

Neurális nyelvi modellek (tovább)fejlesztése

Berend Gábor, Szegedi Tudományegyetem
II. Országos HUN-REN Cloud Találkozó

2025. 11. 14.



Bemutakozás

- Az SZTE-n folytatunk a (nagy) nyelvi modellek működését interpretálhatóbbá (és ezáltal) hatékonyabbá és robusztusabbá tevő eljárásokat
- Csoportunk fő profilja korábban (a 2010-es évek) a magyar nyelvű nyelvi elemzők létrehozása és azok alkalmazásokba ültetése volt



Problémafelvetés

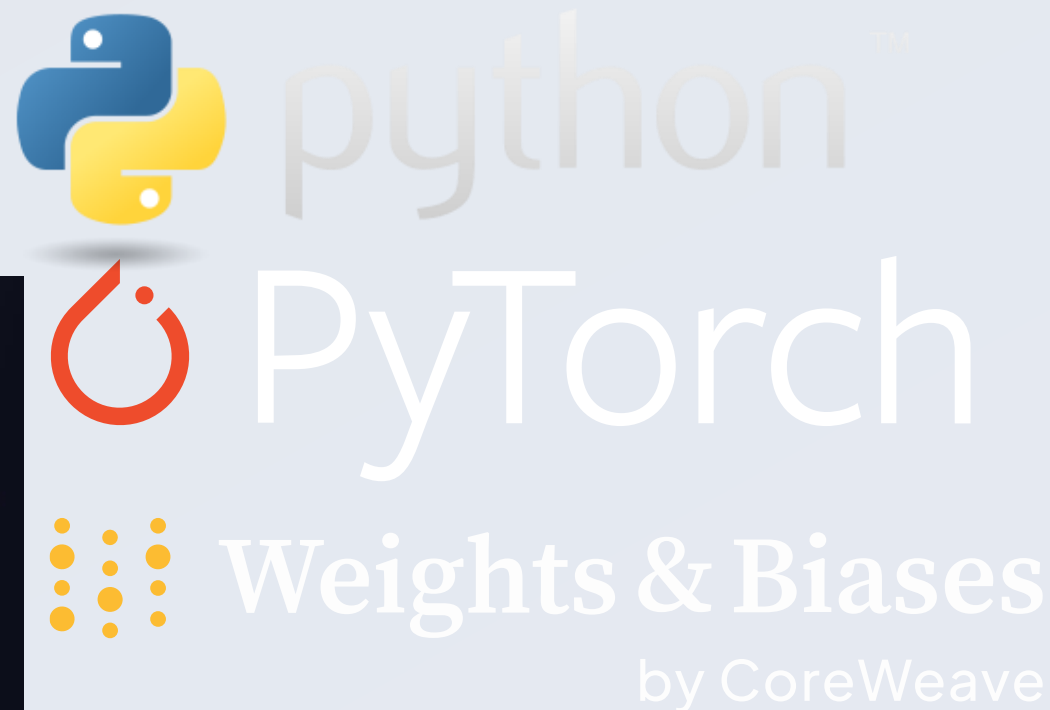
- A neurális nyelvi modellek teljesítménye számos tekintetben kimagasló
- Ugyanakkor több korlátozó tényező (pl. adat,-és erőforráséhség, konzisztens viselkedés hiánya, ...) is jellemzi őket
 - Célunk ezen hiányosságok orvoslása/csökkentése

Igényelt erőforrások

- A projekt 2021-ben indult, az idei évtől ernyőprojektként fut
- Az igényelt erőforrások az idő folyamán (némileg) átalakultak
- A jelenleg használt instance-ek:
 - 3 db g2.2xlarge (Nvidia V100GPU) és 2 db g2.xlarge (1/2 Nvidia V100 GPU)
 - A GPU-kat a (jellemzően enkóder-alapú) nyelvi modellek előtanítására, finomhangolására, illetve a velük történő hatékony inferenciára használjuk

Felhős környezet

- Ernyőprojekt lévén a különböző projektek igényeit különböző szoftveres környezetek elégítik ki



Felhős környezet

- A jobbaink jellege is széles skálán mozog
 - Nagy számú (ezres mennyiségben elvégzett), relatív rövid (legfeljebb pár órás) batch jobok → finomhangolási és vagy modellkiértékelési feladatok
 - Heteken (vagy akár >1 hónapon) át tartó előtanítási feladatok



Tárolási megoldások

- Minden instance külön életet él, saját háttértár + csatolt kötetek
 - Pazarló, mert ugyanazok a modellek minden instance-en megtalálhatók
- A tanítóhalmazokat és a modelleket a HuggingFace Hub-ról/ra töltjük le/föl
- Az egyéb adatmozgatásokra scp/rsync parancsokkal történik a Cloud instancekre, illetve azok között
 - A külső hálózatról a Cloudba történő másolás lassú tud lenni (soktíz, vagy akár száz Gb-nyi adat másolására kerül sor)
- Eredmények naplózása lokálisan, illetve W&B segítségével



Kutatási eredmények

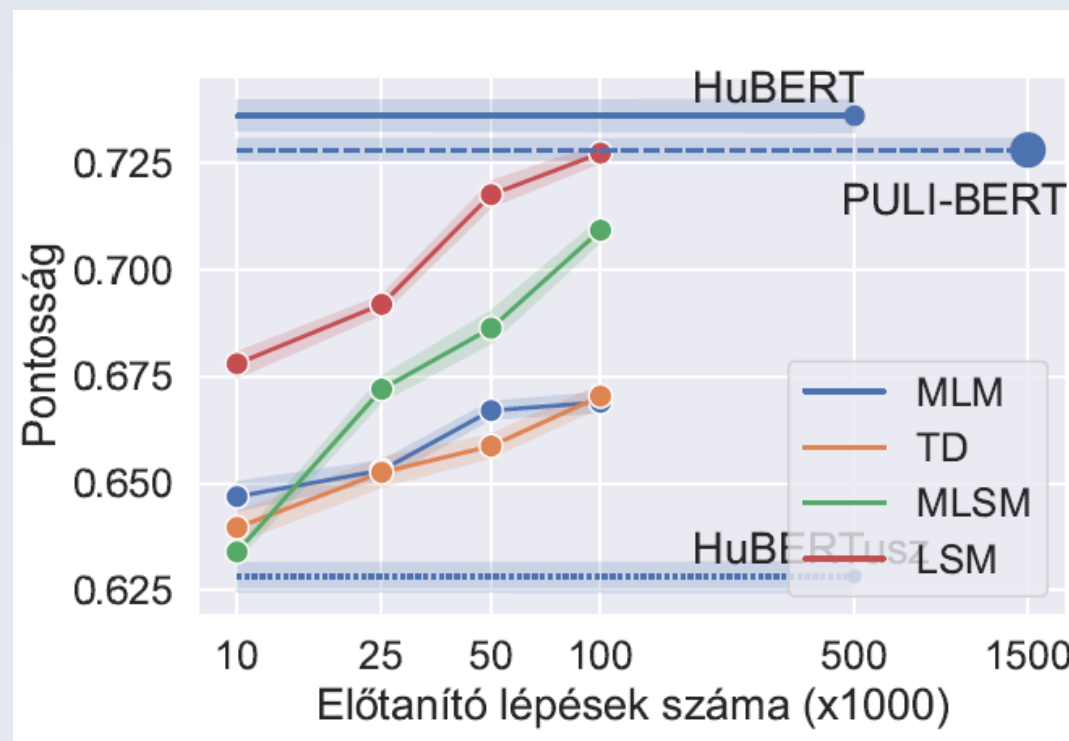
- A maszkolósos nyelvi modellezés egy némileg természetellenes feltevésre alapszik

Kutatási eredmények

- A maszkolósos nyelvi modellezés egy (némileg) természetellenes feltevésre alapszik
 - Ti. egy adott szó helyettesítésére egy adott szövegkörnyezetben kizárólag önmaga alkalmas
 - Praktikus, ugyanakkor túl „merev”
 - Ötlet: az előtanítás során ne a konkrét szóalak rekonstruálását tegyük az előtanítás feladatául, hanem a kitakart szó (szemantikus) tulajdonságaiét

Kutatási eredmények

- A maszkolásos nyelvi modellezés egy (némileg) természetellenes feltevésre alapszik

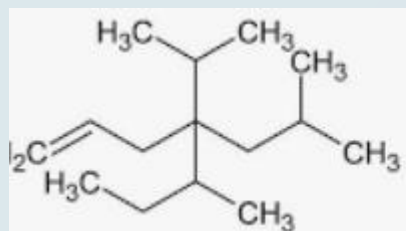


Publikációk

- Összesen 11 darab publikáció (8 nemzetközi)
- Kiemelt publikációk:
 - SUE: Sparsity-based Uncertainty Estimation via Sparse Dictionary Learning. EMNLP 2025.
 - Masked Latent Semantic Modelling: an Efficient Pre-Training Alternative to Masked Language Modeling. Findings of ACL 2023.
 - Combining the Curse of Multilinguality in Cross-Lingual WSD by Aligning Sparse Contextualized Word Representations. NAACL 2022.
 - Better Together: Jointly Using Masked Latent Semantic Modeling and Masked Language Modeling for Sample Efficient Pre-training. BabyLM 2023.
 - Látens szemantikus eloszlások használata a nyelvi modellek előtanítása során. MSZNY 2023.



Jövőbeli tervek: molekulák modellezése



5-methyl-4-(2-methylpropyl)-4-(propan-2-yl)hept-1-ene

CC(C)CC(CC=C)(C(C)C)C(C)CC



 MoleculeNet

text string · lengths	formula string · lengths	NumHDonors int64	NumHAcceptors int64	MolLogP float64	NumHeteroatoms int64	RingCount int64
1=205 100%	10=13 75.3%	0=13 100%	0=27 100%	-14.58=13.3 100%	0=38 100%	0=8 100%
CN1C(=O)C(=Cc2occc2C(=O)c2ccc(-c3ncc(C4=CC=CC5N=C(F)C(F)=NC45)cn3)s2)C(=O)N(C)C1=O	C28H18F2N6O5S	0	10	3.9029	14	6
CN1C(=O)C(=CC2=CC=C(c3cnc(-c4ccc(C(=O)c5ccoc5)s4)nc3)C3N=C(F)C(F)=NC23)C(=O)N(C)C1=O	C28H18F2N6O5S	0	10	3.8161	14	6
CN1C(=O)C(=Cc2ncc(-c3ccc(C(=O)c4cocc4C=Cc4ccnnc4)s3)nn2)C(=O)N(C)C1=O	C25H17N7O5S	0	11	2.8183	13	5
CN1C(=O)C(=Cc2nccc(C=Cc3cocc3C(=O)c3ccc(-c4cncnnc4)s3)n2)C(=O)N(C)C1=O	C25H17N7O5S	0	11	2.8183	13	5
CN1C(=O)C(=Cc2sc(C#N)c(-c3ncc(-c4ccsc4C(=O)c4ccoc4)cn3)c2C#N)C(=O)N(C)C1=O	C26H14N6O5S2	0	11	3.93476	13	5
CN1C(=O)C(=Cc2cocc2C(=O)c2sccc(-c2cnc(-c3c(C#N)csc3C#N)nc2)C(=O)N(C)C1=O	C26H14N6O5S2	0	11	3.93476	13	5
CN1C(=O)C(=Cc2c(F)cc(-c3sc4ccc(C#N)cc4c3C=Cc3ccnnc3)c3nsnc23)C(=O)N(C)C1=O	C28H16FN7O3S2	0	10	4.97798	13	6



Jövőbeli felhős tervek és igények

- A bf32 számábrázolást (és a strukturált ritkaság kezelését) natívan az Ampere mikroarchitektúrával kezdődő GPU-k támogatják, így ilyen (vagy újabb) GPU-k elérhetővé tétele segítené a futtatások hatékonyabbá tételében
- Instance-ek között megosztott hálózati tárolás (az adatbázisok és modellek redundás tárolásának elkerülése érdekében)
- Elosztott tanítás több GPU mentén



A felhővel kapcsolatos tapasztalatok

- Egyszerű használat
- Lassú kommunikáció a külső hálózat és a Cloud között (bár ennek okait a saját tűzfalunk konfigurációja okozza)
- Gyors és hatékony szupport (pl. licenceszerverrel kapcsolatos probléma)

Kutatási területek

Hova tovább?

- A szemantikus előtanítás kiterjesztése generatív modellekre
- A maszkolással kapcsolatos módosított előtanítás vizsgálata molekulák (és tulajdonságaik) modellezése során
- Gépi szövegek felismerésének felügyelet nélküli megvalósítása
- Nyelvi modellek döntéseinek bizonytalanságának detektálása



Köszönöm a figyelmet!

Bármilyen kérdés?

