Nyelvspecifikus transzformer modellek közötti megfeleltetéssel történő zero-shot jelentésegyértelműsítés

Berend Gábor

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

Kivonat Cikkünkben egy nyelvspecifikus transzformer modellekre támaszkodó, a jelentésegyértelműsítési feladatot zero-shot módon elvégző eljárást mutatunk be. A javasolt módszer a nyelvközi tudástranszfert a tanítóadatokkal rendelkező forrás-, valamint a tanítóadatokat nélkülöző célnyelv feldolgozására dedikáltan létrehozott egynyelvű előtanított modellekre épít. A nyelvek közötti kapcsolatot az egynyelvű transzformer modellek rejtett rétegei közötti megfeleltetést szolgáló leképezés tanulásával érjük el. Eredményeink megmutatják, hogy az ilyen módon létrehozott, kizárólag angol nyelvű jelentésegyértelműsített szövegeken tanuló modellek hatékonysága szignifikánsan javítható a többnyelvű maszkolt nyelvi modell alkalmazásához képest.
Kulcsszavak: jelentésegyértelműsítés; zero-shot tanulás

1. Bevezetés

A jelentésegyértelműsítés a természetesnyelv-feldolgozás egy régóta ismert, központi jelentőséggel bíró problémája (Weaver, 1949/1955; Lesk, 1986; Gale és mtsai, 1992; Navigli, 2009). A megoldási kísérletek között találkozhatunk tudásbázisokra támaszkodó, valamint felügyelt tanulást alkalmazó módszerekkel is, amelyek közül tipikusan az utóbbiak teljesítenek jobban. Mindkét fő megközelítésben közös, hogy komoly humánerőforrás-igénnyel rendelkeznek, hiszen mind a tudásbázisok, mind pedig a jelentésegyértelműsítésen átesett tanító-, illetve kiértékelőszövegek létrehozása igen költséges folyamat. Noha angol nyelvre több viszonylag nagy (azonban a jelentések sokszínűségéből adódóan a kívánatostól még így is elmaradó) jelentésegyértelműsítési adatbázis is létezik (Miller és mtsai, 1994; Taghipour és Ng, 2015), a legtöbb nyelvre – köztük a magyarra is – korábban nem létezett megfelelő méretű és részletezettségű adatbázis.[1]

A nemrégiben közreadott XL-WSD adatbázis (Pasini és mtsai, 2021) ezen a helyzeten változtat, ezért is éreztük szükségesnek a különféle kurrens módszereket egymással komplexen összehasonlító jelentésegyértelműsítési kiértékelési kísérletsorozat elvégzését, és bemutatását.

[1] Vincze és mtsai (2008) közreadott ugyan egy jelentésegyértelműsített korpuszt, azonban abban mindössze 39 többértelmű szóalak különböző jelentéseinek előfordulásai voltak megtalálhatók.

2. Kapcsolódó munkák

A kontextualizált szóreprezentációk jelentésegyértelműsítésben való fölhasználhatóságára első ízben (Peters és mtsai, 2018) mutatott rá. Loureiro és Jorge (2019) kísérletei azt igazolták, hogy a BERT (Devlin és mtsai, 2019) maszkolt nyelvi modellből kinyerhető vektorok segítségével egy egyszerű, mégis nagy hatékonyságú 1-legközelebbi szomszédságon alapú osztályozó építhető.

Berend (2020a) egy olyan felügyelet nélküli megoldásra tett javaslatot, ami a kontextuális jelentésvektorokat olyan módon alakítja át, hogy azok a bennük található együtthatók többségében nulla értéket vegyenek föl, az azonos koordináták mentén nemnulla együtthatóval rendelkező szavak pedig tendenciózusan azonos jelentéssel rendelkezzenek. A javasolt módszer segítségével létrehozott nagyfokú ritkasággal jellemzett jelentésvektorokat aztán a Loureiro és Jorge (2019) által is alkalmazott 1-legközelebbi módszerrel kombinálva szignifikáns javulások voltak elérhetők.

Az eddigiekben bemutatott munkák mindegyike az angol nyelven történő jelentésegyértelműsítésre fókuszált, aminek hátterében az áll, hogy magyarra nem létezett korábban kellő méretű és diverzitású jelentésegyértelműsítésre létrehozott tanító-, illetve tesztelő adatbázis. Érdemes megemlíteni a Vincze és mtsai (2008) által megalkotott magyar nyelvű újsághíreket tartalmazó jelentésegyértelműsített HuWSD adathalmazt, azonban ez az erőforrás mindössze 39 többértelmű szóalak vonatkozásában tartalmaz annotációkat, így igazán reprezentatív kiértékelést ez az adatbázis nem tesz lehetővé. A adatbázis méreteiből fakadó limitációi ellenére is születtek többnyelvű transzformer architektúrákat alkalmazó eredmények a HuWSD vonatkozásában is (Berend, 2020b, 2021).

A Pasini és mtsai (2021) által megalkotott XL-WSD adatbázis az angolon kívül 17 további nyelven tartalmaz jelentésegyértelműsítésen átesett diverz szövegeket, amelyek kísérleteink alapjául is szolgáltak.

3. Módszertan

Vizsgálataink során a transzformer modellekből kinyerhető, módosítatlan kontextuális reprezentációkat használó 1-legközelebbi szomszédságon alapuló módszert (Loureiro és Jorge, 2019), valamint a ritkításon átesett kontextuális reprezentációk (Berend, 2020a) használatát hasonlítjuk össze különféle esetekben.

3.1. Kontextuális modellek ritkítása

A kontextuális reprezentációk ritkítása során a Berend (2020a) által leírtak szerint jártunk el, azaz egy transzformer modell valamely rétegéből jövő d dimenziós rejtett reprezentációkat egy $Y \in \mathbb{R}^{d \times n}$ mátrixban összegyűjtve, a

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

feladatot oldottuk meg, ahol $\mathcal{C}$ a legfeljebb 1 normájú oszlopvektorok alkotta $d \times k$ méretű mátrixok konvex halmazát jelöli, λ a csupa nemnegatív értékből álló α együtthatómátrix ritkaságát befolyásoló regularizációs együttható, k pedig a mátrixdekompozíció során alkalmazott atomok számára vonatkozó hiperparaméter.

A ritka kontextuális reprezentációkat a Berend (2020a) által javasolt módon használtuk föl, azaz minden lehetséges s_i jelentéshez társítottunk egy $\phi_{s_i} \in \mathbb{R}^k$ vektort, amely vektor a D szótármátrixban található k jelentéskomponens és a jelentésegyértelműsített tanítókorpuszban s_i jelentésűként megjelölt szavak közötti kapcsolat erősségét fejezi ki a pontonkénti kölcsönös információ (PMI) segítségével. A tesztelés során egy $\alpha_j \in \mathbb{R}^k$ ritkításon átesett kontextuális reprezentációval rendelkező szó kapcsán a modellünk azt az s^* jelentést választja ki az adott szóhoz, amelyre $s^* = \max_{s \in S} s^\intercal \alpha_j$, ahol S az adott szó lehetséges jelentéseinek halmazát jelöli.

3.2. Nyelvspecifikus transzformerek közötti leképezés

Kísérleteink során zero-shot tanulást alkalmaztunk, azaz úgy értékeltük ki a modelljeink jelentésegyértelműsítésben nyújtott teljesítményét, hogy a létrehozásuk során egyáltalán nem támaszkodtunk magyar nyelvű jelentésegyértelműsített szövegekre. Ez komoly előnyt jelent, hiszen a kellően nagy és jó minőségű tanítóadatbázis létrehozása nagyon költséges lenne. Nem véletlen, hogy az általunk használt XL-WSD adatbázisban is csupán a validációs- és teszthalmaz mondatai tekintendők valódi etalonként, a tanítóhalmaz mondatait gépi fordítás segítségével hozták létre a szerzők. Pasini és mtsai (2021) megmutatták, hogy a zero-shot módon, azaz csupán az angol jelentésegyértelműsített tanítókorpusz alapján, illetve az egyidejűleg több eltérő nyelv támogatására képes transzformer modellre (pl. mBERT, XLM-RoBERTa) támaszkodó modelljeik jobb eredmény elérésére voltak képesek, mint a nyelvspecifikus – ám a gépi fordításból adódóan jóval zajosabb – adatokon tanított alternatív modelljeik.

Az általunk vizsgált zero-shot modellek túlmutatnak a korábbiakban létrehozottaktól, ugyanis a nyelvek közötti tudástranszfert nem többnyelvű enkóderek segítségével kívánjuk kezelni, hanem a nyelvspecifikus modellek különböző rétegei mentén kialakuló rejtett reprezentációk közötti lineáris transzformáció alkalmazásával. A javasolt módszer előnye, hogy ezáltal lehetőségünk van kiaknázni a forrás-,valamint a célnyelv feldolgozására specifikusan létrehozott transzformer modellek előnyeit, így elkerülhetővé válik a többnyelvű modellekre jellemző ún. *többnyelvűségi átok* (Conneau és mtsai, 2020).

Amennyiben a tesztelés során egy célnyelvi mondat valamely szavához társuló kontextuális reprezentáció $\boldsymbol{x}$, a célnyelvből a forrásnyelvbe vivő transzformáció pedig W által adott, úgy a forrásnyelvi szó ritka reprezentációját a

$$\min_{\alpha \in \mathbb{R}^k_{\geq 0}} \frac{1}{2} \|W\boldsymbol{x} - D\alpha\|_F^2 + \lambda \|\alpha\|_1, \tag{2}$$

szerint hoztuk létre. Érdemes észrevenni, hogy (1)-el szemben, (2) esetén az optimalizálás már csupán α-ban történik, ami lehetővé teszi a ritka reprezentációk

hatékony meghatározását. Kísérleteinkben az RCSLS algoritmust (Joulin és mtsai, 2018) használtuk a W leképezés meghatározására, amelynek célfüggvénye a következők szerint alakul

$$\min_W \sum_{i=1}^{n} \Big(-2x_i^\intercal W^\intercal y_i + \frac{1}{k} \sum_{y_j \in \mathcal{N}(Wx_i)} x_i^\intercal W^\intercal y_j + \frac{1}{k} \sum_{Wx_j \in \mathcal{N}(y_i)} x_j^\intercal W^\intercal y_i \Big),$$

ahol az (x_i, y_i) kontextuális reprezentációk olyan párosait jelölik, amelyeket a célnyelvre, illetve a forrásnyelvre szabott nyelvi modellből nyertünk ki egy-egy azonos minőségben előforduló fordítási szópár vonatkozásában, $\mathcal{N}$ pedig a tanítás során fölhasznált vektorok közül tér vissza az argumentumában szereplő vektor legközelebbi szomszédjaival.

4. Kísérletek

Azon kísérleteink során, amelyben (triviális eszközökkel) többnyelvű környezetben előtanított nyelvi modellek segítségével teremtettük meg a forrás-,és a célnyelv közötti kapcsolatot, a 24-rétegből álló, több mint 100 eltérő nyelv feldolgozását támogató XLM-RoBERTa (Conneau és mtsai, 2020) modellre (a továbbiakban röviden XLM-R) támaszkodtunk. Azon esetekben, amikor a nyelvek közötti kapcsolatot utólagosan, egy leképezés tanulásával hoztuk létre, olyankor a 24 rétegből álló `bert-large-cased` modellt használtuk az angol tanítószövegek rejtett reprezentációinak meghatározására, míg a teszteléskor a 12 réteg alkotta `huBERT` modellt vettük igénybe. Az említett modelleket a `transformers` (Wolf és mtsai, 2019) könyvtárt használva értük el.

Mivel a transzformer architektúrán alapuló, nagy előtanított nyelvi modellek eltérő rétegei más típusú feladatok elvégzésére specializálódhatnak (Tenney és mtsai, 2019; Reif és mtsai, 2019) – és mivel a szemantikus viszonyok jellemzően a háló kései rétegeiben manifesztálódnak – a kísérleteink során az enkóderek utolsó négy rétegéből (illetve a forrás- és célnyelv kezelésére dedikáltan létrehozott nyelvspecifikus enkódereket használó kísérleteink során ezek kombinációiból) jövő reprezentációk alkalmazását vizsgáltuk.

A ritkítással létrehozott kontextuális reprezentációk megalkotása során a 3. fejezetben leírtak szerint jártunk el. Hiperparamétereinket (Berend, 2020a) nyomán $k = 3000$, valamint $\lambda = 0,05$ értékekben határoztuk meg. A Berend (2020a) által alkalmazott módszertantól azon az egy ponton tértünk el, hogy mi nem alkalmaztuk a jelentésprototípusok reprezentációinak létrehozása során azt a normalizáló lépést (Bouma, 2009), amely előzetes vizsgálataink szerint a jelen felállásban minimálisan rontotta volna az eredményeket.

4.1. A huBERT és BERT közötti leképezés tanulása

Az RCSLS módszer alkalmazása során az azonos kontextusban álló, megegyező jelentéssel bíró szópárokat a nyelvtanulókat segítő Tatoeba platform alapján létrehozott korpuszból (Tiedemann, 2012) nyertük ki a `datasets` (Lhoest és mtsai,

2021) könyvtár segítségével. Az egyes fordított mondatpárokból származó azon (s, t) szópárok kontextuális reprezentációira tekintettünk a megfeleltetés tanulása során alkalmas horgonypontként, amelyekre teljesült, hogy a forrásnyelvi mondatból jövő s szó lehetséges fordításai között megtalálható volt a célnyelvi mondatban szereplő t szó, és ugyanez t irányából nézve is igaz (vagyis s a t szó egy lehetséges fordítása). Annak ellenőrzésére, hogy egy adott szó egy másik szó fordítása-e, a `word2word` erőforrást (Choe és mtsai, 2020) hívtuk segítségül. A leírtak alapján egyebek mellett az alábbi mondatpárból az aláhúzással és azonos színnel megjelölt szópárokat nyertük ki:

{'hu:' 'A csigák lassan másznak.', 'en': ' Snails move slowly.}

A nyelek közötti kapcsolatot megteremtő W mátrixot húszezer azonos kontextusban szereplő fordítási pár, a forrás- és célnyelv utolsó négy rétegének valamelyikéből származó kontextuális reprezentációjának megfeleltetése mentén hoztuk létre.

4.2. A kiértékelő adatbázis

A kiértékelésünk során használt XL-WSD adatbázis magyarra vonatkozó tesztalmazán 3484 különböző lemma (4138 különböző ragozott alakjának) összesen 4428 előfordulásának jelentésegyértelműsítését kell elvégezni a BabelNet (Navigli és Ponzetto, 2012) jelentéskészletével összhangban. A validációs halmazban 1021 egyedi lemma (1084 ragozott formájának) 1107 címkézett előfordulása található. A kiértékelés során alkalmazott minőségi mutató gyanánt a jelentésegyértelműsítés esetén megszokott F-mértéket használtuk. A benchmark adatbázisra vonatkozó bővebb statisztikák az adatbázist bemutató cikkben találhatók (Pasini és mtsai, 2021).

4.3. Eredmények

Az 1. táblázatban az látható, hogy miként alakultak a többnyelvű XLM-R enkóder használata mellett kapott eredményeink a kontextuális reprezentációk módosítatlanul hagyása, valamint a korábbiakban leírtak szerint végrehajtott ritkítása esetén az eltérő rejtett rétegek alkalmazása mellett. Amint az látható, mind a reprezentációk ritkítása, mind pedig az érintetlenül hagyása esetén a 21. réteg szolgáltatta a leghasznosabb információt, továbbá a ritkítás által hozott javulás mértéke jellemzően +5 pont körül mozgott (az utolsó, 24. réteg esetét leszámítva, ahol a változás mértéke csupán +0,5 volt).

A 2. táblázatban annak a megközelítésnek az eredményeit közöljük, amelyeket az általunk javasolt, az egynyelvű modellek rejtett rétegei közötti transzformáció alkalmazásának használatával értünk el. A vizsgálatainkat – a korábban említett módon – a forrásnyelvi, valamint a célnyelvi modellek utolsó négy rétegei közötti kapcsolat megteremése mellett végeztük el, azaz a BERT és a huBERT modellekből a kísérleteink során fölhasznált rétegek a $\{21, 22, 23, 24\} \times \{9, 10, 11, 12\}$ Descartes-szorzatból kerültek ki. A 16 kombináció közül a 2. táblázat azokat az eseteket tartalmazza, amelyek a legjobb teljesítményt voltak képesek nyújtani a

Réteg	Módosítatlan	Ritkítás után
21	65.42	**70.01**
22	64.32	69.17
23	63.69	68.13
24	62.78	63.28

1. táblázat. A többnyelvű XLM-R enkóder (és abból származtatott ritka jelentésreprezentációk) használata mellett a háló eltérő rejtett rétegeiből kinyert kontextualizált reprezentációkkal kapott eredmények.

kiértékelő adatbázis validációra szánt részhalmazán. Ez a módosítatlan vektorok használata esetén a 22. és 12., míg a ritkításon átesett vektorok esetén a 23. és 12. rétegek használatát jelentette a forrásnyelv feldolgozására szolgáló BERT, valamint a tesztelés során látott magyar nyelvű szövegek feldolgozását végző huBERT vonatkozásában. A 2. táblázat eredményeinek az 1. táblázatban foglaltakkal való összehasonlításából egyértelműen kitűnik, hogy a többnyelvű modellek egynyelvű modellekre történő lecserélésével komoly javulásokat tudtunk elérni. Érdemes továbbá megjegyezni, hogy a Pasini és mtsai (2021) által a magyar részkorpuszon elért eredményei 47.29 és 68.36 F-mérték között mozognak, amely hatékonyságot a javasolt eljárásunkkal sikerült jelentősen meghaladni.

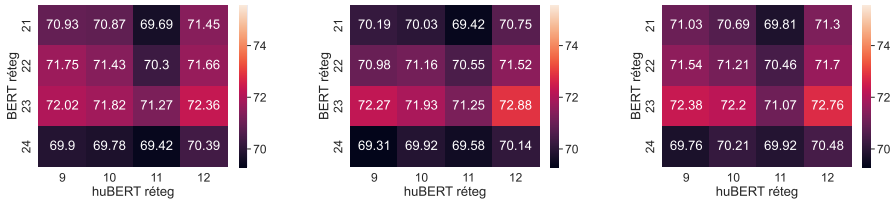
Módosítatlan		Ritkítás után	
Validációs halmaz	Teszt halmaz	Validációs halmaz	Teszt halmaz
73.44	72.76	**74.80**	**75.09**

2. táblázat. A leképezés tanulása mellett a validációs halmaz alapján kiválasztott legjobban teljesítő rendszerek eredménye.

4.4. Szótöredékek kezelése

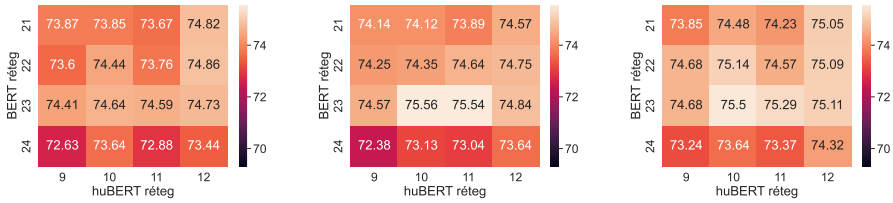
A transzformer alapú modellek jellegzetessége, hogy az inputszekvenciákat a feldolgozásukat megelőzendő szótöredékek (szubtokenek) sorozatára bontják föl. Ebből adódóan ahhoz, hogy a szószintű rejtett reprezentációk megalkotására képessé váljunk, szükségünk van valamilyen aggregáló (pooling) eljárásra, ami az adott esetben több szótöredékre bontott szavak rejtett reprezentációiból kialakítja a szó egészéhez társítandó vektoros kontextuális reprezentációt. Az eddigiekben bemutatott kísérleteink során azt a gyakran használt módszert alkalmaztuk, amelyik a szószintű kontextuális vektorokat az azokat alkotó szótöredékek kontextuális vektorainak átlagolásával hozza létre.

Ács és mtsai (2021) azt vizsgálták, hogy a különféle transzformer alapú előtanított nyelvi modellek magyar szövegeken elvégzett morfoszintaktikai osztályozásának hatékonysága mennyiben függ a tokenszintű kontextuális reprezentációk



(a) Első szubtoken (b) Utolsó szubtoken (c) Szubtokenek átlagolása

1. ábra: A módosítatlan kontextuális reprezentációk használata mellett a különböző forrás-, és célnyelvi enkóderrétegből jövő rejtett reprezentációkra támaszkodó megoldások eredményei eltérő pooling stratégiák alkalmazása esetén.



(a) Első szubtoken (b) Utolsó szubtoken (c) Szubtokenek átlagolása

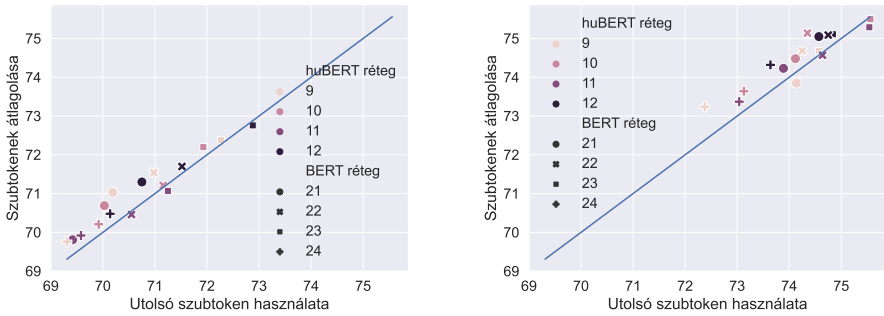
2. ábra: A ritkításon átesett kontextuális reprezentációk használata mellett a különböző forrás-, és célnyelvi enkóderrétegből jövő rejtett reprezentációkra támaszkodó megoldások eredményei eltérő pooling stratégiák alkalmazása esetén.

létrehozása során alkalmazott különböző alternatív aggregáló eljárások megválasztásától. További kísérleteink során a célnyelvünk vonatkozásában mi is az Ács és mtsai által vizsgált lehetőségek összehasonlítását végeztük el, amelyek a következők voltak:

- *első*: tokenen belüli első szubtoken rejtett vektorának használata,
- *utolsó*: tokenen belüli utolsó szubtoken rejtett vektorának használata,
- *átlag*: tokenen belüli szubtokenekhez tartozó rejtett vektorok átlagolása.

Az 1. ábra, valamint a 2. ábra a különféle stratégiák megválasztása esetén elért eredményeinket foglalják össze a BERT és a huBERT modell utolsó négy rétegeiből kinyert módosítatlan, valamint a ritkításon átesett kontextuális vektorok (és a közöttük tanult leképezés) használata esetén. Mindkét ábrából kitűnik, hogy a forrásnyelvi BERT modell tekintetében az utolsó rétegre támaszkodó próbálkozások szerepeltek a legrosszabbul.

Ennek hátterében két magyarázat is állhat: az egyik, hogy a BERT utolsó rétegéből származó reprezentációk már a forrásnyelv esetében is kevésbé alkalmasak a jelentésegyértelműsítési feladat elvégzésére, a másik pedig, hogy ezek a



(a) Hagyományos (b) Ritkítás után

3. ábra: A forrás-és célnyelvspecifikus enkóderek különböző rétegkombinációi esetén kapott eredmények összehasonlítása eltérő pooling stratégiák alkalmazása esetén.

vektorok a forrásnyelv esetében még a többi vizsgált réteg használatával összemérhető eredmény elérésére képesek ugyan, a nyelvközi transzfer minősége azonban lerontja ezen modelleknek a zero-shot helyzetben való alkalmazásának eredményességét. A 3. táblázatban közölt eredmények az első magyarázatot valószínűsítik, az legalábbis mindenképp igaz, hogy a különböző módszerek által az egyes rétegek mentén az angol nyelvű tesztadatokon elért F1-mértékben kifejezett eredmények a BERT utolsó rétegének használata esetén mutatkoztak a legalacsonyabbaknak abban a helyzetben is, amikor a nyelvközi transzfer elvégzésére nem volt szükség, hiszen mind a tanítóadatok mind pedig a tesztelésre szánt adatok angolul álltak rendelkezésre.

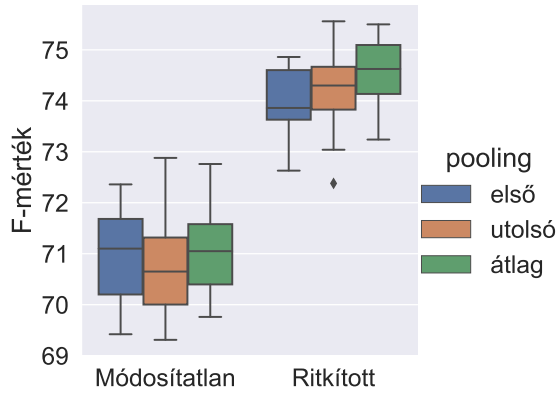
Réteg	Módosítatlan	Ritkítás után
21	74.39	77.45
22	74.87	77.60
23	74.45	77.86
24	73.58	76.21

3. táblázat. A BERT egyes rétegeiből jövő kontextuális vektorok fölhasználásával elért jelentésegyértelműsítési eredmények angol nyelv esetén.

Az 1. ábra, illetve a 2. ábra összevetéséből az is kitűnik, hogy a kontextuális vektorok ritkítása a többnyelvű modellek használatánál látottakhoz hasonlóan az eredmények nagymértékű javulását eredményezte a minden egyéb tekintetben azonosan létrehozott és kiértékelt, de ritkításon át nem esett vektorok használatához képest. Megfigyelhető továbbá, hogy míg a legjobb eredményt az utolsó szubtokenre támaszkodó aggregáló eljárással értük el, összességében nem jelent-

hető ki, hogy a kontextuális szóreprezentációk előállítása során az utolsó szótöredékvektor használata egyértelműen célravezetőbb lenne az egyes szavakhoz tartozó szótöredékvektorok átlagolásánál.

Mindezt a 3. ábra is alátámasztja, ahol egy-egy pont azt reprezentálja, hogy miként viszonyult azon rendszereknek az egymáshoz való eredménye, amelyek minden hiperparaméter vonatkozásában ugyanúgy lettek létrehozva a szóvektorok aggregálása során alkalmazott stratégiát leszámítva. Páros t-próba alkalmazásával úgy találtuk, hogy az átlagolással nyert kontextuális szóvektorok alkalmazása mellett kapott eredmények átlaga szignifikánsan magasabb az utolsó szótöredékből származó kontextuális szóvektorokkal kapott eredményekhez képest mind a módosítatlan vektorok ($p < 0.001$), mind pedig a ritkításon átesett vektorok használata esetén ($p < 0.003$). A különböző vektoraggregálási stratégiák mentén kapott eredmények eloszlását a 4. ábrán is megfigyelhetjük.



4. ábra: A különböző tokenaggregálási stratégiák alkalmazásának hatásai a módosítatlan, valamint a ritkított esetben.

5. Konklúzió

Cikkünkben azt vizsgáltuk, hogy a specializáltan egy adott nyelv feldolgozására létrehozott neurális nyelvmodellek használata milyen előnyökkel jár a soknyelvű nyelvi modellek alkalmazásához képest olyan esetekben, amikor tanítóadatok nem állnak rendelkezésünkre a feldolgozni kívánt forrásnyelven. Mindehhez egy olyan lineáris transzformáció létrehozására tettünk javaslatot, ami az egymástól függetlenül tanított egynyelvű nyelvi modellek reprezentációi közötti kapcsolat megteremtését szolgálja. Az előzőeken túl bemutattuk azt is, hogy a kontextuális jelentésreprezentációk szótártanuláson alapuló, felügyelet nélküli módszerrel

történő ritkításával jelentősen javíthatók a zero-shot módon elvégzett jelentésegyértelműsítési eredmények. Mindezeken felül megvizsgáltuk a különböző szubtokenaggregáló stratégiákat is, és arra jutottunk, hogy az általunk vizsgált feladaton, illetve nyelvközi transzfer alkalmazása esetén a szavakat alkotó szóvektorok átlagolása teljesített a legjobban. A kísérleteink során használt forráskódok az https://github.com/begab/sparsity_makes_sense URL-ről érhetők el.

Köszönetnyilvánítás

A dolgozatban szereplő kutatási eredmények létrejöttét az Innovációs és Technológiai Minisztérium és a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal támogatta a Mesterséges Intelligencia Nemzeti Laboratórium keretében.

Hivatkozások

Ács, J., Kádár, Á., Kornai, A.: Subword pooling makes a difference. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 2284–2295. Association for Computational Linguistics, Online (Apr 2021), https://aclanthology.org/2021.eacl-main.194

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 2020a), https://aclanthology.org/2020.emnlp-main.683

Berend, G.: Word sense disambiguation for Hungarian using transformers. In: Berend, G., Gosztolya, G., Vincze, V. (szerk.) XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020). p. 3–13. Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged (2020b)

Berend, G.: Mitigating the knowledge acquisition bottleneck for Hungarian word sense disambiguation using multilingual transformers. In: XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2021). pp. 77–89. Szeged (2021)

Bouma, G.: Normalized (pointwise) mutual information in collocation extraction. In: From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009. vol. Normalized, pp. 31–40. Tübingen (2009)

Choe, Y.J., Park, K., Kim, D.: word2word: A collection of bilingual lexicons for 3,564 language pairs. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 3036–3045. European Language Resources Association, Marseille, France (May 2020), https://aclanthology.org/2020.lrec-1.371

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp.

8440–8451. Association for Computational Linguistics, Online (Jul 2020), https://www.aclweb.org/anthology/2020.acl-main.747

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

Gale, W.A., Church, K.W., Yarowsky, D.: A method for disambiguating word senses in a large corpus. Computers and the Humanities 26(5), 415–439 (Dec 1992), https://doi.org/10.1007/BF00136984

Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H., Grave, E.: Loss in translation: Learning bilingual word mapping with a retrieval criterion. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2979–2984. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018), https://aclanthology.org/D18-1330

Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In: Proceedings of the 5th Annual International Conference on Systems Documentation. pp. 24–26. SIGDOC '86, ACM, New York, NY, USA (1986), http://doi.acm.org/10.1145/318723.318728

Lhoest, Q., Villanova del Moral, A., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., Davison, J., Šaško, M., Chhablani, G., Malik, B., Brandeis, S., Le Scao, T., Sanh, V., Xu, C., Patry, N., McMillan-Major, A., Schmid, P., Gugger, S., Delangue, C., Matussière, T., Debut, L., Bekman, S., Cistac, P., Goehringer, T., Mustar, V., Lagunas, F., Rush, A., Wolf, T.: Datasets: A community library for natural language processing. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 175–184. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021), https://aclanthology.org/2021.emnlp-demo.21

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

Miller, G.A., Chodorow, M., Landes, S., Leacock, C., Thomas, R.G.: Using a semantic concordance for sense identification. In: HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994 (1994), https://www.aclweb.org/anthology/H94-1046

Navigli, R.: Word sense disambiguation: A survey. ACM Comput. Surv. 41(2) (Feb 2009), https://doi.org/10.1145/1459352.1459355

Navigli, R., Ponzetto, S.P.: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell. 193, 217–250 (Dec 2012), https://doi.org/10.1016/j.artint.2012.07.001

Pasini, T., Raganato, A., Navigli, R.: Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation. Proceedings of the AAAI Conference on Artificial Intelligence 35(15), 13648–13656 (May 2021), https://ojs.aaai.org/index.php/AAAI/article/view/17609

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018), https://www.aclweb.org/anthology/N18-1202

Reif, E., Yuan, A., Wattenberg, M., Viegas, F.B., Coenen, A., Pearce, A., Kim, B.: Visualizing and measuring the geometry of bert. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (szerk.) Advances in Neural Information Processing Systems. vol. 32, pp. 8594–8603. Curran Associates, Inc. (2019)

Taghipour, K., Ng, H.T.: One million sense-tagged instances for word sense disambiguation and induction. In: Proceedings of the Nineteenth Conference on Computational Natural Language Learning. pp. 338–344. Association for Computational Linguistics, Beijing, China (Jul 2015), https://www.aclweb.org/anthology/K15-1037

Tenney, I., Das, D., Pavlick, E.: BERT rediscovers the classical NLP pipeline. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 4593–4601. Association for Computational Linguistics, Florence, Italy (Jul 2019), https://www.aclweb.org/anthology/P19-1452

Tiedemann, J.: Parallel data, tools and interfaces in OPUS. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). pp. 2214–2218. European Language Resources Association (ELRA), Istanbul, Turkey (May 2012), http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf

Vincze, V., Szarvas, Gy., Almási, A., Szauter, D., Ormándi, R., Farkas, R., Hatvani, Cs., Csirik, J.: Hungarian word-sense disambiguated corpus. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). European Language Resources Association (ELRA), Marrakech, Morocco (May 2008)

Weaver, W.: Translation. In: Locke, W.N., Boothe, A.D. (szerk.) Machine Translation of Languages, pp. 15–23. MIT Press, Cambridge, MA (1949/1955), reprinted from a memorandum written by Weaver in 1949.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Brew, J.: Huggingface's transformers: State-of-the-art natural language processing (2019)