyazásokkal

Makrai Márton[1,2]

[1] BME VIK Távközlési és Médiainformatikai Tanszék
[2] MTA TTK Kognitív Idegtudományi és Pszichológiai Intézet
makrai.marton@ttk.hu

1. Európai szótárak egynyelvű korpuszból

Az EFNILEX projekt azt szándékozott felderíteni, hogy a gépi fordítás eszközei hogyan járulnak hozzá szótárak előállításához „közepes" európai nyelveken, vagyis az EU kevesebb nyelvtechnológiai erőforrással rendelkező hivatalos nyelvein. Héja Enikőtől vettem át a stafétát 2014-ben, aki – ahogy ebben a kötetben is írja – párhuzamos korpuszokból készített szótárakat.

Ezekben az években volt a nyelvtechnológia neurális forradalmának első, sekélyebb hulláma, az előtanított (de nem kontextualizált, nem igazán mély) szóbeágyazásoké. A *beágyazás* szó arra utal, hogy a szimbolikus készlet elemei (esetünkben: szókincs) a neurális hálókban valós vektorként vannak reprezentálva. A 2013–14-es hullám első cikkei több mindenre rámutattak: Ezek a módszerek a korábbinál sokkal hatékonyabban állítanak elő egynyelvű szemigigakorpuszból – amilyen magyarra nem utolsó sorban az MNSz.[2] (Oravecz és mtsai., 2014) – jó minőségű szóreprezentációkat: a jelentésben vagy morfoszintaktikailag hasonló szóalakokat egymáshoz közel helyezik el egy pár száz dimenziós euklideszi térben (Mikolov és mtsai., 2013a, 2013b). Továbbá a lexikai szemantika régi álmához, a szótári felbontáshoz is közelebb vittek: Mikolov és mtsai. (2013c) híres és sokat vitatott (Levy és mtsai., 2015; Linzen, 2016) példájával a *királynő* fogalma egy *királyi* és egy *női* elemből áll. Végül – és az EFNILEX e szakasza szempontjából első sorban – a különböző nyelvek egynyelvű korpuszaiból tanított modellek között olyan hasonlóság áll fönn, amely lehetővé teszi az úgynevezett lineáris fordítást, vagyis hogy egy néhány ezer szavas magszótárból föltanítsunk egy lineáris leképezést a két nyelv szavainak vektortere között, amellyel a forrásnyelvi w_s szót a célnyelvi térbe képezve egy olyan vektort kapunk, amelyhez legközelebbi célnyelvi szóvektor formájában megtaláljuk w_s célnyelvi megfelelőjét.

DOI: https://doi.org/10.18135/VT70.12

Enikő bátorított minket, hogy alkalmazzuk a lineáris fordítás módszerét az EFNILEX-ben. A szóvektorok kiértékelésének egyik legnépszerűbb módszerét az analógiás kérdések jelentik, pl. *férfi : nő :: király : ?,* a várt válasz a *királynő.* Elkészítettük és nyíltan közreadtuk (Makrai, 2014) az egyik fő (angol) teszthalmaz magyar megfelelőjét.

Ezeknek a sekély neurális háló segítségével előállított modelleknek a mai mély előtanított neurális nyelvmodellek kontextualizált szóbeágyazásaival szemben (lásd az utolsó szakaszt) az volt a hátrányuk, hogy egy szóalakot, legyen az *poliszém* vagy akár *homonim*, egyetlen vektor reprezentált, a különböző jelentések a legjobb esetben szuperponálódtak, vagy a ritkább jelentés elveszett, esetleg káros módon keveredtek. Prószéky Gábor példájával élve: tudomásul kellett venni, hogy a *daru* egy olyan entitás, amelyik olykor fészket rak, máskor betondarabokat emelget. 2015-től az EFNILEX-ben, majd fiatal kutatóként ezen a hiányosságon igyekeztünk-igyekeztem javítani kétféleképpen.

2. Szótári háromszögek egyértelműsítése

A gépi szófordítás (avagy szótárindukció) egyik bevett eszköze az úgynevezett *háromszögelés (triangulation)* vagy *sarokkőmódszer (pivot-based method).* Abból, hogy a cseh *zvíře* angol fordítása *animal*, az *animal* magyar fordítása pedig *állat*, arra lehet következtetni, hogy a *zvíře* magyarul *állat*. Ebbe a módszerbe több úton is zajt hoz a többértelműség. A középső nyelv homonímiái hamis háromszögeket vezetnek be (német *was* – magyar *mi* – angol *we*). A vektoros módszer viszont csak a forrás- és a célnyelv többértelműségeire érzékeny, így a két hiba kompenzálja egymást. Makrai (2016) hamis háromszögeket szűrt ki szóvektorok segítségével a német–magyar nyelvpáron. Megmutattuk, hogy a lineáris leképezésből kapott pontszámok simább mértékét adják a fordítások jóságának, mint ha csak megszámoljuk, hogy hány nyelven keresztül háromszögelhető az adott szópár. A nyíltan közreadott, megbízhatósági pontszámokkal ellátott német–magyar erőforrás tudomásunk szerint a legnagyobb szabad elérésű szólista volt akkor.

3. Egy fiatal kutató és a túl finom jelentéskészlet

2015-től 2018-ig az intézet fiatal kutatója voltam Tamás vezetésével. Nagyon hálás vagyok, hogy 3 éven át teljes állásban kutathattam a témámat, és a legjobb konferenciákon publikálhattam az eredményeket. Mint

mondtuk, a szóvektorok a szokásos esetben egy-egy szóalakhoz tartoznak, így a többértelmű szavak vektora rosszabb minőségű. Ezt a problémát hivatottak megoldani a többjelentésű szómodellek *(multi-sense word embedding, MSE)*, amelyek a szóalakok különféle jelentéseit különböző vektorokkal ábrázolják. Ebben a paradigmában annak a megállapítása is a felügyeletlen modell feladata, hogy mely szavak többértelműek, és azoknak hány jelentése van. Az alkalmazásban legjobbnak bizonyuló modellek vektorai közül azonban sok nem felel meg a motiváló várakozásoknak: jobb esetben olyan jelentések között tesznek különbséget, melyeket intuitíve ugyanazon jelentés különböző kontextusokban való használatának tekintenénk, vagy akár puszta zajt képviselnek.

Ezért a szerzőtársaimmal (Borbély és mtsai., 2016) két új módszert javasoltunk az MSEk szemantikai szemcsésségének mérésére. Az egyik egynyelvű szótárakat használ, a másik pedig azon az elven alapszik, hogy egy szó akkor többértelmű, ha a feltételezett jelentések más nyelvre való fordítása különböző. Az utóbbit Makrai és Lipp (2019) bontotta ki két pontosságmértéket formalizálva. Az egyik bünteti a duplumokat, a másik pedig azért van, hogy a vektorok ne mossanak össze jelentéseket. A kísérletek igazolták, hogy a két mérték között csereviszony van: minél specifikusabb egy vektor, annál könnyebb lefordítani, csak persze ha túl specifikus, akkor egybeeshetnek a fordítások. Tehát a két mérték számszerűsíti, hogy egy többjelentésű szóbeágyazás mennyire jól ragadja meg a lexikai struktúrát (Borbély és mtsai., 2016; Makrai és Lipp, 2019).

A kutatás egy másik ágában Berend Gáborral hipernimákat (a fölérendelt fogalmat, pl. hogy a kutya egy állat) nyertünk ki szóvektorokból. Ritka szóreprezentációkon alapuló módszerünkkel megnyertünk több kategóriát a szakma évente megrendezésre kerülő legrangosabb versenyének (SemEval) egyik feladatában. Abban az évben New Orleansben volt a SemEval, így nem tudtam volna prezentálni a posztert, ha nincs az a nagyon stabil és bőséges anyagi keret, amit Tamás biztosított.

Bár nem tartozott projektbe, témája miatt kedves volt Tamásnak egy olyan cikk, amely egy általam bírált szakdolgozatból született: szónál kisebb elemek beágyazásán alapuló magyar nyelvmodelleket hasonlított össze a hallgató (Döbrössy és mtsai., 2019).

4. Evezz a mélyre

2018-ban újabb, mélyebb hullámot vetett a nyelvtechnológia (NLP) neurális forradalma. Mély neurális hálóval való tanulás alatt azt értjük,

hogy a gépi tanulás eredménye egy olyan számítási modell, amely rétegekből áll, és az input rétegtől rejtett rétegeken át az output réteg felé haladva egyre magasabb szintű jellemzőket számít ki. A mélytanulás először a beszédtechnológiában (Dahl és mtsai., 2011) és a gépi látásban (Krizhevsky és Sutskever, 2012) hozott áttörtést. 2018-ban az NLP-ben is elérkezett az, amit Sebastian Ruder *ImageNet pillanat*nak[1] nevez.

> „A gépi látás (*computer vision*, CV) kutatóközössége évek óta tanít fel teljes modelleket alacsony és magas szintű jellemzők előtanításával. Leggyakrabban ez úgy történik, hogy a nagy ImageNet adatkészlet képeinek osztályozását tanítják meg. Az ULMFiT, az ELMo és az OpenAI transzformer most elhozta a nyelv ImageNet-jét, vagyis egy olyan feladatot, amely lehetővé teszi a modellek számára, hogy a nyelv magasabb szintű aspektusait is megtanulják a modellek, hasonlóan ahhoz, ahogy az ImageNet lehetővé tette olyan CV-modellek feledzését, amelyek a képek általános célú jellemzőit tanulják meg."

Az utóbbi két évről kiváló áttekintést adnak Qiu és mtsai. (2020). A számítógépes nyelvész számára különösen érdekes a modellek nyelvészeti tudásának letapogatására irányuló kutatás, amit Rogers és mtsai. (2020) foglalnak össze. 2020-ban elindult egy magyar mély nyelvmodellek létrehozására, kiértékelésére, és nyelvészeti tartalmának felderítésére irányuló projekt is (HILBERT, Feldmann és mtsai., 2021).

Bibliográfia

Borbély, G., Makrai, M., Nemeskey, D. M., Kornai, A.: Evaluating multi-sense embeddings for semantic resolution monolingually and in word translation. In: Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP. pp. 83–89. Association for Computational Linguistics, Berlin (2016), http://www.aclweb.org/anthology/W16-2515

Dahl, G. E., Yu, D., Deng, L., Acero, A.: Large vocabulary continuous speech recognition with context-dependent dbn-hmms. In: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. pp. 4688–4691. IEEE (2011)

Döbrössy, B., Makrai, M., Tarján, B., Szaszák, G.: Investigating sub-word embedding strategies for the morphologically rich and free phrase-order Hungarian. In: Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019). pp. 187–193. Association for Computational Linguistics, Florence, Italy (2019) https://www.aclweb.org/anthology/W19-4321

[1] https://ruder.io/nlp-imagenet/

Feldmann, Á., Váradi, T., Hajdu, R., Indig, B., Sass, B., Makrai, M., Mittelholcz, I., Halász, D., Zijian, G. Y.: HILBERT, magyar nyelvű bert-large modell tanítása felhő környezetben. In: Berend G., Gosztolya G., Vincze V. (szerk.) XVII. Magyar Számítgépes Nyelvészeti Konferencia, pp. 29–36. Szegedi Tudományegyetem TTIK, Informatikai Intézet, Szeged (2021) MSZNY (2021)

Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 25 (NIPS 2012) (2012)

Levy, O., Remus, S., Biemann, C., Dagan, I.: Do supervised distributional methods really learn lexical inference relations? In: Mihalcea, R., Chai, J., Sarkar, A. (eds.) Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 970–976. Association for Computational Linguistics (2015)

Linzen, T.: Issues in evaluating semantic spaces using word analogies. In: RepEval (2016)

Makrai, M.: Deep cases in the 4lang concept lexicon. In: Tanács, A., Varga, V., Vincze, V. (szerk.) X. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2014). pp. 50–57. Szegedi Tudományegyetem, Szeged (2014)

Makrai, M.: Filtering wiktionary triangles by linear mapping between distributed models. In: LREC (2016)

Makrai, M., Lipp, V.: Do multi-sense word embeddings learn more senses? In: Gyuris, B., Mády, K., Recski, G. (eds.) K + K = 120 Workshop Dedicated to László Kálmán and András Kornai on the occasion of their 60th birthdays. pp. 385–398. (2019) MTA Research Institute for Linguistics, Budapest,

Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation (2013a), arXiv preprint arXiv:1309.4168

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. (eds.) Advances in Neural Information Processing Systems 26, pp. 3111–3119. Curran Associates, Inc. (2013b), https://bit.ly/39HikH8

Mikolov, T., Yih, W.t., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013). pp. 746–751. Association for Computational Linguistics, Atlanta, Georgia (2013c)

Oravecz, C., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In: Calzolari, N. et al. (eds.) Proceedings of Ninth International Conference on Language Resources and Evaluation (LREC 2014). pp. 1719–1723. Reykjavik. ELRA. (2014) http://www.aclweb.org/anthology/L14-1536

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., Huang, X.: Pretrained models for natural language processing: A survey. arXiv preprint arXiv:2003.08271 (2020)

Rogers, A., Kovaleva, O., Rumshisky, A.: A primer in bertology: What we know about how bert works. arXiv preprint arXiv:2002.12327 (2020)