A látens szemantikus tulajdonságok segítségével előtanított nyelvi modellek vizsgálata

Berend Gábor

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

Kivonat A szóalakok helyett fogalmi egységek mentén előtanított nyelvi modellek mintahatékonyságát több korábbi munka is igazolta. Ezek a vizsgálatok azonban a létrehozott modelleket csupán finomhangolhatóságuk szempontjából vizsgálták. Jelen cikkben a látens szemantikus tulajdonságok jelentette kváziszimbolikus információ mentén történő előtanításnak a nyelvi modellezés elvégzésére gyakorolt (negatív) hatását mutatjuk be, illetve javaslunk két módszert is, amely használatával a modellek nyelvi modellezési képességei javíthatók a fogalmakra való érzékenységének megtartása mellett.

1. Bevezetés

A tradicionálisan előtanított maszkolt nyelvi modellek létrehozásuk során olyan kimeneti eloszlások meghatározásával tudják a veszteségfüggvényüket minimalizálni, amelyek a teljes valószínűségi tömeget pontosan egy megadott – az aktuális mondatban a kimaszkolást megelőzően az adott szó helyén eredetileg ott álló – szóra helyezik. Vagyis ha a maszkolás után előálló *Péternek [MASK] gyereke van.* mondatban az eredetileg a [MASK] token helyén álló szó a *három* számnév volt, úgy a nyelvi modell abban az esetben tudja minimalizálni az adott maszk tokenre vonatkozó veszteségét, ha annak behelyettesítésével kapcsolatosan kizárólag a *három* szó ottlétéhez rendel nullától különböző valószínűséget, holott a valóságban más szavak is alkalmasak lennének a mondatból kitakart szó helyettesítésére. A betanított maszkolt nyelvi modellek kimeneti eloszlásai természetesen már nem annyira szélsőségesek, mint amennyire szélsőséges kimeneti eloszlások minimalizálják a veszteségfüggvényüket, amit annak köszönhetnek, hogy az előtanításuk nagy mennyiségű, változatos szövegeken történik, és hosszú távon az egyes kimenetek mentén elvárt szélsőséges eloszlások „átlagát" adják vissza.

Shani és mtsai (2023) véleménycikkükben amellett érveltek, hogy a nyelvi modellek viselkedése szempontjából kívánatos lenne, ha azok működése a konkrét szóalakok helyett emberi fogalmak mentén történne. Egy ilyen megoldást adott (Berend, 2024), amely egy már tradicionálisan betanított nyelvi modell rejtett reprezentációi segítségével határozta meg azon látens szemantikai tulajdonságok körét, amely alapján aztán az enkóderalapú neurális nyelvi modellek tanításának elvégezését javasolta. A módosított előtanítással nyert neurális nyelvi modellek mintahatékonynak bizonyultak abban az értelemben, hogy megegyező mennyiségű szövegen történő előtanítást követően, változatos célfeladatok mentén való

finomhangolásuk során a hagyományos módon előtanított modelleknél átlagosan jóval kedvezőbb eredmények elérésére voltak képesek.

A Berend (2024) által létrehozott modellek mintahatékonysága azonban nem terjedt ki a létrejövő modellek maszkolt nyelvi modellezési képességeire, hiszen azáltal, hogy a javasolt előtanítás teljességgel függetlenítette magát a konkrét szóalakok előrejelzésétől, így ennek a feladatnak az elvégzésére teljesen alkalmatlanak voltak a modelljei.

A korábban javasolt látens szemantikus tulajdonságok mentén történő előtanítás egy további tulajdonsága, hogy egy már előtanításon átesett segéd nyelvi modellre támaszkodik. Azon túl, hogy egy már előtanított modell létezésének feltételezése egy új modell létrehozása során nem feltétlen elegáns, maga után vonja azt a negatív következményt is, hogy a látens szemantikus tulajdonságokkal operáló neurális nyelvi modell szótárának szükségképpen a segédmodell szótárával megegyezőnek kell legyen, ezzel egy potenciálisan komoly korlátot állítva a módszer gyakorlati alkalmazhatósága elé.

Jelen cikkben olyan kiterjesztését, illetve módosítását adjuk a maszkolt látens szemantikus modellezési (MLSM) előtanítási eljárásnak, amely segítségével létrehozott neurális nyelvi modellek a tokenek fogalmi csoportosításán túlmenően a konkrét maszkolt nyelvi modellezési feladat ellátására is képesek. A javasolt új módszerek közül az egyik még azzal a további kedvező tulajdonsággal is rendelkezik, hogy nem igényli egy már korábban betanított nyelvi modell meglétét, ezáltal lehetőséget teremt arra is, hogy akár teljesen új – a számunkra releváns új doménhez adaptált – tokenizálóval hozzunk létre neurális nyelvi modelleket mintahatékony módon.

2. Kapcsolódó munkák

A maszkolt nyelvi modellezés során sokáig a BERT (Devlin és mtsai, 2019) modell létrehozása során lefektetett stratégiára támaszkodtak, ami az inputban található szótöredékek véletlenszerűen kiválasztott 15%-a alapján hajtja végre az előtanítást. A maszkolt nyelvi modellezéssel előtanított nyelvi modellek feladata az, hogy az eredeti szövegek 15%-ának lecserélését követően legyen képes helyreállítani a szövegből kitakart szóalakok pontos kilétét. A közelmúltban azonban számos alternatívát javasoltak a rekonstruálandó szótöredékek véletlenszerű kiválasztása helyett (Joshi és mtsai, 2020; Levine és mtsai, 2021; Yang és mtsai, 2023). Wettig és mtsai (2023) az inputszekvencia 15%-ára irányuló kiválasztási stratégiát vizsgálta, és azt találták, hogy érdemes az inputban található tokeneknek a megszokottnál nagyobb hányadát rekonstruálandó tokenként kezelni.

Berend (2023) egy olyan előtanítási feladatot vezetett be, amely során a tanulás tárgyát nem a kiválasztott inputtokenek rekonstruálása képzi, hanem az azok kontextusa mentén meghatározott látens szemantikus jellemzések meghatározása a célja. A szemantikus jegyek meghatározására egy segédmodell rejtett reprezentációi alapján került sor ritka kódolás használatával. A módszer sajátossága, hogy mivel az előtanítás során alkalmazott modellezés fókuszát áthelyezi a szóalakokról a kváziszimbolikus szemantikus jegyekre, így a klasszikus nyel-

vi modellezés feladatát nem képesek ellátni az ezzel a módszerrel létrehozott modellek.

3. Vizsgált módszerek

Cikkünkben a klasszikus maszkolt nyelvi modellezéssel előtanított transzformeralapú enkóder modelleknél jobb mintahatékonysággal rendelkező maszkolt látens szemantikus modellezési előtanítási eljárás továbbfejlesztési lehetőségeit vizsgáljuk. Célunk, hogy a látens szemantikus kategóriák kezelésére, valamint a szó(be)helyettesítési feladat elvégzésére *egyaránt* képes modelleket hozzunk létre. A következőkben áttekintjük a kísérleteinkben előforduló előtanítási eljárásokat, illetve azoknak az általunk javasolt továbbfejlesztéseit.

3.1. Maszkolt nyelvi modellezés (MLM)

A klasszikus maszkolt nyelvi modellezési feladat során a kategorikus keresztentrópia hibatagot használjuk annak számszerűsítésére, hogy az előtanított modell aktuálisan mekkora outputvalószínűséget rendel a megadott pozíción az eredeti inputban szereplő token visszahelyettesítésével kapcsolatban. Ez a klasszikus előtanítási paradigma húzódik meg számos modell, pl. a BERT (Devlin és mtsai, 2019), vagy a magyar huBERT (Nemeskey, 2021) létrehozása mögött.

3.2. Maszkolt látens szemantikus modellezés (MLSM)

Az MLSM egy segédmodellre támaszkodik a maszkolásra véletlenszerűen kiválasztott tokenek kváziszimbolikus látens szemantikus profiljának meghatározása során.

Az eljárás első előkészítő lépése a segédmodell h-dimenziós rejtett állapotaiból mintavételezéssel létrehozott $\boldsymbol{X} \in \mathbf{R}^{h \times n}$ mátrixra támaszkodik, amelyre megkeressük a

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

feladat optimumát, és amely alapján egy korábbiakban nem látott $\boldsymbol{x}_i \notin X$ rejtett reprezentációval rendelkező vektorra a $\boldsymbol{D} \in \mathbf{R}^{h \times k}$ mátrixot már ismertnek tekintve az

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

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

A λ regularizációs tag azt eredményezi, hogy a $\boldsymbol{\alpha_i}$-beli együtthatók többsége pontosan 0 lesz, míg a nemnulla együtthatók pozíciói olyanok lesznek, hogy azok – a kísérleteink által is a későbbiekben alátámasztott módon – nagyfokú együttállást mutatnak a szemantikus tulajdonságaikkal.

UT-MLSM: Utólagos tanításon átesett maszkolt látens szemantikus modellezés A kváziszimbolikus látens szemantikus tulajdonságok segítségével előtanított modellek a modellezés hiperparamétereként megválasztott k darab látens szemantikus tulajdonságra vonatkozó eloszlásokat produkálnak. Amennyiben a látens tulajdonságok valóban a szavak fogalmi tulajdonságait ragadják meg, és az MLSM előtanítás megfelelően konvergált, úgy feltételezhető, hogy az MLSM célfüggvénnyel előtanított modellek által szolgáltatott eloszlásoknak alkalmasak a konkrétan kimaszkolt szavak rekonstruálására is.

Ezen hipotézis ellenőrzésének érdekében egy olyan modellvariánst hoztunk létre, amely egy hagyományos MLSM előtanításon átesett modellt olyan módon egészít ki, hogy annak paramétereinek lefagyasztását követően egy további osztályozó fejjel egészíti ki a modellt. Ennek az osztályozó fejnek a feladata az, hogy a látens szemantikus tulajdonságok teréből a modell által használt szótárára vonatkozó logit értéket hozzon létre, vagyis lényegében egyetlen, a modellezés során használt látens tulajdonságok számaszor a modellben megkülönböztetett tokenek számaszoros lineáris leképezést tanulunk a modellhez úgy, hogy mindeközben a korábbiakban az előtanításon átesett súlyait érintetlenül hagyjuk.

SM-MLSM: Segédmodellmentes maszkolt látens szemantikus modellezés A következőnek javasolt modellvariánsunk fontos tulajdonsága, hogy az nem igényli egy az előtanítás megkezdésének pillanatában már rendelkezésre álló betanított neurális nyelvi modell létezését. A segédmodellmentes maszkolt látens szemantikus modellezés (SM-MLSM) elvén működő módszerünk a maszkolt nyelvi modellezés és a látens szemantikus kategóriák kialakításának alternáló végrehajtásának az elvét követi.

Az SM-MLSM módszer egy teljesen véletlenszerű súlyokkal inicializált neurális nyelvi modellből indul ki, majd az előtanítás során elvégzendő lépések számának az első meghatározott (esetünkben $12,5\%$-nak választott) lépésében kizárólagosan a klasszikus maszkolt nyelvi modellezés feladatát látja el. Az előtanítás a hiperparaméterben meghatározott darabszámú kisebb előtanítási szakaszból áll, és minden egyes szakasz végén, az adott szakaszban meghatározott utolsó 200 kötegben előfordult tokeneknek a transzformer modell utolsó rétegében előálló rejtett vektorai alapján egy (1)-ben található alakú optimalizálás problémát old meg annak érdekében, hogy a következő előtanítási szakaszban – a klasszikus maszkolt nyelvi modellezési feladat ellátása mellett párhuzamosan – a látens szemantikus tulajdonságok modellezésére vonatkozó előtanítási feladatot is el tudja látni a modell a frissített szótármátrixra támaszkodva.

Mivel a modell korai verziói még vélhetőleg kevésbé hasznos rejtett állapotokat produkálnak, ezért a maszkolt nyelvi modellezés és a látens szemantikus tulajdonságokra támaszkodó modellezésből jövő veszteség egymáshoz való viszonyát az előtanítás előre haladtával szakaszosan változtatjuk. A legelső szakaszban kizárólagosan csak a maszkolt nyelvi modellezésből jövő veszteségfüggvényt vesszük figyelembe a modell frissítése során, míg a következő szakaszokban ezen komponens azonos ütemben csökkenve egyre kisebb szerephez jut, az előtanítás végére elérve egy előzetesen meghatározott arányt (kísérleteink során $0,25$-et).

4. Kísérletek

A következőkben bemutatjuk az előtanítással kapcsolatos kísérleti beállításainkat, majd a modellekkel végzett kísérleteinket közöljük.

4.1. Az előtanítás

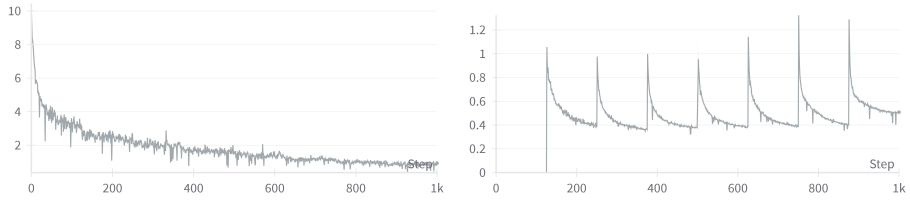
Modelljeink előtanítása során az input szövegek maximális szekvenciahosszát 128 szubtokenben maximalizáltuk, 16-os gradiens akkumuláció mellett 64-es kötegméretet, valamint 1e–4 maximális tanulási rátát alkalmaztunk. A maszkolásra kiválasztott tokenek arányát megemeltük 0,15-ről 0,25-re, összhangban (Wettig és mtsai, 2023) ajánlásaival. Minden elkészült modellünknek elkészítettük az előtanítási lépéseinek 10%-a, 25%-a, 50%-a, valamint 100%-a után előálló változatát. Az előtanítást miden esetben a huBERT (Nemeskey, 2021) tokenizálójára és a Webcorpus2.0(Nemeskey, 2020) szövegeire támaszkodva végeztük el. A korpusz hozzávetőlegesen tizenhatmilliárd tokenből áll.

UT-MLSM előtanítások Saját UT-MLSM modellünk[1] létrehozása során (Berend, 2024) publikusan elérhetővé tett, MLSM módszerrel előtanított DeBERTa architektúrájú (He és mtsai, 2021) modelljéből[2] indultunk ki. Mivel az UT-MLSM modell esetén a kiinduló modell súlyai lefagyasztásra kerültek, és éppen arra voltunk kíváncsiak, hogy az eredeti modell által visszaadott látens szemantikus tulajdonságokra vonatkozó tokenreprezentációk mennyire alkalmasak a maszkolásos nyelvi modellezési feladat hatékony megtanulására, ebben az esetben mindössze 10000 előtanítási lépést hajtottunk végre, ami a Webcorpus2.0 körülbelül 7,5%-ának felel meg.

SM-MLSM és tradicionális MLM előtanítások Mivel ezek a modellek nem egy már korábbi előtanított modell továbbtanításiból jöttek létre, itt az előtanításra fordított frissítési lépéseink számát százezerre állítottuk. Mindez hozzávetőlegesen tizenkétmilliárd token (a Webcorpus2.0 hozzávetőlegesen háromnegyedének) a fölhasználását jelentette.

A tradicionális MLM modell[3] előtanításra vonatkozó hiperparamétereink nem tértek el a korábban megadottaktól, míg a segédmodellmentes MLSM modell[4] számára specifikus hiperparaméterek a következők szerint alakultak: a szótármátrix frissítését 12500 lépésenként (a teljes előtanítás ideje alatt tehát összesen 8 alkalommal) végeztük el, az alkalmazott látens kategóriák számát a modell rejtett vektorainak dimenziószámának kétszeresére (1536-nak) választottuk, és a maszkolásos nyelvi modellezés veszteségfüggvényének relatív súlya az előtanítás végére a 0,25-ös értéket vette fel. Az SM-MLSM előtanítás tulajdonképpen egy többfeladatos tanulási problémaként értelmezhető. Az aggregált veszteségfüggvényünk egyes komponenseinek alakulását az 1. ábrában mutatjuk be.

[1] https://huggingface.co/SzegedAI/huDeBERTa-MLSM-UT
[2] https://huggingface.co/SzegedAI/huDeBERTa-MLSM
[3] https://huggingface.co/SzegedAI/huDeBERTa-MLM-v2_100k
[4] https://huggingface.co/SzegedAI/huDeBERTa-aux-free-MLSM

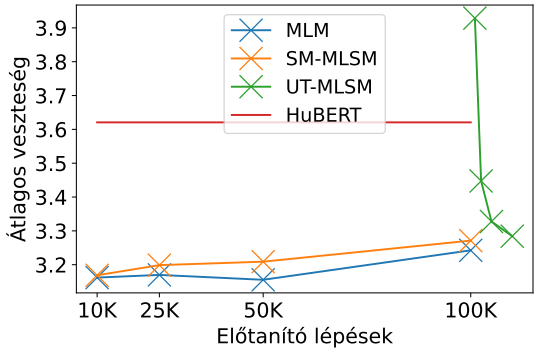


(a) Az MLM veszteségfüggvény alakulása (b) Az MLSM veszteségfüggvény alakulása

1. ábra: A segédmodell fölhasználása nélkül létrehozott SM-MLSM modell veszteségének komponenseinek alakulása a tanítás előrehaladtának függvényében: (a) a klasszikus MLM tanítás által (is) használt kategorikus keresztentrópia hibatag, valamint a (b) MLSM által használt látens szemantikus tulajdonságok céleloszlásához viszonyított Kullback-Leibler divergencia célfüggvényértékének változása.

4.2. A modellek maszkolt nyelvi modellezési képességei

Mint azt korábban említettük, a kváziszimbolikus alapokon nyugvó eredeti MLSM modellezéssel előtanított modellek egyáltalán nem alkalmasak a maszkolt nyelvi modellezési feladat ellátására, vagyis a kimaszkolt tokenek helyére nem képesek a token helyén álló valódi tokenekre vonatkozó kimeneti eloszlásokat produkálni. Az UT-MLSM és SM-MLSM módszerekkel előtanított modelljeink különböző mechanizmusok útján ezek képességgel kapcsolatos hiányosságaira adnak egy megoldást a látens szemantikus tulajdonságok segítségével előtanított modellek viselkedésével kapcsolatban. A modellek tanítására nem használt fejlesztési korpusz mentén a különböző módon előtanított modellek átlagos veszteségfüggvényértékét közöljük a 2. ábrán.

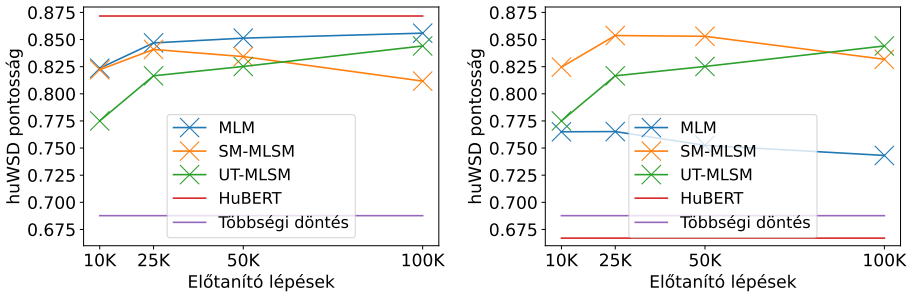


2. ábra: A különböző előtanított modelleknek a tanításukra föl nem használt szövegeken elért veszteségfüggvényének átlagos értéke a kumulált előtanító lépések számának függvényében. Az UT-MLSM modell esetén egy már előzetesen 100 ezer előtanítási lépésen átesett modellből indulunk ki.

Az 1. ábra alapján megállapíthatjuk, hogy az előtanítás során a tokenek látens fogalmi tulajdonságait is figyelembe vevő modelljeink maszkolt nyelvi modellezési képességei összehasonlíthatók a kizárólag a maszkolásos nyelvi modellezési feladat ellátását elvégezni képes tradicionális MLM módon előtanított modellével. Az UT-MLSM modellel kapcsolatban az a tény, hogy már a teljes Webcorpus2.0 kevesebb, mint 1%-ának a fölhasználásával nemtriviális nyelvi modellezési képességekre tett szert (4.0 alatti veszteségfüggvény-érték) azon hipotézisünket támasztja alá, hogy a huDeBERTa-MLSM modell által előzetesen megalkotott látens fogalmi hierarchiája hasznos kiindulási alapot tudott biztosítani a maszkolt nyelvi modellezési képességek elsajátítására nézve.

4.3. Jelentésegyértelműsítési eredmények

Megvizsgáltuk a különböző modellek által a huWSD jelentésegyértelműsítési feladaton elért eredményeket. A nyelvi modellek által létrehozott reprezentációk jelentésegyértelműsítési feladatba történő felhasználásának egy egyszerű, de meglepően jó teljesítményt produkálni képes változata az, amikor az egyes ismert jelentéskategóriákhoz egy-egy centroidot rendelünk az adott kategóriába tartozó tanítóhalmazbeli szavakhoz a nyelvi modell által rendelt rejtett reprezentációkból, és tesztelés során egy egyszerű legközelebbi szomszédsági elven döntünk (Loureiro és Jorge, 2019). Az ezzel a módszerrel kapott eredményeinket a különböző módszerekkel előtanított nyelvi modelljeink eltérő elkészültségi fokai mellett a 3a. ábra tartalmazza. Látható, hogy mindegyik modell kifejezetten magas pontosság elérésére képes.



(a) Utolsó rejtett réteg vektora (b) Modell kimeneti eloszlása

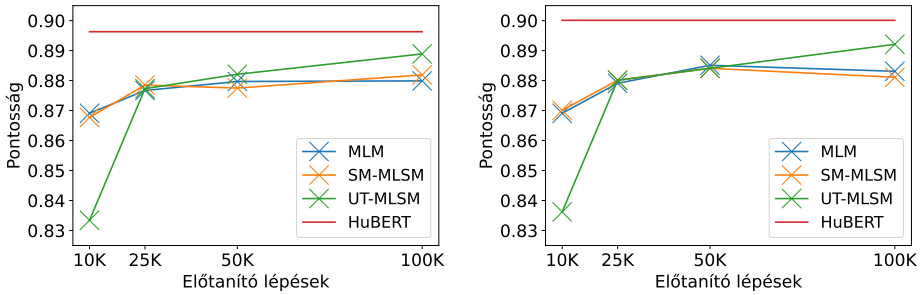
3. ábra: A különböző vizsgált modelleknek a huWSD adatbázisán elért jelentésegyértelműsítési eredményei: (a) a modell utolsó rétegéből jövő rejtett állapotok, illetve (b) a modell kimeneti eloszlásának fölhasználásával elért eredmények.

Egy kevésbé gyakori megközelítés, ugyanakkor a modellezést a kváziszimbolikus látens szemantikus fogalmi tulajdonságok mentén (is) elvégző előtanítási

módszerek esetén azonban kínálja magát az a megoldás is, hogy a modell belső rejtett reprezentációi helyett konkrétan magára a modell kimenetén produkált eloszlásvektorokra hagyatkozva hozzuk meg a döntéseinket. Ebben az esetben a legközelebbi szomszédság megítélését az Hellinger-távolság kiszámolásával végeztük. Az így kapott eredményeinket a 3b. ábra foglalja össze. Az látható, hogy a látens tulajdonságokat alkalmazó modellek ebben a fajta kiértékelésben sokkal jobb teljesítmény elérésére voltak képesek a látens szemantikus tulajdonságokat az előtanítás során figyelembe nem vevő társaikhoz képest.

4.4. Finomhangolásos eredmények

Noha kutatásunkban a modelleket elsődlegesen nem finomhangolhatóságuk szempontjából vizsgáltuk, a következőkben közöljük az opinhubank (Miháltz, 2013) szentimentosztályozási feladatán elért eredményüket. A modelleket 5 alkalommal finomhangoltuk, és közöljük az egyes modellek mentén kapott pontosság átlagát (4a), valamint a független finomhangolással létrehozott modellek szavaztatásával kapott eredményeinket (4b) a különböző módszerrel előtanított neurális nyelvi modelljeinkre. A kapott finomhangolási eredményeinkről elmondható, hogy azok mindegyike összehasonlítható, vagyis a javasolt módosításokkal is a finomhangolásra nézve jól teljesítő nyelvi modelleket kaphatunk.



(a) Egyedi modellek eredménye (b) Modellek szavaztatásának eredménye

4. ábra: A különböző vizsgált modelleknek az opinhubank szentimentosztályozási adathalmazon elért eredményei.

5. Konklúzió

Cikkünkben a látens szemantikus tulajdonságokra támaszkodva megalkotott neurális nyelvi modellek azon hiányosságaira adtunk megoldásokat, miszerint azok nem alkalmasak a klasszikus maszkolásos nyelvi modellezési feladat ellátására. A javasolt módszereink egyike (SM-MLSM) ráadásul nem csak ezen limitációján segít a látens kategóriák mentén történő nyelvi modellezésnek, hanem azt

is lehetővé teszi, hogy a nyelvi modell előtanításának megkezdésekor ne kelljen rendelkezésünkre állnia egy betanított már átesett segéd nyelvi modellnek.

Köszönetnyilvánítás

A cikk a Bolyai János Kutatási Ösztöndíj támogatásával készült. 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

Berend, G.: Masked latent semantic modeling: an efficient pre-training alternative to masked language modeling. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (szerk.) Findings of the Association for Computational Linguistics: ACL 2023. pp. 13949–13962. Association for Computational Linguistics, Toronto, Canada (Jul 2023), https://aclanthology.org/2023.findings-acl.876

Berend, G.: Neurális nyelvi modellek látens szemantikus információ alapján történő maszkolásmentes előtanítása. In: XX. Magyar Számítógépes Nyelvészeti Konferencia. pp. 85–95. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2024), https://acta.bibl.u-szeged.hu/78418/

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

He, P., Liu, X., Gao, J., Chen, W.: DeBERTa: Decoding-enhanced BERT with disentangled attention. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=XPZIaotutsD

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

Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L., Levy, O.: SpanBERT: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics 8, 64–77 (2020), https://aclanthology.org/2020.tacl-1.5

Levine, Y., Lenz, B., Lieber, O., Abend, O., Leyton-Brown, K., Tennenholtz, M., Shoham, Y.: PMI-masking: Principled masking of correlated spans. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=3Aoft6NWFej

Loureiro, D., Jorge, A.: Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation. In: Korhonen, A., Traum, D., Màrquez, L. (szerk.) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 5682–5691. Association for Computational Linguistics, Florence, Italy (Jul 2019), https://aclanthology.org/P19-1569

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

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)

Shani, C., Vreeken, J., Shahaf, D.: Towards concept-aware large language models. In: Bouamor, H., Pino, J., Bali, K. (szerk.) Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 13158–13170. Association for Computational Linguistics, Singapore (Dec 2023), https://aclanthology.org/2023.findings-emnlp.877

Wettig, A., Gao, T., Zhong, Z., Chen, D.: Should you mask 15% in masked language modeling? In: Vlachos, A., Augenstein, I. (szerk.) Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. pp. 2985–3000. Association for Computational Linguistics, Dubrovnik, Croatia (May 2023), https://aclanthology.org/2023.eacl-main.217

Yang, D., Zhang, Z., Zhao, H.: Learning better masking for better language model pre-training. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (szerk.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 7255–7267. Association for Computational Linguistics, Toronto, Canada (Jul 2023), https://aclanthology.org/2023.acl-long.400