Látens szemantikus eloszlások használata a nyelvi modellek előtanítása során

Berend Gábor

Szegedi Tudományegyetem, Informatikai Intézet
berendg@inf.u-szeged.hu

Kivonat Cikkünk egy olyan variánsát mutatja be a nyelvi modellek előtanításának, amely során a maszkolás tárgyául nem a véletlenszerűen kiválasztott tokenek rekonstruálását, hanem azok szemantikus kategóriájának megállapítását tűzzük ki célul. A javasolt módon létrehozott modelljeink finomhangolását változatos benchmarkokon elvégezve azt találjuk, hogy azok szignifikánsan jobb eredmény elérésére képesek hagyományos társaikhoz képest.
Kulcsszavak: előtanítás; szemantikus klaszterek

1. Bevezetés

A komoly eredményeket felmutató modern természetesnyelv-feldolgozó modellek alapjául tipikusan a nyelvi modellezési feladat ellátására önfelügyelt módon létrehozott, transzformer architektúrát használó (Vaswani és mtsai, 2017) modellek szolgálnak. A modellek a nyelvi modellezés elsajátítása során először egy általános nyelvmodellező képességre tesznek szert, amit aztán egy-egy specializált feladat megoldására transzferálhatunk a célfeladat doménjéből jövő felügyelt tanítópéldák felhasználása segítségével. Az előbbi lépést szokás szerint előtanításnak, míg az utóbbit a célfeladatra történő finomhangolásnak hívjuk.

Az előtanítás során két fő megközelítést szokás alkalmazni: az autoregresszív, valamint a maszkolásos módszereket. Az autoregresszív modellek a megkezdett szövegek minél valósághűbb folytatását tűzik ki feladatukul, a céljuk tulajdonképpen a minél kisebb perplexitással rendelkező szövegek generálása.

A maszkolást alkalmazó módszerek ezzel szemben egy teljes mondatot (vagy szövegrészt) kapnak inputul a tanítás során, majd az ezekből a mondatokból véletlenszerűen kitakart szavak (illetve szótöredékek) rekonstruálását tűzik ki célukul az előtanítás során, ami gyakorlatilag az autoenkóderek működéséhez hasonló viselkedést eredményez – annyi különbséggel, hogy a maszkolásos nyelvi modellek nem folytonos értékek, hanem diszkrét szimbólumok rekonstruálására törekszenek.

Munkánkban a maszkolással tanított nyelvi modellek előtanításának egy alternatíváját kívánjuk bemutatni, amely során nem a konkrét kitakart szó(töredék)et, hanem annak (látens) szemantikai jellemzésének megállapítását adjuk a nyelvi modellezést elvégző neurális háló feladatául. Noha cikkünkben az autoregresszív modellekkel nem foglalkozunk, a javasolt eljárás azokra is kiterjeszthető lenne.

A javasolt módszert motiválandó, vegyük azt az egyszerű mondatot, hogy „`Mari süteményt eszik`". Amennyiben az előtanítás során a kitakarásra véletlenszerűen kiválasztott szimbólum a „`süteményt`" szó, úgy a neurális modellünk veszteségfüggvénye abban az esetben lesz a lehető legkisebb *ezen példa mentén*, ha a háló a kitakart szó vonatkozásában kizárólag a „`süteményt`" szót tartaná behelyettesítésre alkalmasnak, míg más – hétköznapi tudásunkkal egyébként teljességgel összeegyeztethető – szavak (mint pl. „`sóletet`", „`disznósajtot`", *stb.*) vonatkozásában a kitakart szó helyén való előfordulás eshetőségét teljességgel kizárná.

A nyelvi modell végső célja szempontjából egy természetesebb, és vélhetőleg jobb mintahatékonysággal bíró – gyakorlati szempontból ugyanakkor jóval problematikusabb – célfüggvény a modell által egy-egy adott szó kitakarása mentén visszaadott, a szótár elemei fölött értelmezett eloszlást azzal a – jellemzően nem a teljes valószínűségi tömeget egy adott szó mentén tartalmazó – eloszlással vethetné össze, amely a kitakart szó adott kontextusa mentén az egyes szavak helyettesítőként való előfordulásának valószínűségét tartalmazná. Kellő megbízhatóságú tanítóadatot minderre a nyelvhasználatot jellemző adatritkaságból fakadóan szinte lehetetlen lenne találni.

Egy alternatív előtanítási folyamatot úgy is elképzelhetünk, hogy valamilyen szemantikai erőforrásra támaszkodva (pl. WordNet (Fellbaum, 1998; Miháltz és mtsai, 2008) vagy ConceptNet (Speer és Havasi, 2012)) a kitakart szavak ontologikus tulajdonságainak meghatározását várnánk el a nyelvi modellek előtanítása során. Ebben az esetben a korábban látott példamondat („`Mari süteményt eszik`") esetén a `sütemény` szó kimaszkolása esetén nem a konkrét szó rekonstruálása lenne a feladatunk, hanem például annak meghatározása, hogy a kitakart szó helyén egy *ehető* fogalom állt. Egy ilyen módon működő előtanítás azonban feltételezi egy kellően expresszív tudásbázis meglétét, mi több, a nyelvi modell tanítására használt szövegek vonatkozásában az az információ is rendelkezésre kell álljon, hogy az egyes szóelőfordulásokra az éppen adott kontextusuk mentén a különböző lehetséges jelentéseik közül mely szemantikus tulajdonság(ok) teljesül(nek).

Ilyen szemantikus részletességgel annotált tanítóadatbázis meglétében az előtanításhoz szükséges mennyiségben aligha reménykedhetünk. Az általunk javasolt módszer éppen ennek a szemantikus annotációval ellátott előtanító anyag kiváltására tesz javaslatot oly módon, hogy a kontextuális szóreprezentációk alapján felügyelet nélküli módon látens tulajdonságok teljesülését rendeljük az egyes szóalakokhoz, megspórolva ezzel a módosított előtanításhoz szükséges szemantikus emberi annotáció elvégzésének szükségességét.

2. Kapcsolódó munkák

A különböző transzformer architektúrák (Vaswani és mtsai, 2017) mára meghatározó jelentőségűvé váltak a mesterséges intelligencia számos területén. A kontextusérzékeny jelentésreprezentációkat megalkotni képes BERT modell (Devlin és mtsai, 2019) mutatott rá először az angol nyelv vonatkozásában az – erede-

tileg gépi fordításra kifejlesztett – architektúra általános és változatos nyelfeldolgozási feladatokon való alkalmazhatóságára. A modell sikerein felbuzdulva, mára rengeteg nyelv támogatására készítettek BERT variánsokat (Vilares és mtsai, 2021; Martin és mtsai, 2020; Le és mtsai, 2020; Ulčar és Robnik-Šikonja, 2020). A nyelvspecifikus BERT modellek egy számunkra kifejezett fontossággal bíró képviselője a dedikáltan a magyar nyelv feldolgozására specializált HuBERT (Nemeskey, 2021).

A különböző forrásokból származó külső tudások előtanított nyelvi modellekbe történő integrálására számos kísérlet született már (Mihaylov és Frank, 2018; Bauer és mtsai, 2018; Peters és mtsai, 2019; Ye és mtsai, 2019; Yang és mtsai, 2019; Qiu és mtsai, 2019; Levine és mtsai, 2020; Liu és mtsai, 2020). A javasolt megoldások széles spektrumon mozognak aszerint, hogy a külső tudást milyen módon használják föl, az azonban közös bennük, hogy mind valamilyen explicit tudásra, pl. manuálisan létrehozott tudásbázisokra támaszkodnak.

A mi megközelítésünk alapvetően abban különbözik a korábbi megoldásoktól, hogy az nem igényel explicit külső szemantikus erőforrást, az előtanulás során használt szemantikus kategóriákat felügyelet nélkül hozzuk létre. Ezek az adatvezérelt szemantikus kategóriák bizonyos tekintetben hasonlítanak a Brown klaszterekre (Brown és mtsai, 1992), fontos különbség azonban, hogy míg a Brown klaszterek kialakítására a szóalakok szintjén kerül sor, addig a mi módszerünkkel a szó*előfordulások* szintjén kaphatunk egy szemantikus kategorizációt.

3. Módszer

Az egyes szóelőfordulásokhoz tartozó látens szemantikus kategóriákat kontextusérzékeny módon előállító eljárásunkat a Berend (2020) által javasol módon hajtottuk végre. Az eljárás első lépése, hogy a transzformer modell által a felügyelet nélküli modellalkotáshoz fölhasznált mondatok inputszimbólumaihoz rendelt, egységnyi hosszúvá normalizált h-dimenziós rejtett állapotaiból képzett $\boldsymbol{X} \in \mathbf{R}^{h \times n}$ mátrixra az alábbi optimalizálási feladatot oldjuk meg

$$\min_{\boldsymbol{D},\boldsymbol{\alpha} \in \mathbb{R}^{k \times n}_{\geq 0}} \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{\alpha}\|_F^2 + \lambda\|\boldsymbol{\alpha}\|_1, \tag{1}$$

amely alapján egy későbbiekben látott $\boldsymbol{x}_i \in \mathbf{R}^h$ kontextuális reprezentációs vektorhoz már a

$$\min_{\boldsymbol{\alpha_i} \in \mathbb{R}^{k}_{\geq 0}} \frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{D}\boldsymbol{\alpha_i}\|_2^2 + \lambda\|\boldsymbol{\alpha_i}\|_1 \tag{2}$$

módon határozható meg annak látens szemantikus jellemzését adó $\boldsymbol{\alpha_i} \in \mathbb{R}^{k}_{\geq 0}$ vektora.

Fontos különbség (1) és (2) között, hogy az utóbbi esetében $\boldsymbol{D}$-ben nem optimalizálunk, azaz a kontextuális modell használata során kapott rejtett vektorokat a korábbi lépésben meghatározott $\boldsymbol{D}$ mátrixra támaszkodva igyekszünk ℓ_1 regularizáció használata mellett minél pontosabban felírni a $\boldsymbol{D}$-beli vektorok lineáris kombinációjaként.

A regularizációs tag azt eredményezi, hogy a $\boldsymbol{\alpha_i}$-beli együtthatók között több is pontosan 0 értéket fog fölvenni, míg a nemnulla együtthatók elhelyezkedése olyan lesz, hogy abból szemantikus tulajdonságok olvashatók ki (Berend, 2020).

Mivel az $\boldsymbol{\alpha_i}$-beli együtthatókkal kapcsolatban egy nemnegativitási feltételt is kikötöttünk, $\boldsymbol{\alpha_i}$ vektorra (a benne található együtthatók összegével való normalizálást követően) viszonylag természetes módon egy olyan eloszlásként is tekinthetünk, ami azt az információt hordozza, hogy egy kérdéses szóelőfordulás mekkora valószínűséggel tekinthető a k különböző felügyelet nélküli módon az (1) segítségével meghatározott látens szemantikus kategóriába tartozónak.

A módosított előtanítási eljárás során a humán annotáció útján rendelkezésre álló szemantikus tudás hiányát ezzel a (normalizáláson átesett) $\boldsymbol{\alpha_i}$ vektorral helyettesíthetjük. Mindehhez a transzformer alapú nyelvi modellünk szótárába fölveszünk k új speciális szimbólumot, amelyek mindegyike egy-egy látens szemantikus kategóriának felel meg, és ahelyett, hogy a maszkolásos nyelvi modellezés során a kitakart szó(töredék) identitásának helyreállítására törekednénk, a k speciális szimbólumra támaszkodva azt vizsgáljuk meg a KL-divergencia használatával, hogy a modell által a k speciális szimbólumra visszaadott multinomiális eloszlása mennyire tér el az adott szóelőfordulás mentén a (2) alapján meghatározott $\boldsymbol{\alpha_i}$ szemantikus eloszlástól.

4. Kísérletek

Kísérleteinket a HuBERT modellen (Nemeskey, 2021) végeztük a `transformers` (Wolf és mtsai, 2020) könyvárra támaszkodva. Az (1) megoldásához a Hungarian Webcorpus 2.0 (Nemeskey, 2020) Wikipédiáról származó 100 000 véletlenszerűen kiválasztott mondatából azon szóelőfordulások kontextuális reprezentációiból alkottuk meg az $\boldsymbol{X}$ inputmátrixot, amelyek olyan szóalakokhoz tartoztak, amelyeknek az adott mintán belüli előfordulása nem haladta meg az 1 000 darabot.

A kiválasztott mondatokban szereplő mintegy $2,7$ millió szó(töredék) nagyjából kétharmadát, közel $1,8$ millió szó(töredéke)t használtunk föl a $\boldsymbol{D}$ mátrix meghatározásához. A λ, illetve k hiperparamétereket (Berend, 2020) alapján rendre 0,05-nek, illetve 3000-nek választottuk, a rejtett kontextuális reprezentációkat pedig a transzformer modell utolsó (tizenkettedik) rétegéből vettük.

Az előtanítást a Hungarian Webcorpus 2.0 20 millió véletlenszerűen kiválasztott mondatán hajtottuk végre gradiens akkumuláció használata mellett, 2048-as virtuális batchmérettel. Az előtanított modellt a HuBERT modell súlyaival inicializáltuk, és a javasolt előtanítási módszert használva 10 000 modellfrissítési lépést hajtottunk végre. Az előtanítás során a tanulási ráta, valamint az egyes tokenek kimaszkolási valószínűsége gyanánt az irodalomban gyakran alkalmazott 0,0001, illetve 0,15 értékeket használtuk.

Baselineként egy olyan, a klasszikus – szemantikus eloszlások helyett szótöredéken alapuló – maszkolt nyelvi modellezési feladattal továbbtanított modellt is létrehoztunk, ami a célfeladattól eltekintve az általunk javasolt modellel teljesen azonos módon került továbbtanításra.

A továbbiakban a szemantikus eloszlások mentén a KL-divergenciával továbbtanított modellünket HuBERTkl-ként, a klasszikus maszkolt nyelvi modellezéssel továbbtanított baseline modellünket pedig HuBERTmlm-ként fogjuk hivatkozni. A különböző célfüggvények használatával továbbtanított modellek összehasonlíthatóságát azzal teremtettük meg, hogy mindkét modellt ugyanazokkal a kezdeti súlyokkal inicializáltuk, a frissítésük során pedig megegyező tartalmú batcheket dolgoztunk föl.

A modellünk alapját képző szemantikus kategóriákat kialakító felügyelet nélküli módszert először kvalitatív szempontból vizsgáljuk, majd az erre építő előtanítással létrehozott modellünk kvantitatív kiértékelését mutatjuk be különböző feladatokon történő finomhangolás mentén.

4.1. Kvalitatív eredmények

Ebben a fejezetben azt kívánjuk illusztrálni, hogy az (1) megoldásával előálló látens szemantikus kategóriák, illetve eloszlások alkalmasak a szóelőfordulások jelentéseik mentén történő csoportosítására, illetve a kontextuális hasonlóság mérésére. Kvalitatív vizsgálatainkat az 1. táblázatban található mondatokban kiemeléssel jelzett szóelőfordulások mentén közöljük. A mondatokban található | szimbólumok a tokenizálás során kialakuló szótöredékek közötti határokat jelölik. A több szótöredék alkotta szavak reprezentálására a szótöredékekhez tartozó kontextuális vektorok átlagát határoztuk meg.

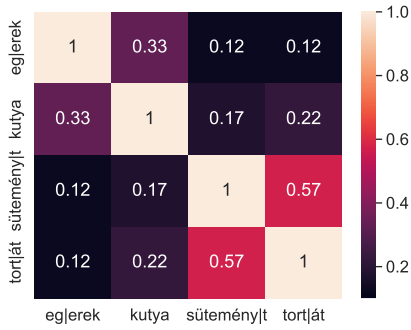
Az **eg|erek** szeretik a sajtot|.
A **kutya** megker|get|te a macsk|át|.
Mari **sütemény|t** eszik|.
Klára nem kér több **tort|át|**.

1. táblázat. Példamondatok, amikben a kiemelt szavak viselkedését vizsgáljuk, a | pedig a szótöredékek határát jelzi.

Megvizsgáltuk, hogy a példánkban vizsgált szópárok mentén miként alakulnak a koszinusz hasonlóságok, ha annak alapját az eredeti transzformer modell által létrehozott kontextuális reprezentációk, vagy a (2) alapján létrejövő $\boldsymbol{\alpha_i}$ ritka vektorok képzik. A páronkénti összehasonlítás eredményeit az 1. ábra foglalja össze. Jól látható, hogy míg a közvetlenül a transzformer modell rejtett állapotából jövő $\boldsymbol{x}_i$ vektorok mentén történő hasonlósági értékek homogén viselkedést mutatnak (1a. ábra), addig a ritka $\boldsymbol{\alpha_i}$ vektorok mentén történő hasonlóságszámítás esetén a szóelőfordulások szemantikus csoportjainak kialakulása figyelhető meg (1b. ábra).

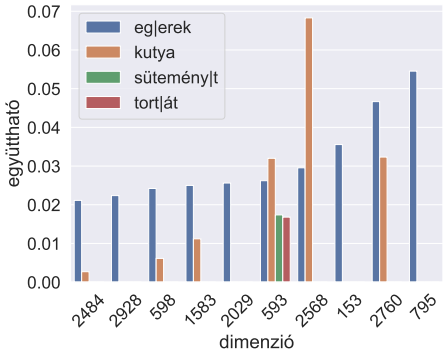
A 2. ábrán az látható, hogy a vizsgált szavakhoz a (2) segítségével kapott szemantikus eloszlások mely látens dimenziók mentén veszik föl a legnagyobb értékeket, illetve, hogy ugyanezen dimenziók mentén a többi vizsgált szó milyen értékekkel rendelkezik. Jól kivehető, hogy a hasonló jelentéssel bíró szóelőfordulásokhoz hasonló szemantikus eloszlások kerültek meghatározásra.

(a) Az eredeti modell $\boldsymbol{x}_i$ vektorai alapján

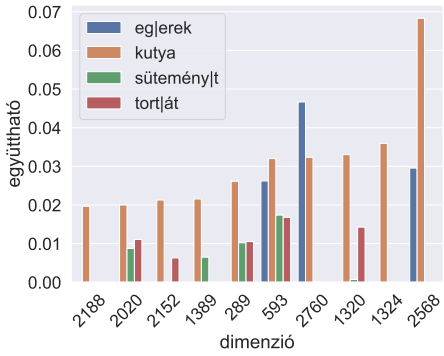


(b) A ritka $\boldsymbol{\alpha}_i$ vektorok alapján

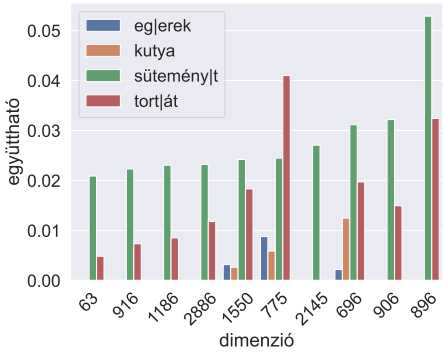
1. ábra: A vizsgált szópárok közötti páronkénti koszinusz hasonlóságok.



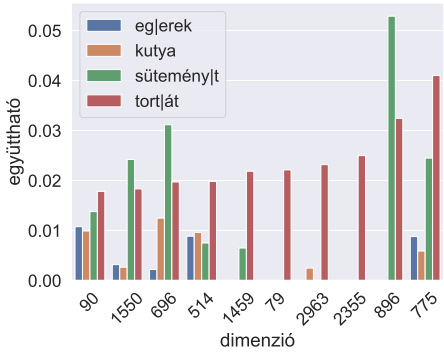
(a) eg|erek



(b) kutya



(c) sütemény|t



(d) tort|át

2. ábra: A vizsgált szóelőfordulásokhoz rendelt szemantikus eloszlások domináns együtthatóinak vizsgálata. Minden ábrán az adott szóelőfordulás mentén kialakuló 10 legmagasabb együtthatóval bíró dimenziója látható, illetve a többi szó ugyanezen dimenziók mentén való viselkedése.

4.2. Kvantitatív eredmények

A létrehozott modelleket változatos célalkalmazási feladatokra való finomhangolás során is kiértékeltük. A kiértékelés során a szakirodalomból megszokott hiperparamétereket választottunk ki: az előtanított modellek finomhangolását 32-es batchmérettel, 0,00002 tanulási ráta mellett hajtottuk végre 3 epochon keresztül az összes vizsgált adatbázison. Mivel a neurális modellek finomhangolása során kapott eredmények nagy szórással rendelkeznek (Dodge és mtsai, 2020), célszerű a finomhangolás során kapott eredményeket több véletlenszerűen inicializált osztályozó fej tekintetében vizsgálni. Éppen ezért, minden vizsgált adathalmazon mindkét modellünket $50 - 50$ kiértékelésnek vetettük alá, eredményeinket több formában is közöljük.

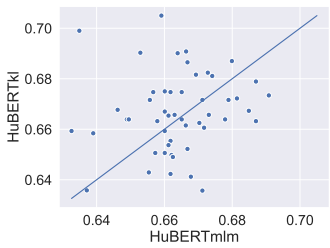
A HuLU (Ligeti-Nagy és mtsai, 2022) három alfeladatán (HuCOLA, HuCoPA, HuSST2), valamint az OpinHuBank (Miháltz, 2013) szentimentosztályozási feladatán az egyes modellek által elért eredmények összefoglalását a 2. táblázat tartalmazza. Láthatjuk, hogy az átlagos teljesítmény tekintetében a szemantikus eloszlások segítségével létrehozott HuBERTkl modell minden vizsgált feladaton jobban teljesít a baselineként használt HuBERTmlm modellnél, amely különbségek az input mondatok grammatikai helyességének eldöntésére irányuló HuCOLA esetét leszámítva szignifikánsnak tekinthetők.

Feladat	Metrika	HuBERTmlm	HuBERTkl	p-érték
HuCOLA	Matthew korreláció	0,664	0,667	0,282
HuCoPA	Pontosság	0,564	0,603	2,4*e-7
HuSST	Pontosság	0,741	0,746	3,1*e-7
OpinHuBank	Pontosság	0,897	0,900	0,005

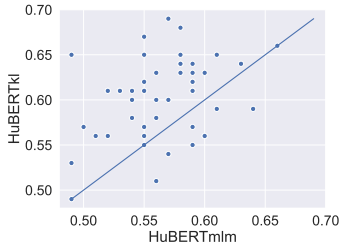
2. táblázat. A modellek átlagos teljesítménye a vizsgált feladatonkénti lebontásban, és modellek teljesítménybeli különbségének szignifikanciájára vonatkozó p-értékek.

A 4. ábra az eltérő módon előtanított, de azonosan inicializált osztályozó fejekkel rendelkező modellpárosok finomhangolás során nyert eredményeinek feladatonkénti összehasonlítását teszi lehetővé. A HuCOLA, HuCoPA, HuSST és az OpinHuBank feladatokon elvégzett modelltípusonkénti 50 kísérletpáros közül rendre 25, 42, 41 és 34 alkalommal tapasztaltuk azt, hogy a HuBERTkl modellre építő finomhangolás megegyezően inicializált osztályozó fejekből kiindulva legalább olyan eredményt produkált, mint a HuBERTmlm modell.

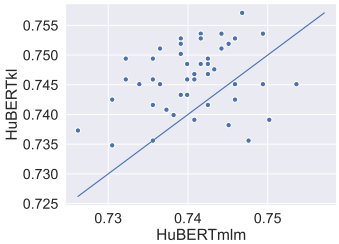
A különböző előtanítási megközelítések alkalmazása mellet létrehozott modellek egyes feladatokon a finomhangolási kísérletek során elért eredményességi mutatóinak eloszlását a 4. ábrán látható dobozdiagramok szemléltetik. Az ábra vizsgálatából kitűnik, hogy nem csak az átlagos teljesítmény tekintetében, hanem a medián teljesítmény vonatkozásában is eredményesebbek a HuBERTkl finomhangolásával nyert modellek.



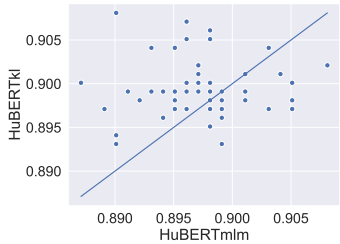
(a) HuCOLA



(b) HuCoPA

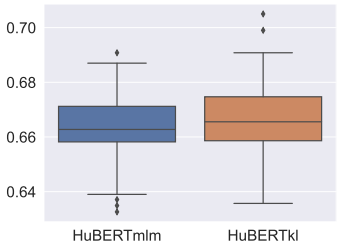


(c) HuSST

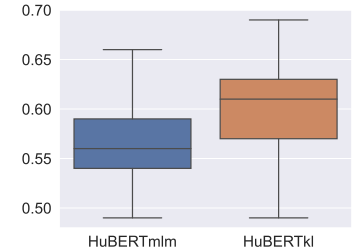


(d) OpinHuBank

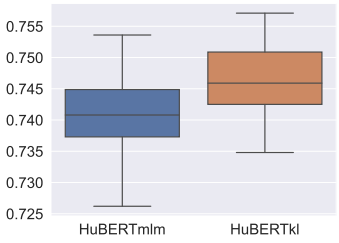
3. ábra: Az eltérő módon előtanított, azonosan inicializált osztályozó fejekkel rendelkező modellpárosok eredményei feladatonkénti lebontásban.



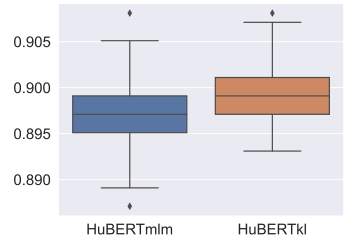
(a) HuCOLA



(b) HuCoPA



(c) HuSST



(d) OpinHuBank

4. ábra: A feladatonként és modellenkénti 50-50 kísérleti eredményeinek eloszlása.

5. Konklúzió

Cikkünkben egy olyan transzformer architektúrára tettünk javaslatot, ami az előtanítása során a konkrét szótöredékek rekonstruálása helyett (látens) szemantikus információk meghatározását tűzi ki célul. A javasolt módszert mind kvalitatív, mind pedig kvantitatív kiértékelésnek alávetettük, amelyek alapján úgy találtuk, hogy ígéretes alternatívát képes nyújtani a klasszikus szótöredék-alapú maszkolt nyelvi modellezéssel történő előtanításnak. Az elkészült modelljeinket[1,2], valamint a létrehozásukhoz használt forráskódot[3] egyaránt elérhetővé tettük.

Köszönetnyilvánítás

A kutatás az Európai Unió támogatásával valósult meg, az RRF-2.3.1-21-2022-00004 azonosítójú, Mesterséges Intelligencia Nemzeti Laboratórium projekt keretében. A MILAB SZTE nyelvtechnológia projekt nevében köszönetet mondunk az ELKH Cloud (lásd: Héder és mtsai (2022); https://science-cloud.hu/) használatáért, ami hozzájárult a publikált eredmények eléréséhez.

Hivatkozások

Bauer, L., Wang, Y., Bansal, M.: Commonsense for generative multi-hop question answering tasks. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 4220–4230. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018), https://aclanthology.org/D18-1454

Berend, G.: Sparsity makes sense: Word sense disambiguation using sparse contextualized word representations. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 8498–8508. Association for Computational Linguistics, Online (Nov 2020), https://aclanthology.org/2020.emnlp-main.683

Brown, P.F., Della Pietra, V.J., deSouza, P.V., Lai, J.C., Mercer, R.L.: Class-based n-gram models of natural language. Computational Linguistics 18(4), 467–480 (1992), https://aclanthology.org/J92-4003

Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), https://www.aclweb.org/anthology/N19-1423

[1] https://huggingface.co/SzegedAI/HuBERTmlm
[2] https://huggingface.co/SzegedAI/HuBERTkl
[3] https://github.com/szegedai/MLSM

Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., Smith, N.: Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping (2020), http://arxiv.org/abs/2002.06305, cite arxiv:2002.06305

Fellbaum, C.: WordNet: An Electronic Lexical Database. Bradford Books (1998)

Héder, M., Rigó, E., Medgyesi, D., Lovas, R., Tenczer, S., Török, F., Farkas, A., Emődi, M., Kadlecsik, J., Mező, Gy., Pintér, Á., Kacsuk, P.: The past, present and future of the ELKH cloud. Információs Társadalom 22(2), 128 (aug 2022), https://doi.org/10.22503/inftars.xxii.2022.2.8

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., Schwab, D.: Flaubert: Unsupervised language model pre-training for french. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 2479–2490. European Language Resources Association, Marseille, France (May 2020), https://www.aclweb.org/anthology/2020.lrec-1.302

Levine, Y., Lenz, B., Dagan, O., Ram, O., Padnos, D., Sharir, O., Shalev-Shwartz, S., Shashua, A., Shoham, Y.: SenseBERT: Driving some sense into BERT. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4656–4667. Association for Computational Linguistics, Online (Jul 2020), https://aclanthology.org/2020.acl-main.423

Ligeti-Nagy, N., Ferenczi, G., Héja, E., Jelencsik-Mátyus, K., Laki, L.J., Vadász, N., Yang, Z.Gy., Váradi, T.: HuLU: magyar nyelvű benchmark adatbázis kiépítése a neurális nyelvmodellek kiértékelése céljából. In: XVIII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 431–446. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2022)

Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H., Wang, P.: K-BERT: Enabling language representation with knowledge graph. Proceedings of the AAAI Conference on Artificial Intelligence 34(03), 2901–2908 (Apr 2020), https://ojs.aaai.org/index.php/AAAI/article/view/5681

Martin, L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., Sagot, B.: CamemBERT: a tasty French language model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7203–7219. Association for Computational Linguistics, Online (Jul 2020), https://aclanthology.org/2020.acl-main.645

Miháltz, M.: OpinHuBank: szabadon hozzáférhetö annotált korpusz magyar nyelvű véleményelemzéshez. In: IX. Magyar Számítógépes Nyelvészeti Konferencia. pp. 343–345. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2013), http://acta.bibl.u-szeged.hu/58859/1/msznykonf_009_343-345.pdf

Miháltz, M., Hatvani, C., Kuti, J., Szarvas, Gy., Csirik, J., Prószéky, G., Váradi, T.: Methods and results of the Hungarian wordnet project. In: Proceedings of The Fourth Global WordNet Conference. pp. 311–321 (2008)

Mihaylov, T., Frank, A.: Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 821–832. Association for Computational Linguistics, Melbourne, Australia (Jul 2018), https://aclanthology.org/P18-1076

Nemeskey, D.M.: Natural Language Processing Methods for Language Modeling. Ph.D.-értekezés, Eötvös Loránd University (2020), https://hlt.bme.hu/en/publ/nemeskey_2020

Nemeskey, D.M.: Introducing `huBERT`. In: XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2021). pp. 3–14. Szeged (2021)

Peters, M.E., Neumann, M., Logan, R., Schwartz, R., Joshi, V., Singh, S., Smith, N.A.: Knowledge enhanced contextual word representations. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 43–54. Association for Computational Linguistics, Hong Kong, China (Nov 2019), https://aclanthology.org/D19-1005

Qiu, D., Zhang, Y., Feng, X., Liao, X., Jiang, W., Lyu, Y., Liu, K., Zhao, J.: Machine reading comprehension using structural knowledge graph-aware network. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 5896–5901. Association for Computational Linguistics, Hong Kong, China (Nov 2019), https://aclanthology.org/D19-1602

Speer, R., Havasi, C.: Representing general relational knowledge in conceptnet 5. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012). European Language Resources Association (ELRA) (2012)

Ulčar, M., Robnik-Šikonja, M.: FinEst BERT and CroSloEngual BERT. In: Sojka, P., Kopeček, I., Pala, K., Horák, A. (szerk.) Text, Speech, and Dialogue. pp. 104–111. Springer International Publishing, Cham (2020)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (szerk.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017)

Vilares, D., García, M., Gómez-Rodríguez, C.: Bertinho: Galician BERT representations. Proces. del Leng. Natural 66, 13–26 (2021), http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6319

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020), https://aclanthology.org/2020.emnlp-demos.6

Yang, A., Wang, Q., Liu, J., Liu, K., Lyu, Y., Wu, H., She, Q., Li, S.: Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 2346–2357. Association for Computational Linguistics, Florence, Italy (Jul 2019), https://aclanthology.org/P19-1226

Ye, Z., Chen, Q., Wang, W., Ling, Z.: Align, mask and select: A simple method for incorporating commonsense knowledge into language representation models. CoRR abs/1908.06725 (2019), http://arxiv.org/abs/1908.06725