Nesze semmi, fogd meg jól!

Zéró kopulák automatikus felismerése neurális gépi fordítással

Dömötör Andrea[1,3], Yang Zijian Győző[2,3], Novák Attila[2,3]

[1]Pázmány Péter Katolikus Egyetem Bölcsészet- és Társadalomtudományi Kar
2087 Piliscsaba, Egyetem u. 1.
[2]Pázmány Péter Katolikus Egyetem Információs Technológiai és Bionikai Kar
[3]MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
1083 Budapest, Práter u. 50/a.
{domotor.andrea, yang.zijian.gyozo, novak.attila}@itk.ppke.hu

Kivonat Kutatásunkban a nominális mondatok zérókopula-jelenségével foglalkozunk, miszerint bizonyos default esetekben a predikatív névszók önmagukban, testes segédige jelenléte nélkül is betölthetik az állítmányi funkciót. Ennek gépi kezelésére létrehoztunk egy eszközt, amely a zéró kopulás mondatok automatikus felismerésére alkalmas, mi több képes a zéró kopulát a mondatok megfelelő helyére beilleszteni. Az általunk létrehozott eszköz in-domain, azaz a tanítóanyaggal megegyező forrásból származó tesztanyagban közel 90%-os pontossággal képes a zéró kopulák helyes beillesztésére.
Kulcsszavak: zéró kopula, szintaxis, gépi tanulás, gépi fordítás, korpusznyelvészet

1. Bevezetés

A zéró kopula jelensége, miszerint bizonyos default esetekben a predikatív névszók önmagukban, testes segédige jelenléte nélkül is betölthetik az állítmányi funkciót, számos nyelvben ismert. A magyarban a kijelentő mód, jelen idő, 3. személy ilyen default eset.

(1) a. Fábián elvtárs is vámpír **Ø**.
b. Lehetetlen, hogy Fábián elvtárs is vámpír **legyen**.
c. Fábián elvtárs is vámpír **volt**.
d. Én is vámpír **vagyok**.

Kutatásunk célja egy olyan eszköz kidolgozása, amely képes az (1a) típusú mondatok megfelelő helyére beilleszteni a zéró kopulát. „Megfelelő hely" alatt azt a pozíciót értjük, ahol nem default esetben a testes kopula lenne (vö. 1a és c).

Fontos megjegyezni, hogy jelen kutatásban csak azokat tekintjük kopulás mondatnak, ahol az adott default esetben soha nem fordul elő testes ige. Nem soroljuk ide a (2)-típusú létige–semmi váltakozásokat (az alábbi példák forrása az MNSz2 (Oravecz és mtsai, 2015)). Más szóval jelen tanulmány csak a nominális mondatok zérókopula-jelenségével foglalkozik, és nem terjed ki az opcionálisan elhagyható egzisztenciális létigékre (2a-b) vagy a címek szerkezeti sajátosságaira (2c).

(2) a. Ott (van) a csodarendszer, hát alkalmazzák!

b. Ennek semmi értelme (nincs), csak eszembe jutott.

c. Veszélyben (van) az olajellátás.

Célunk elsősorban a korpusznyelvészeti kutatások támogatása: egy olyan eszközt szeretnénk kifejleszteni, amely segítségével a nominális mondatok kvantitatív vizsgálatára alkalmas, nagy méretű korpusz hozható létre. Emellett a zéró kopulás mondatok automatikus felismerése a számítógépes mondatelemzés és szövegfeldolgozás számára is hasznos információ lehet.

2. Kapcsolódó munkák

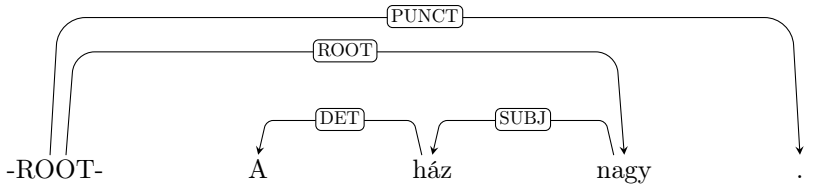
Simkó és Vincze (2017) a zéró kopulás mondatok függőségi elemzésében háromféle megközelítéssel kísérletezik: a funkciófejes, a tartalmas fejes és a komplex címkés elemzéssel. A funkciófej-elemzést követi a Szeged Treebank (Vincze és mtsai, 2010), ahol a névszói állítmány egy üres fejhez kapcsolódik, így a zéró kopulás mondatok a testes kopulásokkal analóg elemzést kapnak. A tartalmasfej-elemzés ezzel szemben nem enged meg üres fejeket, ezért a kopulás mondatok feje minden esetben a névszói állítmány. Ez egyébként megfelel a Universal Dependencies elveinek (Nivre, 2014). A komplex címkés elemzésben a kopula hiányát a névszói állítmány speciális ROOT-VAN-PRED címkéi jelzik. A tanulmány kísérletei során a Bohnet parsert (Bohnet, 2010) tanították be a szerzők a három elemzési módszerre. Eredményeik szerint a parser a funkciófejes elemzést tanulta meg a legsikeresebben, ami azt mutatja, hogy a zéró kopulák beillesztése valóban hasznos a függőségi elemzés számára.

Az üres funkciófejek automatikus beillesztésével kapcsolatos kísérleteket mutat be Seeker és mtsai (2012). Kutatásuk célja a zéró szóalakok (ellipszis miatt hiányzó szavak vagy zéró kopulák) megjóslása a függőségi elemzés során. A cikk három módszert mutat be: az elsőben az üres fejeket a parser illeszti be az elemzés során, a másodiknál az üres fejek a címkekészletben vannak kódolva, míg a harmadik esetén az üres fejek szükségességéről egy osztályozó dönt az elemzés előtt. Számunkra ez utóbbi módszer a leginkább érdekes, hiszen itt a miénkhez hasonló feladatról van szó. A fő különbség, hogy Seeker és mtsai (2012) az üres fejeket a tagmondat elejére illeszti be, és nem oda, ahol a felszíni szerkezetben a zéró kopula helye lenne, így valóban csak osztályozást végez. A másik különbség, hogy a cikkben ismertetett osztályozó elemzett szövegekkel dolgozik, azaz morfológiai információt is használ, míg esetünkben csak az elemzetlen mondatok

állnak rendelkezésre. Csak a zéró kopulás mondatokat tekintve Seeker és mtsai (2012) 83,6%-os pontosságot, 69,2%-os fedést és 75,8-os F-mértéket ért el. A saját módszerünkre is készítettünk ezzel összevethető, csak az osztályozás sikerességét mérő kiértékelést.

Tudomásunk szerint a zéró kopula felszíni szerkezetbe való beillesztésére alkalmas eszköz nem áll rendelkezésre a magyar nyelvre. Az elérhető szintaktikai elemzők sem alkalmazhatók erre a feladatra, ezek ugyanis, a Universal Dependencies elveit követve, nem illesztenek zéró kopulát az elemzéseikbe. Ennek megfelelően az `e-magyar` (Váradi és mtsai, 2018) elemző a zéró kopulás mondatok esetén a tartalmasfej-elemzést alkalmazza (3. példa).

(3)



A korpuszokat tekintve említettük, hogy a Szeged Treebankben vannak zéró kopulát pótló üres fejek. Ezek azonban nem „valódi” mondatrészként, csak virtuális csomópontokként szolgálnak, így a mondatbeli pozícióknak nem tulajdonítottak jelentőséget a korpusz készítői. Az üres fejeket jelölő szimbólumok így gyakorlatilag véletlenszerű helyeken jelennek meg a felszíni szerkezetben, ezért ezeket nem tudjuk közvetlenül zéró kopulás tanítóanyagként felhasználni az általunk kitűzött célra. Előnye azonban a Szeged Treebanknek, hogy bináris osztályozási feladatra (zéró kopulás-e a mondat vagy nem) gold standard adatként rendelkezésre áll. A korpusz 16003 darab zéró kopulás mondatot tartalmaz, ami a teljes méretének nagyjából 17%-a. Ezt az empirikus arányt használtuk fel a tanító- és tesztanyagaink összeállításában.

3. Módszer

Kutatásunkban a feladat megoldásához a gépi fordítás módszerét alkalmaztuk, melynek lényege, hogy transzformációt képez tetszőleges forrás- és célnyelvi mondatok között, ahol a rendszer betanításához nem kell más, mint egy kétnyelvű párhuzamos korpusz.

A gépi fordítás módszerével való megközelítés indokolt, hiszen a forrás- és a célnyelvi mondat azonos, kivéve a zéró kopulás mondatpárokat, melyben a célnyelvi mondatban a zéró kopula helyén egy <zerokop> címke áll.

Munkánk során a Marian NMT (Junczys-Dowmunt és mtsai, 2018) nevű keretrendszert használtuk, ami egy c++ nyelven íródott szabadon hozzáférhető programcsomag. Könnyen telepíthető, jól dokumentált, memória- és erőforrás-optimális implementációjának köszönhetően [1] az akadémiai felhasználók és fejlesztők által leggyakrabban használt eszköz (Barrault és mtsai, 2019).

[1] https://marian-nmt.github.io/

Az NMT tanításához a jelenleg „state-of-the-art" Transformer (Vaswani és mtsai, 2017) modellt és Sentece Piece (Kudo és Richardson, 2018) tokenizálót használtuk. A rendszer beállításai és paraméterei a következők:

- Sentence Piece: szótárméret: 16000; egy szótár a forrás- és egy a célnyelvi korpusznak; karakter lefedettség a teszt korpuszon: 100%
- Transformer modell: enkóder és dekóder rétegeinek száma: 6; transformer-dropout: 0.1;
- learning rate: 0,0003; lr-warmup: 16000; lr-decay-inv-sqrt: 16000;
- optimizer-params: 0,9 0,98 1e-09; beam-size: 6; normalize: 0,6
- label-smoothing: 0.1; exponential-smoothing

4. A tanítóanyag

A rendszer tanításához olyan tanítóanyagra van szükség, ahol a zéró kopulák jelölve vannak a mondatokban. Mint említettük, eddig egy ilyen korpusz létezik, a Szeged Treebank, ez azonban egyrészt túl kicsi, másrészt a zéró kopulák helyét tekintve rendszertelen. Elkerülhetetlen volt tehát, hogy új, saját zérókopula-korpuszt hozzunk létre.

Az alapötlet az volt, hogy csináljunk zéró kopulás mondatokat a testes kopulás mondatokból. Ehhez először is arra volt szükség, hogy a testes kopulákat elkülönítsük a lexikális VAN létige azonos alakjaitól (azaz a lokatív, az egzisztenciális és a birtokos igétől). Első kísérletként kipróbáltuk, hogyan oldja meg ezt a feladatot az `e-magyar` automatikus szövegelemző rendszere. Az igék osztályozásához elkészíttettük egy 1000 mondatos teszthalmaz függőségi elemzését, és a kapott elemzés szerint osztályoztuk a kérdéses létigéket. (A teszthalmaz egy random minta volt, 598 mondatban szerepelt benne kopula, és 402-ben lexikális ige.) Ha a kérdéses létigéhez tartozott vele `PRED` relációban álló névszó, akkor *kopula* címkét kapott, egyéb esetben pedig *lexikális*at. A kiértékelés során a kopulás találatok pontosságát és fedését mértük. A módszer 87,8%-os pontosságot és 81,0%-os fedést ért el.

Az eredmények szerint a különböző létigetípusok megfelelő elemzésének kiválasztása nem triviális feladat egy automatikus elemzőrendszer számára. A nehézség egyik oka az lehet, hogy a magyar nem konfigurációs nyelv, így a szórend nem segít abban, hogy megállapítsuk az egyes szavak szintaktikai szerepét. A magyar nyelv másik, automatikus feldolgozást nehezítő tulajdonsága a pro-drop, emiatt nem lehet a létigetípusok megkülönböztetését az esetrag nélküli névszók számára alapozni, hiszen ha csak egyetlen ilyen is van, akkor is könnyen lehetséges, hogy ez az egyetlen névszó predikatív, az alany pedig nincs jelölve.

Alternatív módszerként egy angol-magyar párhuzamos korpuszt használtunk kopulás mondatok gyűjtésére, az angol nyelv konfigurációs jellegének köszönhetően ugyanis az angol mondatokon sokkal könnyebb lokális információk alapján, szabályok segítségével szintaktikai döntéseket meghozni. Vagyis a magyar mondatok angol megfelelői segítenek abban, hogy meghatározzuk a létige aktuális mondatbeli funkcióját (típusát).

Az adatgyűjtéshez egy lemmatizált, morfológiailag elemzett és egyértelműsített, szó szinten megfeleltetett angol-magyar párhuzamos korpuszt használtunk (Novák és mtsai, 2019). Ennek alapja az OPUS OpenSubtitles korpusz (Lison és Tiedemann, 2016), amely összesen 644,5 millió tokenből álló megfeleltetett mondatpárokat tartalmaz. Az angol oldal elemzése a morpha lemmatizálóval (Minnen és mtsai, 2001) és a Stanford taggerrel (Toutanova és mtsai, 2003) történt. A magyar oldalon a PurePos (Orosz és Novák, 2013) és a Humor (Novák, 2014) eszközök végezték el a morfológiai elemzést. Az elemzett szövegekben mindkét oldalon minden eredeti tokent két token reprezentál: (1) a szótő és a fő szófajcímke, illetve (2) az egyéb morfológiai címkék. Az előfeldolgozott mondatokon a fast align program (Dyer és mtsai, 2013) segítségével szó szintű megfeleltetések készültek. Ehhez a morfológiai címkék külön tokenként ábrázolása előnyös, mert így lehetőség van arra, hogy bizonyos, lexikális megfelelővel nem rendelkező szavakat a morfológiai címkékkel kössük össze, például az angol prepozíciókat a magyar oldalon az esetragokat jelző címkékhez.

Az előzőekben leírt párhuzamos korpuszból kiválasztottuk azokat a mondatokat, ahol a magyar oldalon a létige vagy a kopula valamilyen múlt idejű, harmadik személyű alakja (továbbiakban VOLT) szerepel. Ezeket a mondatokat egy szabályalapú algoritmussal címkéztük.

A címkéző algoritmus első lépésben megnézi a VOLT-nak megfeleltetett angol tokeneket. Ha ezek között szerepel nem segédigei *have* vagy expletív *there*, a mondat `lexikális` címkét kap. Ha a VOLT *be*-vel van megfeleltetve, akkor a mondat lexikális igés és kopulás egyaránt lehet, ekkor tehát további vizsgálati lépésekre van szükség. Ha az előbbi tokenek egyike sem szerepelt a VOLT-nak megfeleltetett angol tokenek között, akkor a mondat kikerült a címkézendő anyagból, ekkor ugyanis feltehetően az eredetitől nagyon eltérő fordításról lehet szó, így a mondat angol megfelelője nem megbízható kiindulópontja a címkézésnek.

Ha a VOLT az angol oldalon *be*-nek felel meg, akkor a program megkeresi az angol mondatban azt a „kulcsszót", amely alapján a címkézés megtörténik. A kulcsszó feltételezésünk szerint vagy egy névszói állítmány, vagy egy nem nominatívuszi argumentum (illetve ezek része). Az algoritmus tehát ezeknek az elemeknek a kanonikus pozícióját keresi az angol mondatban (ami a magyar mondat esetén a szórend kiszámíthatatlansága miatt lehetetlen lenne).

A kulcsszó kiválasztásához először is meg kell állapítani, hogy kijelentő vagy kérdő szórendű mondatról van-e szó. Ezt a kérdőszó megléte, illetve az ige pozíciója alapján ellenőrzi az algoritmus. Kijelentő mondatok esetén a kulcsszó a *be*-t követő első olyan token, amely nem tagadószó vagy NP-t módosító elem (például: *very, more*). Az „alkalmasság" megállapítása elsősorban a szófajcímkék alapján történik. (Ld. 4. példa) Eldöntendő kérdés vagy a *what*, *who*, *whose*, *which*, *how* és *why* kérdőszavak esetén a program az előzőekhez hasonlóan jár el, a szórendváltás miatt plusz egy tokent átugorva. (Ld. 5. példa) Az egyéb kérdószóval (pl. *where*, *when* stb.) bevezetett mondatok `lexikális` címkét kapnak.

(4) a. *Régen ez egy kvalitás volt.*
It used to be **a** quality.
b. *Nem volt otthon.*
He was not **at** home.

(5) a. *Mi volt ez a zaj?*
What was that **noise**?
b. *Miről volt szó?*
What was it **about**?

A kulcsszó kiválasztása után az algoritmus megnézi a kulcsszóhoz rendelt magyar tokeneket, és ezek szófaj- illetve morfológiai címkéi alapján megállapítja a kérdéses létige típusát. Ha a kulcsszóhoz tartozik egy nem nominatívuszi esetet jelölő morfológiai címke, akkor a mondat a `lexikális` címkét kapja. Ha a kulcsszónak névelő vagy nominatívuszi névszó felel meg a magyar oldalon, akkor a mondat címkéje *kopula* lesz.

Némely esetben az algoritmus speciális lexikális szabályokat is tartalmaz, a szófajcímkék ugyanis félrevezetőek lehetnek, például az időjárást és egyéb „környezeti helyzeteket" leíró szerkezetek esetén. Ezeknél a kulcsszókeresés értelemszerűen rossz eredményt ad (6). Ezek a szerkezetek ezért lexikális kivételként vannak kezelve egy névszólista alapján, amely az MNSZ2 kollokációkeresőjével készült.

(6) a. ***Sötét** volt és köd.*
It was **dark** and foggy.

A „környezeti kopulás" szerkezetek mellett van még néhány olyan speciális eset, ahol a kulcsszókeresés félrevezető lehet, ezeket leginkább „állandó fordítási különbségnek" nevezhetnénk. Ez alatt azokat a szerkezeteket értjük, amelyek angolul kopulásak, magyarra viszont lexikális VAN-nal fordítjuk őket. Ennek leggyakoribb esete a *being right* 'igaza van' mondat, de ez alá a speciális lexikális szabály alá soroltam a *being lucky* 'szerencséje van', *being necessary* 'szükség van' és a *being ready* 'kész van' szerkezeteket is.

Az algoritmus teljesítményét ugyanazon az 1000 mondatos mintán értékeltük ki, amelyeket az `e-magyarra` alapozott teszthez is használtunk. Az osztályozó 90,8%-os pontosságot és 91,1%-os fedést ért el, azaz jobban teljesített a függőségi elemzésre alapozott módszernél. Az elért pontosság azonban még így sem közelíti meg egy gold standard tanítókorpusz minőségét.

A hibák elemzése során kiderült, hogy a hibás címkék nagy része nem az algoritmusból, hanem valamilyen „külső körülményből" ered. Tipikusan ilyenek például a hibás szófaji címkézés vagy szómegfeleltetés. Szintén gyakori külső hibaforrás volt az angol eredetitől jelentősen eltérő magyar fordítás. Bár az algoritmus igyekszik az ebből származó problémákat kiküszöbölni azzal, hogy figyelmen kívül hagyja az olyan mondatokat, ahol a VOLT-hoz sem *be*, sem *have* nem volt hozzárendelve, ez a megszorítás azonban még mindig sok olyan mondatot „átenged", ahol a mondatszerkezeti vagy akár jelentésbeli különbségek megnehezítik a címkézést.

Ezeket a címkézési hibákat nehezen lehetne elkerülni, így a továbbiakban ezzel a 90%-os pontosságú kimenettel dolgoztunk. A program által kopulásnak

címkézett mondatokban a testes kopulákat egy `<zerokop>` jelre cseréltük, ezek adták a tanítóanyag pozitív példáit (318843 mondat). Ehhez hozzáadtunk az OpenSubtitles korpuszból 1 millió random mondatot (természetesen ügyelve arra, hogy ne legyen átfedés a zérókopulás mondatokkal), és ebből tanítottuk be az alapmodellt. Ez elég jó pontosságot (89,6), viszont gyenge fedést (58,2) produkált, ami nem meglepő, hiszen a negatív példának szánt 1 millió mondatban valószínűleg sok „valódi", jelöletlen zéró kopula volt. Ennek kiküszöbölésére az OpenSubtitles korpuszból kiszűrtük azokat a mondatokat, ahol a *be* valamilyen jelen idejű alakja szerepel nem segédigei funkcióban az angol oldalon harmadik vagy második személyű alannyal (az utóbbinak az esetek nagy részében harmadik személyű a magyar fordítása). Ugyan az angol kopulás mondatok fordítása nem szükségszerűen kopulás a magyar oldalon és fordítva, ezzel a módszerrel viszonylag hatékonyan ki tudtuk szűrni a hamis negatív példák nagy részét a tanítóanyagból. További szűrésképpen a maradékon lefuttattuk az alapmodellt. A továbbiakban azokat a mondatokat (illetve ezek egy részhalmazát) használtuk negatív példának, amelyekbe a modell nem illesztett be zéró kopulát. Végül, az alapmodellt lefuttattuk a kiszűrt (azaz az angol oldalon *be*-t tartalmazó) mondatokon is, így összesen további 161223 darab zéró kopulás mondatot nyertünk.

Az így rendelkezésre álló adatokból a következő modelleket építettük fel[2]:

- **Eredeti szűrt**: ez a modell a címkéző algoritmus segítségével előállított 318843 zéró kopulás, és 1 millió random kiválasztott nem zéró kopulás mondatot tartalmaz. A tanítóanyagba nem kerültek egyszavas, illetve speciális karaktereket tartalmazó mondatok. Ennek a szűrésnek a zajcsökkentés volt a célja.
- **Bővített szűrt**: A pozitív példákhoz hozzáadtuk az alapmodellel előállított zéró kopulás mondatokat (így összesen 477082 zérót tartalmazó mondatunk lett), a negatív példákat pedig újabb 1 millió random nem zéró kopulás mondattal bővítettük.
- **Eredeti javított**: Az eredeti szűrt modell tanítóanyagába visszakerültek az egyszavas mondatok és néhány olyan speciális karakter, amely a Szeged Treebankben gyakorinak bizonyult (pl. §). Ezen kívül javítottunk néhány, a tesztadatok átnézése során feltűnő könnyen kiszűrhető hibát, amelynek altípusait a (7) alatti példákkal szemléltetjük. Az eredeti modell által zéró kopulásnak jelölt mondatok közül kiszűrtük azokat, ahol a magyar oldalon az eredeti mondatban szereplő *volt(ak)* szóalak valamilyen minden esetben hangsúlytalan (de nem enklitikus) összetevő: kötőszó (7a), vonatkozó névmás (7b) vagy névelő (7d, 7e) után következik. Az előbbieket csak a mindig hangsúlyos lexikális (egzisztenciális (7a) vagy birtokos (7c)) *volt* vagy *voltak* igealak követheti, az utóbbiak pedig gyakorlatilag biztosan az 'ex' jelentésű szintén hangsúlyos *volt* melléknév hibás annotációjával keletkeztek. Tulajdonképpen az első eset speciális változata ezen kívül a vesszőt követő vagy mondat eleji *volt* (7f, 7g). Ebben az esetben ugyanúgy általában kizárt a zéró kopulával való helyettesítés (kivéve a beágyazott vonatkozó mellékmondatok esetét). Ahogy a (7) alatti példák is mutatják, az eredeti algoritmus

[2] A modellek elérhetők: http://nlpg.itk.ppke.hu/projects/zerokopula

nagyrészt akkor hibázott így, ha az eredeti mondat múlt idejű névszói állítmányt és hangsúlyos (lexikális vagy melléknévi) *volt*-ot is tartalmazott, és a szóösszerendelési modell az angol kopulát hibásan hozzákapcsolta az utóbbihoz (is). Emellett az is egy érdekes eset, amikor ugyan kopuláról van szó, de ugyanakkor ellipszis is van a mondatban (7h), ezért múlt időben a *volt* a fókuszos kontrasztív szerkezet miatt kötelezően hangsúlyos, azonban jelen időben muszáj lenne egy másik elemnek megjelennie, hogy legyen, amit fókuszálni lehet. Ez a modell így 314607 zéró kopulás, és 1515204 nem zéró kopulás mondatot tartalmaz. (Azaz, ez a tanítóanyag már megfelel a Szeged Treebankből megállapított empirikus aránynak.)

– **Bővített javított**: Az előző anyaghoz hozzáadtuk az alapmodellel generált zéró kopulás mondatokat, és a negatív példákat is kibővítettük a 17%-os aránynak megfelelően, így ez a modell 475830 zéró kopulás, és 2574207 nem zéró kopulás mondatot tartalmaz.

(7) a. De igen, a legtöbb az volt, de **volt** fehér is, valamint ilyen bébifosbarna is.

b. Az a belépőkártya... volt minden, amim **volt**.

c. Házas volt, és **volt** egy fia.

d. Megértjük, hogy a **volt** férje és Ő üzleti társak voltak.

e. A különleges osztag egy **volt** kihallgatótisztje csinálta.

f. Egy szinttel lejjebb voltak a testőrök, **volt** ejtőernyősök vagy idegenlégiósok.

g. Ahhoz képest, hogy fiú vagy, **voltak** jó válaszok.

h. Nem olyan, mint **volt**.

5. Eredmények

A kiértékeléshez kétféle teszthalmazt használtunk. Az egyik ugyanabból a korpuszból készült, mint a tanítóanyag (OPUS), a másik pedig a Szeged Treebankből. Ez utóbbi azt a célt szolgálja, hogy kipróbáljuk, hogy működik a rendszer más domaineken.

Mindkét tesztkorpusz 2000 mondatot tartalmaz, ebből, a 17%-os empirikus aránynak megfelelően, 340 zéró kopulás. A tesztmondatokat ellenőriztük és kézzel javítottuk, ahol szükséges volt, illetve a Szeged Treebank jelölt zéró kopuláit szintén kézzel kellett a helyükre illeszteni. A végeredmény tehát két, méretben és arányban megegyező, de különböző forrásból származó, és ezáltal különböző szövegtípusokat tartalmazó gold standard tesztkorpusz lett.

A Szeged Treebank mondatain tesztelve azonban váratlan nehézségekbe ütköztünk. Egyrészt kiderült, hogy ezek a mondatok olyan karaktereket is tartalmaznak, amelyek nem szerepelnek az első két modell tanítóanyagában (pl. §),

és ezeket a fordító nem tudta kezelni. Emiatt a javított modellekbe visszakerültek a zajcsökkentés céljával kizárt, speciális karaktereket tartalmazó mondatok. Másrészt gondot okoztak a hosszú mondatok is, az OPUS korpuszból készült tanítóanyag mondatai ugyanis – műfaji sajátosságuknál fogva – jellemzően rövidek voltak. Erre a problémára valószínűleg a tanítóanyag hosszú mondatokkal való kiegészítése lehetne megoldás, de a Szeged Treebankből készült tesztanyagon sajnos nem értünk el olyan eredményt, ami lehetővé tette volna, hogy a program outputját tanítóanyagként használjuk.

Mindegyik modellt lefuttatuk mindkét teszthalmazon, és a kimeneten kétféle kiértékelést végeztünk. Az egyik esetén osztályozási feladatnak tekintettük a zéró kopulák beillesztését, azaz csak azt mértük, hogy hány esetben találja el a program, hogy kell-e zéró kopula a mondatba, azt nem értékeltük, hogy jó helyre illeszti-e be azt. A filmfeliratkorpusz viszonylag egyszerű mondatai esetében egyébként szinte minden esetben, ahol nem a referenciával azonos helyre szúrta be a kopulát az algoritmus, helyes az általa javasolt megoldás is. (4. táblázat) A másik kiértékelésnél már csak azokat a zéró kopulákat tekintettük jó találatnak, amelyek a mondat referenciával azonos pozíciójába kerültek. Az eredményeket az 1. és 2. táblázatok tartalmazzák.

	Osztályozás			**Beszúrás**		
	P	**R**	**F1**	**P**	**R**	**F1**
eredeti szűrt	95,5	84,2	89,5	89,7	79,1	84,1
bővített szűrt	94,9	78,3	85,8	88,7	73,2	80,2
eredeti javított	93,6	82,8	87,9	85,6	75,7	80,4
szűrt javított	94,1	76,0	84,1	86,4	69,8	77,2

1. táblázat. Az OPUS tesztkorpusz eredményei a különböző modellekkel

Az in-domain tesztkorpusz esetén az osztályozási feladaton mindegyik modell magas pontosságot ért el. A fedés tekintetében érdekes módon a korpusz bővítése jelentős visszaesést eredményezett. A beszúrási feladatnál mind pontosságban, mind fedésben az eredeti szűrt modell volt a legjobb, 90, illetve 80%-ot közelítő eredménnyel.

	Osztályozás			**Beszúrás**		
	P	**R**	**F1**	**P**	**R**	**F1**
eredeti szűrt	52,9	29,5	37,8	37,8	21,0	27,0
bővített szűrt	59,6	28,5	38,5	45,1	21,5	29,2
eredeti javított	59,4	25,7	35,9	42,3	18,3	25,6
bővített javított	70,5	29,0	41,1	52,4	21,5	30,5

2. táblázat. A Szeged tesztkorpusz eredményei a különböző modellekkel

Más domaineken azonban már egyik modell sem volt ilyen sikeres. Bár az osztályozási feladat pontosságán sokat javított a korpusz bővítése és javítása (részben talán a speciális karakterek visszakerülésének köszönhetően), az eredmény még így is csak 70% körüli. A beszúrást is a bővített javított modell oldotta meg legjobban, 52%-os eredménnyel. A fedés láthatóan mindkét feladatnál és mindegyik modellnél gyenge, a 30%-ot sem éri el. Ennek egyik oka lehet a fentebb említett technikai problémák miatti „veszteség". Másrészt a Szeged tesztkorpusz különböző műfajú szövegei sok olyan mondattípust tartalmaznak, ami a modellek számára ismeretlen, mert a filmfeliratok műfajára egyáltalán nem jellemző.

	Osztályozás			**Beszúrás**		
	P	**R**	**F1**	**P**	**R**	**F1**
eredeti szűrt	77,7	55,0	64,4	68,0	48,2	56,4
bővített szűrt	80,8	51,7	63,1	71,3	45,7	55,7
eredeti javított	81,4	52,4	63,7	70,1	45,1	54,9
bővített javított	85,4	50,9	63,8	73,9	44,1	55,2

3. táblázat. A tesztkorpuszok összesített eredményei a különböző modellekkel

A két tesztkorpusz eredményeit összegezve (3. táblázat) az látszik, hogy a javított modellek a szűrtekhez képest valamivel magasabb pontosságot, de alacsonyabb fedést produkáltak. Az F-mértéket tekintve nincsenek jelentős különbségek a modellek között, a pontosság- és fedésértékek eltérései nagyjából kiegyenlítik egymást.

	OPUS			**Szeged**		
Modell	**helyes**	**valószínűtlen**	**helytelen**	**helyes**	**valószínűtlen**	**helytelen**
Eredeti szűrt	53.9%	7.7%	38.5%	22.2%	22.2%	55.6%
Bővített szűrt	66.7%	6.7%	26.7%	36.4%	9.1%	54.6%
Eredeti javított	63.6%	4.6%	31.8%	50.0%	6.3%	43.8%
Bővített javított	70.0%	10.0%	20.0%	38.9%	11.1%	50.0%

4. táblázat. Helyes-e a zéró kopula modell által javasolt helye, ahol az nem egyezik a referenciával?

6. Hibaelemzés

A hibásan beillesztett zéró kopulák áttekintésekor feltűnt néhány mondattípus, amelyeket a program látszólag következetesen (vagy legalábbis gyakran) elrontott. Ilyenek voltak például az egyszavas mondatok, ezekbe mindegyik modell hajlamos volt hibás zéró kopulákat beszúrni: **Mióta Ø?; *Elnézést Ø!; *Értem*

Ø. stb. Ez érthető azoknál a modelleknél, ahol a tanítóanyagból kiszűrtük az egyszavas mondatokat, de ugyanezeket az outputokat kaptuk az egyszavas szűrő kikapcsolása után is. Úgy tűnik tehát, hogy a rendszernek mindenképpen nehézséget okoz a nagyon rövid mondatok kezelése.

Hasonlóan általános hibajelenségnek bizonyultak a vonatkozó névmással vagy *mint* kötőszóval összekapcsolt összetett mondatok is (8a és 8b). Ennek valószínűleg a zéró kopulák hasonló helyzetben való gyakorisága lehet az oka. Ugyanígy jellemző hiba, hogy a rendszerek minden esetben zéró kopulát szúrnak be az *ez a(z)...* típusú szerkezetekbe, feltehetően szintén azért, mert ezek gyakran valóban zéró kopulásak (8c). A gyakoriság problémakörébe tartozik még a *tagadószó + NP* szerkezetek hibás zéró kopulával való ellátása is (8d).

(8) a. *Hallottam, hogy megtiltottad a belépést a szentélybe, és a boszorkány Ø, aki magához vette a herceget, eltűnt.

b. *Erzsi híreinek valóságtartalma Ø, mint valami méreg, szívódott fel a szervezetébe.

c. *...mert ez Ø az út a hegy másik felére vezetett el...

d. *A mű nem Ø üzletszerű többszörözése és terjesztése a szabad felhasználás körébe tartozik.

Más forrású lehet, de szintén nagyon gyakori hiba a halmozott jelzős szerkezetek hibás értelmezése is (9). Ezekbe mindegyik modell hajlamos az első jelző után zéró kopulát beilleszteni. Ennek az a valószínű oka, hogy az OPUS korpuszra az ilyen leírások nem jellemzőek, így a rendszernek nem volt esélye ezek helyes kezelését megtanulni.

(9) a. *A bennszülött bivalyszerű Ø, fekete nyakizmai kidagadtak az erőfeszítéstől.

b. *A fordulat a második Ø, 1994-es választás után következett be.

Az előbbihez valamelyest hasonló az a hibatípus, amelyet akár pszicholingvisztikai motiváltságúnak is nevezhetnénk, ezek a mondatok ugyanis a szekvenciális feldolgozás egy pontján (vagy csak egy adott tagmondatot tekintve) valóban zéró kopulásnak tűnhetnek. Tipikusan ilyenek az értelmező szerkezetek (10a), illetve az ellipszisek és a koordinált névszói állítmányok (10b-10d).

(10) a. *Kedden a tokiói tőzsde vezető részvényindexe Ø, a Nikkei 225 mintegy 280 pont, azaz 2,7 százalékos erősödést jelzett.

b. *És mi tudna ezen változtatni,... ha nem egy újabb vihar Ø.

c. *Nem szeretnék valami hatalmas villában vagy panellakásban élni, mert az előbbi túl nagy Ø, az utóbbi túl kicsi lenne a család számára.

d. *A Pénzügyi Szervezetek Állami Felügyelete engedélyezte, hogy Szobonya Csaba Zoltán, az OTP Lakástakarékpénztár Rt. igazgatósági tagja Ø, egyben vezérigazgatója legyen.

Végül meg kell említeni a csak a bővített modellekre jellemző hibajelenségeket. Ezekben előfordult a megszólítások zéró kopulás értelmezése, ami az eredeti modelleknél nem merült fel. Továbbá ezeknél a modelleknél találtuk a „legindokolatlanabb" hibákat is, a rendszer hajlamos volt akár ragozott igék után is zéró kopulát beilleszteni. Ez arra utal, hogy a több iterációs tanítóanyag-gyártásnál fennáll a hibák terjedésének és halmozódásának veszélye.

7. Összegzés

Kutatásunk során létrehoztunk egy eszközt, amely a zéró kopulás mondatok automatikus felismerésére és a zéró kopulák mondatba való beillesztésére alkalmas. In-domain tesztkorpusz esetén az eszköz közel 90%-os pontossággal tudta megfelelő helyre beilleszteni a zéró kopulát. A kutatást kiterjesztettük a Szeged Treebankre, amely jelentős mennyiségű a – főleg egyszerű beszélt nyelvi szövegekből álló – tanítóanyagunktól nagymértékben különböző, és jóval bonyolultabb szerkezeteket tartalmazó jogi, irodalmi, illetve sajtószöveget tartalmaz. Ennek következtében ezen a korpuszon jóval gyengébb teljesítményt mértünk. Cikkünk hibaelemzést is tartalmaz, amelyben áttekintettük rendszerünk jellemző hibatípusait.

Köszönetnyilvánítás

Jelen kutatás a Nemzeti Kutatási Fejlesztési és Innovációs Alap által biztosított támogatással az FK 125217 számú projekt keretében az FK 17 pályázati program, valamint a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal Mesterséges Intelligencia Nemzeti Kiválósági Programja támogatásával a 2018-1.2.1NKP-2018-00008 azonosítójú projekt keretében valósult meg.

Hivatkozások

Barrault, L., Bojar, O., Costa-jussà, M.R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., Zampieri, M.: Findings of the 2019 conference on machine translation (wmt19). In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). pp. 1–61. Association for Computational Linguistics, Florence, Italy (August 2019)

Bohnet, B.: Top accuracy and fast dependency parsing is not a contradiction. In: Proceedings of Coling 2010. pp. 89–97 (2010)

Dyer, C., Chahuneau, V., Smith, N.A.: A Simple, Fast, and Effective Reparameterization of IBM Model 2. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 644–648. Association for Computational Linguistics (2013), http://aclweb.org/anthology/N13-1073

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A.F., Bogoychev, N., Martins, A.F.T., Birch, A.: Marian: Fast neural machine translation in C++. In: Proceedings of ACL 2018, System Demonstrations. pp. 116–121. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)

Kudo, T., Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 66–71. Association for Computational Linguistics, Brussels, Belgium (Nov 2018)

Lison, P., Tiedemann, J.: OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In: Chair), N.C.C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (szerk.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France (may 2016)

Minnen, G., Carroll, J.A., Pearce, D.: Applied morphological processing of English. Natural Language Engineering 7(3), 207–223 (2001), https://doi.org/10.1017/S1351324901002728

Nivre, J.: Nonverbal Predication and Copulas in UD v2. http://universaldependencies.org/v2/copula.html (2014), accessed: 2020-01-04

Novák, A.: A New Form of Humor – Mapping Constraint-Based Computational Morphologies to a Finite-State Representation. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (szerk.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 1068–1073. European Language Resources Association (ELRA), Reykjavik, Iceland (May 2014), aCL Anthology Identifier: L14-1207

Novák, A., Laki, L.J., Novák, B.: Mit hozott édesapám? Döntést – Idiomatikus és félig kompozicionális magyar igei szerkezetek azonosítása párhuzamos korpuszból. In: XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019). pp. 63–71. Szeged University, Szeged (2019)

Oravecz, Cs., Sass, B., Váradi, T.: Mennyiségből minőséget. Nyelvtechnológiai kihívások és tanulságok az MNSz új változatának elkészítésében. In: XI. Magyar Számítógépes Nyelvészeti Konferencia. pp. 109–121. Szegedi Tudományegyetem, Szeged (2015)

Orosz, Gy., Novák, A.: PurePos 2.0: a hybrid tool for morphological disambiguation. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013). pp. 539–545. Incoma Ltd. Shoumen, Bulgaria, Hissar, Bulgaria (2013)

Seeker, W., Farkas, R., Bohnet, B., Schmid, H., Kuhn, J.: Data-driven dependency parsing with empty heads. In: Proceedings of COLING 2012: Posters. pp. 1081–1090. The COLING 2012 Organizing Committee, Mumbai, India (Dec 2012), https://www.aclweb.org/anthology/C12-2105

Simkó, K.I., Vincze, V.: Hungarian copula constructions in dependency syntax and parsing. In: Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017). pp. 240–247. Linköping University Electronic Press, Pisa,Italy (Sep 2017), https://www.aclweb.org/anthology/W17-6527

Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1. pp. 173–180. NAACL '03, Association for Computational Linguistics, Stroudsburg, PA, USA (2003), http://dx.doi.org/10.3115/1073445.1073478

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 30, pp. 5998–6008. Curran Associates, Inc. (2017)

Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian Dependency Treebank. In: Proceedings of LREC 2010. ELRA, Valletta, Malta (May 2010)

Váradi, T., Simon, E., Sass, B., Mittelholcz, I., Novák, A., Indig, B., Farkas, R., Vincze, V.: E-magyar – A Digital Language Processing System. In: chair), N.C.C., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T. (szerk.) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 7-12, 2018 2018)