Subecz Zoltán, Nagyné Csák Éva

Szemantikus szerepek címkézése természetes szövegekben

Zoltán Subecz – Éva Nagyné Csák: Semantic role labeling in natural texts

Abstract
The event extraction and semantic role labeling are important areas in information extraction from natural language texts. Semantic role labeling, sometimes also called shallow semantic parsing, is a task in natural language processing consisting of the detection of the semantic arguments associated with the predicate or verb of a sentence and their classification into their specific roles. We created a program for this task in Java language, in which we applied data mining and artificial intelligence algorithms. We introduce in this article the principles and results of this application.
Keywords: Event detection, Information extraction, Data Mining, Text mining, Programming

ÖSSZEFOGLALÓ

A természetes szövegekből történő információkinyerés egyik fontos részterülete az események detektálása és a megtalált eseményeknél a szemantikus szerepek címkézése. A szemantikus szerepek címkézése az események szemantikus kapcsolatainak, vagy szemantikus szerepeinek detektálását és osztályozását jelenti. Ehhez a feladathoz Java nyelven készítettünk programot, amiben adatbányászati és mesterséges intelligencia algoritmusokat használtunk fel. Programunk egy adott eseményhez megkeresi a hozzá tartozó szemantikus szerepeket. A cikkben bemutatjuk a program készítési elveit, lépéseit és eredményeit.
Kulcsszavak: Eseménydetektálás, Információ-kinyerés, Adatbányászat, Szövegbányászat, Programozás

BEVEZETÉS

A természetes szövegekből történő *információkinyerés* alatt (Information extraction) a szövegbányászati feladatok egy speciális esetét értjük, ahol a cél az adott feladat szempontjából fontos szövegrészek (információk, tények) kigyűjtése a dokumentumokból, azaz strukturálatlan szövegekből *strukturált információ* előállítása. Az Információkinyerés egyik fontos feladata a névelemek felismerése mellett az *események detektálása*. A szövegekben lévő események felismerése, analizálása, és hogy hogyan viszonyulnak egymáshoz időben, fontos a szöveg tartalmának megismerésében. Egy esemény-előfordulás olyan kifejezés, ami egy cselekvésre, vagy történésre utal.
Az események detektálása mellett fontos azok szemantikus kapcsolatainak, vagy szemantikus szerepeinek megtalálása is (*szemantikus szerepek címkézése*). Ez a szemantikus kapcsolatok azonosítását jelenti egy szemantikus kereten belül (semantic frame). A keretek eseményeket írnak le azok szereplőinek szintaktikai és szemantikai megkötésein keresztül.

A szemantikus szerepek címkézéséhez Java nyelven készítettünk programot, amiben adatbányászati és mesterséges intelligencia algoritmusokat használtunk fel. Programunk egy adott eseményhez megkeresi a hozzá tartozó szemantikus szerepeket. Alkalmazásunk bemenetére érkezik a mondat, a célszó (target word: az adott esemény), és az adott keret. Az

alkalmazás az adott célszóhoz bejelöli a hozzá tartozó szemantikus kapcsolatokat.

1. SZEMANTIKUS KERETEK ÉS A SZEMANTIKUS SZEREPEK

Sok információ kinyerő rendszer manapság *tárgykör (domain)* specifikus *keretekkel* dolgozik. Egy-egy tárgykör eseményeit célszerű egy *kereten* belül vizsgálni, hiszen ugyanazok a *szerepek* tartoznak minden eseményhez, ami egy adott csoporthoz tartozik. Például egy repülőjegy foglalásokat feldolgozó rendszer a következő *szerepeket* használhatja: indulási időpont, érkezési időpont, célállomás, indulási állomás, távolság, ár. Egy repülőjegy foglalásokkal kapcsolatos *keret célszavai* lehetnek például: foglal, lefoglal, előjegyez, vált. A célszavak nem csak igék, hanem főnevek és igenevek is lehetnek a mondatokon belül. Például: foglalás, foglalni, foglaló, foglalva.
Munkánkban a *vállalati vásárlások, tulajdonváltozások keretével* foglalkoztunk és csak *igei és főnévi igenévi* célszavakhoz kerestük ki a szereplőket. A következő igei célszavakat vizsgáltuk meg az adott kereten belül: *vesz, vásárol, szerez, bekebelez, gyarapít, ad, átruház, értékesít, forgalmaz*. Valamint e célszavak minden igekötős, módbeli és időbeli változatát is.
A célszavakhoz a mondatokon belül a következő szerepeket kerestük meg: *vevő, eladó, áru, ár, idő*.
Néhány *példa* a szerepekre a vállalati vásárlások tárgykörben. A példákban vastag betűvel vannak kiemelve a célszavak és szögletes zárójelben a szerepek találhatóak. Alsóindexben szerepel az adott szerep típusa.

1. *[A svéd Electrolux]$_{Eladó}$ **eladja** [motorgyártó részlegét]$_{Árú}$ [az olasz Appliance Components Companies részvénytársaságnak]$_{Vevő}$ - tájékoztatott az Electrolux.*
2. *[A Deutsche Börse AG]$_{Vevő}$ pénteken bejelentette, hogy teljesen **megveszi** [a luxemburgi Clearstream elszámoló-házat]$_{Árú}$.*
3. *[A Royal Dutch Shell csoport]$_{Vevő}$ [400 millió dollárért]$_{Ár}$ **megvenni** készül [a legnagyobb kínai offshore-földgáz- és olajmező 20 százalékát]$_{Árú}$.*
4. *[A svéd Ericsson]$_{Eladó}$ bejelentette, hogy [a német Infineonnak]$_{Vevő}$ **adja** el [chipgyártó részlegét]$_{Árú}$, [400 millió euróért]$_{Ár}$.*
5. *[A többnyire szárazföldi szállítással foglalkozó magyar tulajdonban lévő Cronus Kft.]$_{Vevő}$ [a közelmúltban]$_{Idő}$ **megvásárolta** [a Magyar Államvasutak Rt.-től]$_{Eladó}$ [a debreceni székhelyű MÁV Hajdú Vasútépítő-mélyépítő Kft.-t]$_{Árú}$ - jelentette be szerdán Debrecenben a Cronus Kft. tulajdonosa.*

A példákban látszik, hogy egy szerep általában *több szóból* áll és a mondatok általában nem tartalmazzák mind az öt szerepet.

2.1 Felhasznált Korpusz

Az alkalmazásunk teszteléséhez a Szeged Korpusznak [Csendes,2005:409-412] a gazdasági rövidhírek csoportjának egy olyan változatát használtuk fel, amelyikben annotálva vannak a vállalati vásárlásokra a szemantikus szerepek. Az annotált korpuszban a tanításhoz be vannak jelölve a szemantikus szerepek az eseményekhez. Ezek közül 670 mondatot használtunk fel. A tanításhoz és kiértékeléshez 10-szeres keresztvalidációt alkalmaztunk: a kiinduló mondathalmazt 10 egyenlő részre osztottunk fel véletlen kiválasztással, és minden lépésnél a kiválasztott 10%-os rész volt a kiértékelő, a maradék 90% pedig a tanító halmaz.

2.2 Statisztikai adatok

Mondatok száma összesen: 670 db
Mondatok száma a Tanító korpuszon: 603 db
Mondatok száma a Kiértékelő korpuszon: 67 db

Azon mondatok száma, amelyek tartalmazzák az adott szerepet:
Vevő: 518 db Eladó: 379 db Áru: 649 db
Ár: 204 db Idő: 194 db

2.3 Felhasznált programcsomagok

A feladatokat *bináris osztályozásra* vezettük vissza. Az osztályozáshoz a *Weka*

programcsomagnak (http://www.cs.waikato.ac.nz/ml/weka/) (Data Mining Software in Java) a J48-as döntési fa elemzőjét használtuk fel. A Weka adatbányászati feladatokhoz készített gépi tanuló algoritmusok gyűjteménye.

A feladathoz felhasználtuk még a Magyarlanc 2.0 programcsomagot is. [Zsibrita,2013:368–374] A csomag magyar szövegek mondatra és szavakra bontására, a szavak morfológiai elemzésére, majd szófaji egyértelműsítésére, és mondatok függőségi nyelvtan szerinti szintaktikai elemzésére alkalmazható.

2.4 A Magyarlanc elemzésének bemutatása

A Magyarlanc programcsomag a bemenetére érkező mondatoknak elkészíti az előző pontban leírt elemzését. A mondat minden szavához külön sorba elkészíti az elemzést (1. és 4. ábra). Minden szóról megadja a következő információkat: *sorszám, szó, lemma, szófaj, morfológiai kódok.* A sor végén megadja, hogy az adott szó melyik szóval van *szintaktikai kapcsolatban*, és hogy milyen a *kapcsolat típusa*. A szintaktikai kapcsolatok alapján a mondatok egy *függőségi fát* alkotnak.

Az elemzések után megjelenítettük *vizuális elemzővel* a szintaktikai kapcsolatok alapján a mondat függőségi fáját a program *online elemzőjével* (http://www.inf.u-szeged.hu/rgai/magyarlanc-service/) (2. és 5. ábra). Az elemzés és a vizuális ábrázolás egymásnak megfelelően megadja a szavak közötti szintaktikai kapcsolatokat. Az ábrán a nyilak címkéje jelzi a kapcsolatokat.

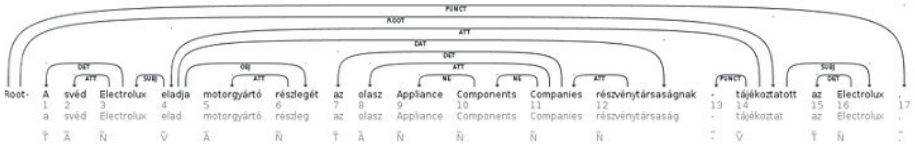
2.4.1 Az 1. fejezet első két mondatának elemzése.

1. [A svéd Electrolux]$_{Eladó}$ ***eladja*** *[motorgyártó részlegét]$_{Árú}$ [az olasz Appliance Components Companies részvénytársaságnak]$_{Vevő}$ - tájékoztatott az Electrolux.*

```
1       A       a       T       SubPOS=f                3       DET
2       svéd    svéd    A       SubPOS=f|Deg=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none       3       ATT
3       Electrolux      Electrolux      N       SubPOS=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none     4       SUBJ
4       eladja  elad    V       SubPOS=m|Mood=i|Tense=p|Per=3|Num=s|Def=y       14      ATT
5       motorgyártó     motorgyártó     A       SubPOS=f|Deg=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none       6       ATT
6       részlegét       részleg N       SubPOS=c|Num=s|Cas=á|NumP=s|PerP=3|NumPd=none   4       OBJ
7       az      az      T       SubPOS=f                11      DET
8       olasz   olasz   A       SubPOS=f|Deg=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none       11      ATT
9       Appliance       Appliance       N       SubPOS=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none     10      NE
10      Components      Components      N       SubPOS=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none     11      NE
11      Companies       Companies       N       SubPOS=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none     12      ATT
12      részvénytársaságnak     részvénytársaság        N       SubPOS=c|Num=s|Cas=d|NumP=none|PerP=none|NumPd=none     4       DAT
13      -       -       -       -       _       14      PUNCT
14      tájékoztatott   tájékoztat      V       SubPOS=m|Mood=i|Tense=s|Per=3|Num=s|Def=n       0       ROOT
15      az      az      T       SubPOS=f                16      DET
16      Electrolux      Electrolux      N       SubPOS=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none     14      SUBJ
17      .       .       .       _       0       PUNCT
```

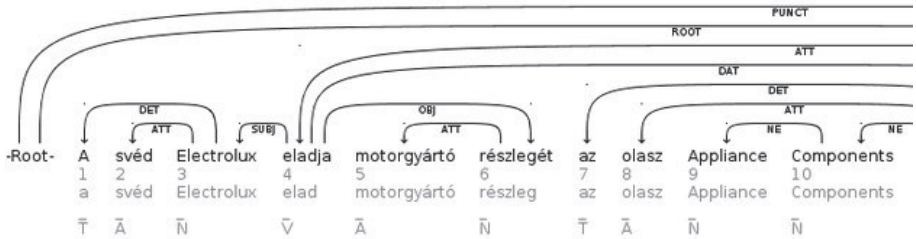
1. ábra: Az első mondat szöveges elemzése

Forrás: saját elemzés



2. ábra: Az első mondat vizuális elemzése

Forrás: saját elemzés



3. ábra: A vizuális elemzés első fele kiemelve

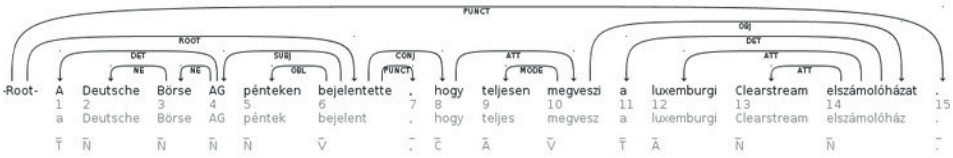
Forrás: saját elemzés

*2. [A Deutsche Börse AG]$_{Vevő}$ pénteken bejelentette, hogy teljesen **megveszi** [a luxemburgi Clearstream elszámolóházat]$_{Árú}$.*

1	A	a	T	SubPOS=f	4	DET			
2	Deutsche		Deutsche	N	SubPOS=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none	3	NE		
3	Börse	Börse	N	SubPOS=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none	4	NE			
4	AG	AG	N	SubPOS=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none	6	SUBJ			
5	pénteken	péntek	N	SubPOS=c|Num=s|Cas=p|NumP=none|PerP=none|NumPd=none	6	OBL			
6	bejelentette	bejelent	V	SubPOS=m|Mood=i|Tense=s|Per=3|Num=s|Def=y	0	ROOT			
7	,	,	,	6	PUNCT				
8	hogy	hogy	C	SubPOS=s|Form=s|Coord=p	6	CONJ			
9	teljesen	teljes	A	SubPOS=f|Deg=p|Num=s|Cas=w|NumP=none|PerP=none|NumPd=none	10	MODE			
10	megveszi	megvesz	V	SubPOS=m|Mood=i|Tense=p|Per=3|Num=s|Def=y	8	ATT			
11	a	a	T	SubPOS=f	14	DET			
12	luxemburgi	luxemburgi	A	SubPOS=f|Deg=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none	14	ATT			
13	Clearstream	Clearstream	N	SubPOS=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none	14	ATT			
14	elszámolóházat	elszámolóház	N	SubPOS=c|Num=s|Cas=a|NumP=none|PerP=none|NumPd=none	10	OBJ			
15	.	.	.	_	0	PUNCT			

4. ábra: A második mondat szöveges elemzése

Forrás: saját elemzés



5. ábra: A második mondat vizuális elemzése

Forrás: saját elemzés

Az elemzésekből látszik, hogy a függőségi elemző egy *szabályos elemző fát* készít kimenetként. (A vizuális elemző a fát 90º-al elforgatva ábrázolja a szavak egymás utáni sorrendjének megfelelően.) A fa legfelső eleme a *Root*. A fa *csomópontjaiban* vannak a mondat szavai, az *ágak* a szavak közötti *szintaktikai kapcsolatokat* reprezentálják. A fában kiemelt szerepe van az *igéknek*. A *főige* (a példákban az *eladja* és *bejelentette*) általában a Root alatt helyezkedik el, a szintaktikai kapcsolatokon keresztül ehhez kapcsolódnak a többi elemek.

Ha a szerep *több szóból* áll, akkor ezek a szavak egy *részfát* alkotnak a mondat fáján belül. A részfa a *kiemelt szaván* (fejszó, headword) keresztül kapcsolódik a fa többi részéhez.

Van, amikor a szerep kiemelt szava (headword) a célszóhoz kapcsolódik közvetlenül. Mint például az első mondatban az *eladó* szerep *Elektrolux* szava szintaktikai kapcsolatban van az *eladja* célszóval(alany, SUBJ). Ilyen esetben könnyebb megtalálni a szerepet. Van, amikor a szerep kiemelt szava nem a célszóhoz kapcsolódik közvetlenül. Például a második példamondatnál a *vevő* kiemelt szava nem kapcsolódik szintaktikailag a *megveszi* célszóhoz az elemzőfában, hanem a *bejelentette* igén keresztül. Ilyenkor nehezebb megtalálni a szerepet. Minél távolabb van a szerep a célszótól a mondaton vagy az elemzőfán belül, annál kisebb a valószínűsége a szerep azonosításának.

Bár a Magyarlanc program elkészíti a mondatoknak a szintaktikai elemzését, de a példákon is láttuk, hogy a szintaktikai kapcsolat típusából nem következik a szemantikai szerep. Például a *vesz* célszónak az alanya általában a vevő, az *elad* célszónak az alanya általában az eladó. Így a szintaktikai kapcsolatok mellett több más tulajdonságot is meg kell figyelni a mondatban. A feladatot megnehezíti, hogy a Magyarlanc elemző is hibával dolgozik, így ez a hiba és a hibákból eredő hamis döntések megjelennek a mi eredményeinkben is.

3. AZ OSZTÁLYOZÁS BEMUTATÁSA

Az osztályozáshoz *bináris osztályozót* használtunk. A célszavakra a következő szerepeket vizsgáltuk: *vevő, eladó, áru, ár, idő*. Minden bemeneti mondatnál adott volt a célszó, és hogy a mondat tartalmazza-e az adott szerepet. Az osztályozóknál a jelöltek a függőségi elemzőfa csomópontjai voltak. Egy mondat annyi csomópontot tartalmaz, ahány szóból áll (plusz a Root) és általában egy csomópont ezek közül a keresett szerep kiemelt

szava(head word). Az osztályozásnál ezek az *igaz (true)* esetek. A többi csomópont pedig a *hamis (false)* eset.

A tanító és a kiértékelő halmazon a jelöltekhez *jellemzőket* vettünk fel. A tanítóhalmazon a jelöltek jellemzői és a jelölt *true* vagy *false* tulajdonsága alapján az osztályozó szabályokat készít (osztályozó tanítása), amiket majd a kiértékelő halmaz jelöltjeire alkalmaz (a tanított osztályozó alkalmazása a döntésekben). Az osztályozó a megtanult összefüggéseket a kiértékelő halmaz ismeretlen mondataira alkalmazza.

3.1 A Jellemzőtér kiválasztása

Olyan *jellemzőket* kerestünk, amelyek segíthetnek az adott szerep megtalálásában. A jellemzőknél felhasználtuk a *függőségi elemzőfát* is, a jelölt és a célszó viszonyát a függőségi fában, mert ez gyakran egy fontos tulajdonsága az adott szerepnek.

A következő jellemzőket választottuk ki:

- **Felszíni jellemzők**: *Bigramok, trigramok*: A vizsgált szavak végén lévő 2-es, 3-as betűcsoportok.
  Pozíció: a jelölt a célszó előtt vagy után áll a mondatban.
  Távolság-mondatban: a jelölt és a célszó szótávolsága a mondaton belül.
- **Morfológiai jellemző**: *Szófaj*, *Lemma*: a jelölt és a célszó szófaja és lemmája.
- **Szintaktikai jellemzők**: *Szintaktikai kapcsolat*: Ha a jelölt és a célszó közvetlen szintaktikai kapcsolatban áll egymással, akkor a kapcsolat típusát megadtuk. (-1 ha nincs kapcsolat)
- **Jellemzők az elemzőfa alapján**: Itt a jelölt és a célszó viszonyát vizsgáltuk a függőségi elemzőfában. A jelölt és a célszó is egy-egy csomópont az elemzőfában. *Jelölt-célszó-távolság-elemzőfában*: A jelölt és a célszó csomópontjai közötti csomópontok száma az elemzőfában. *Lemma-útvonal*: Egymás után írtuk a jelölt és a célszó csomópontok közötti csomópontok lemmáját, feljegyezve azt is, hogy az elemzőfában felfelé, vagy lefelé haladtunk az adott kapcsolatnál. Például: Budapesti↑Értéktőzsde↑honlap↑közöl ↓megvásárol. *Szófaj-útvonal*: Mint az előző, de itt nem a szó lemmáit, hanem a szófajok útvonalát vettük fel. Például: C↑S↑V↑C↑V↑V↓V↓N↓N↓A. (V:ige, N:főnév, stb.) *Szintaktikai-kapcsolat-útvonal*: Az előzőkhöz hasonlóan itt azt vettük fel, hogy a jelölt és a célszó között az elemzőfában milyen szintaktikai kapcsolatokon keresztül tudunk eljutni. Például:
  ↑COORD*SUBJ↓ATT↓INF↓OBJ↓ATT.
  (SUBJ: alany, OBJ: tárgy, stb.). *Uralkodó-kategória-szófaja*: A jelölt és a célszó közötti útvonalon megkerestük a legmagasabban fekvő csomópontot, és feljegyeztük a hozzá tartozó szó szófaját. *Jelölt-alatti-részfában-van-e-névelem*: Névelemek azok a szavak, vagy szósorozatok, amelyek a világ valamely egyedére egyedi módon hivatkoznak. Például személyek, intézmények, földrajzi helyek nevei. A Magyarlanc program az elemzésében jelöli, ha talált névelemeket a mondatban. Mivel a vállalati tulajdonváltozások témakörében gyakran találkozunk vállalati névelemekkel, ezért felvettük, hogy a jelölt, vagy az alatta levő részfa tartalmaz-e névelemet? *Jelölt-alatti-részfában-névelem-távolság*: az előzőhöz hasonlóan megadtuk a részfában azt a mélységet, ahol először találtunk névelemet. (-1: ha nincs alatta névelem)

3.2 Statisztikai arány felhasználása az osztályozásnál

Lehetett volna egyszerűen az előző pontban megadott jellemzőket felhasználni az osztályozásnál, ezeket adni az osztályozó bemenetére. E helyett a következő módszert választottuk - amellyel *jelentősen csökkentettük az osztályozó vektorterének méretét és így a futási időt* is. A tanító halmazon a jelölteket és azokon belül minden jellemzőt végignézve minden jellemző egyedhez feljegyeztük, hogy az adott egyed hány alkalommal fordult elő és

ebből hányszor volt igaz (*true*) eset. A két szám alapján kiszámítottuk a hozzá tartozó *előfordulási arányt*.
Például a *Jelölt-lemma* jellemzőnél a *jelölt-lemma=Corp*. eset 11-szer fordult elő és ebből 7-szer volt igaz eset, (4-szer pedig hamis). Így hozzá a 0,64-es előfordulási arány tartozott. Az osztályozónak ezeket az arányokat adtuk meg. Az előző példához *Jelölt-lemma-arány=0,64*.

Az osztályozó a tanítás során *szabályokat készített* azokhoz a jellemző-arányokhoz, amelyeket fontosnak talált annak eldöntéséhez, hogy a jelölt a keresett szerep, vagy sem. A szabályokat *döntési fában* adta meg. Szemléltetésül az egyik *döntési fa egy kis részlete*:

```
LemmaParseTreePathArany <= 0
|  JeloltLemmaArany <= 0.07692
|  |  PosKodParseTreePathArany <= 0.14286
|  |  |  JeloltVegenTrigramArany <= 0.19457
|  |  |  |  JeloltLemmaArany <= 0: semmi (16227.0/3.0)
|  |  |  |  JeloltLemmaArany > 0
|  |  |  |  |  SzintaktikaiKapcsolatArany <= 0.02857: semmi (152.0)
|  |  |  |  |  SzintaktikaiKapcsolatArany > 0.02857
|  |  |  |  |  |  JeloltVegenBigramArany <= 0: semmi (2.0)
|  |  |  |  |  |  JeloltVegenBigramArany > 0: valami (6.0/1.0)
|  |  |  JeloltVegenTrigramArany > 0.19457
|  |  |  |  EdgeTypeParseTreePathArany <= 0.07143: semmi (19.0)
|  |  |  |  EdgeTypeParseTreePathArany > 0.07143: valami (7.0/1.0)
|  |  PosKodParseTreePathArany > 0.14286
|  |  |  JeloltVegenBigramArany <= 0.02885: semmi (23.0)
……..
```

Az osztályozás minden esetéhez egy ilyen *döntési fát* készített el.
A *kiértékelő halmazon* hasonlóan jártunk el. Minden jelöltet és azon belül minden jellemzőt végignézve a tanítóhalmaznál elkészített előfordulási arányokból megkerestük, *hogy az adott jellemző előfordulás milyen aránnyal szerepelt a tanító halmazon* (-1 ha nem fordult elő). A kiértékelőnek ezeket az arányokat adtuk meg. Az osztályozó a tanító halmaz alapján elkészített szabályokat felhasználva *osztályozta a kiértékelő-halmaz jelöltjeit*.

4. EREDMÉNYEK

A méréseket *két fő esetre* bontottuk. Az első esetben a célszavakat egy csoportként kezeltük. Második esetben a célszavakat két csoportra bontottuk. A *vevő* és az *eladó* szerepek viselkedését meghatározza, hogy az adott célszónál az alany általában vevő vagy eladó. Ezért a következő szabályt alkalmaztuk a csoportosításra:

A *vevő-centrikus* csoportba azok a szavak kerültek, amelyeknél az alany általában a *vevő*: vesz, vásárol, szerez, bekebelez, gyarapít. Az *eladó-centrikus* csoportba pedig azok, amelyiknél az alany általában az *eladó*: ad, átruház, értékesít, forgalmaz. Ez a felosztás segítette a *vevő* és az *eladó* szerepek megtalálását. Minden esethez külön bináris osztályozót készítettünk.

A kiértékelésnél a megszokott mértékeket határoztuk meg:
TP – True Positive, *TN* – True Negative, *FP* – False Positive, FN – False Negative
Ezekből az ismert metrikákat számítottuk ki:
Pontosság (Precision): P=TP/(TP+FP); *Fedés* (Recall): R=TP/(TP+FN);

Az F-mértéket a pontosság és a fedés harmonikus közepe adja, így egy mérőszámmal jól jellemezhető a rendszer hatékonysága:
F-mérték: F=2*P*R/(P+R)

Vevő-centrikus				
Szerepnév	**Esetek száma**	**Pontosság**	**Fedés**	**F-mérték**
Vevő	345	72,59	55,21	**62,61**
Eladó	148	79,61	51,07	**61,34**
Áru	373	76,31	73,13	**74,50**
Ár	149	92,35	80,51	**85,95**
Idő	121	81,16	64,37	**71,44**

1. táblázat: Eredmények a vevő-centrikus célszavakra (%)

Forrás: saját elemzés

Eladó-centrikus				
Szerepnév	**Esetek száma**	**Pontosság**	**Fedés**	**F-mérték**
Vevő	173	71,11	57,43	**63,18**
Eladó	231	64,90	47,64	**54,29**
Áru	276	78,07	75,21	**76,25**
Ár	55	86,62	79,83	**82,33**
Idő	73	69,55	35,66	**45,08**

2. táblázat: Eredmények az eladó-centrikus célszavakra (%)

Forrás: saját elemzés

Szerepnév	**Esetek száma**	**Pontosság**	**Fedés**	**F-mérték**
Vevő	518	70,97	53,78	**61,05**
Eladó	379	70,46	42,27	**52,35**
Áru	649	78,75	77,04	**77,75**
Ár	204	89,29	80,70	**84,52**
Idő	194	78,05	57,78	**65,86**

3. táblázat: Eredmények a célszavak csoportosítása nélkül (%)

Forrás: saját elemzés

4.1 Az eredmények kiértékelése

A mérések eredményei az *1,2,3-as táblázatban* láthatóak.

A vizsgált mondatokban sokkal több mondatban volt a célszó vevő-centrikus, mint eladó-centrikus.

Legjobb eredményeket az *ár* szerep megtalálásánál ért el a modell (mindhárom esetben 80 fölötti F-mérték). *Második legjobb eredményt* az *áru* szerepnél értünk el (mindhárom esetben 70 fölötti F-mérték). *Leggyengébb eredményt az idő szerepnél* ért el a modell az eladó-centrikus célszavaknál (45,08). Ennek magyarázata lehet a vizsgált esetek kis száma ezen a területen (73). Az esetek kis száma miatt valószínű a modell túltanulta magát a tanító adatokon (over-fitting), és e szabályok alkalmazása nem vezetett jó eredményre a kiértékelő halmazon. Ezt támasztja alá, hogy ahol már jóval több tanító adat volt az idő szerepnél: vevő-centrikus eset (121) és csoportosítás nélkül (194), a modell sokkal jobban teljesített (71,44, 65,86). Az *eladó-centrikus esetnél az ár szerep* is kevés vizsgált esetet tartalmazott (55), de az eredményen látszik (82,33), hogy a tanító korpusz alapján kialakított szabályokat az osztályozó eredményesen tudta alkalmazni a kiértékelő korpuszon is.

A *célszavaknak a két csoportra bontása* a vevő és az eladó szerepek megtalálásánál is segített. A másik három szerepnél nem jelentett minden esetben segítséget a két csoportra bontás. Ezeknél egyes esetekben csoportosítás nélkül

jobb eredményeket kaptunk, mint csoportosítással. Ennek oka lehet, hogy a csoportosítás nélküli esetben az esetek száma jóval nagyobb, mint a csoportosított esetekben külön-külön, ami segítette az általánosabb, az ismeretlen mondatokon is jól alkalmazható összefüggések megtalálását.

A kapott eredményeket jónak értékeljük annak ismeretében, hogy a gyakran hosszú mondatokban, sok esetben a szeresett szerepek távolabb távol helyezkedtek el a célszótól a mondatban és az elemzőfában is.

Az eredmények jelentősen javultak volna, ha az elő-feldolgozást, szintaktikai elemzést nem programmal készíttetjük el (Magyarlanc), hanem emberi annotálással. De ez a módszer sokkal költségesebb és időigényesebb lett volna, ami miatt ezt csak ritkán lehet alkalmazni.

ÖSSZEGZÉS

Munkánkban szemantikus szerepek címkézésével foglalkoztunk. A szemantikus szerepek címkézése az események szemantikus kapcsolatainak, vagy szemantikus szerepeinek detektálását és osztályozását jelenti. Ehhez a feladathoz Java nyelven készítettünk programot, amiben adatbányászati és mesterséges intelligencia algoritmusokat használtunk fel. Programunk egy adott eseményhez megkereste a hozzá tartozó szemantikus szerepeket. A cikkben bemutattuk a program készítési elveit, lépéseit és eredményeit. A kapott eredményeket jónak értékeljük.

IRODALOMJEGYZÉK

[1] Csendes Dóra, Alexin Zoltán, Csirik János, Kocsor András (2005): A Szeged Korpusz és Treebank verzióinak története. In: Alexin Zoltán, Csendes Dóra (szerk.): III. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2005). Szeged, Szegedi Tudományegyetem, pp. 409-412.

[2] Ehmann Bea, Lendvai Piroska, Mihăltz Márton, Vincze Orsolya, László János (2013): Szemantikus szerepek a narratív kategoriális elemzés (NARRCAT) rendszerében. IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, p. 121-123

[3] Farkas Richárd, Konczer Kinga, Szarvas György (2004): Szemantikus keret illesztés és az IE-rendszer automatikus kiértékelése. In: Alexin Zoltán, Csendes Dóra (szerk.): II. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY). Szeged, Szegedi Tudományegyetem, pp. 49-53.

[4] D. Gildea, D. Jurafsky (2002): Automatic labeling of semantic roles. Computational linguistics, Vol. 28, No. 3, Pages 245-288

[5] L. Màrquez X., C. Kenneth, C. Litkowski, S. Stevenson: Semantic Role Labeling (2008): An Introduction to the Special Issue. Computational linguistics, 2008, Vol. 34, No. 2, Pages 145-159

[6] Zsibrita, J., Vincze, V., Farkas,R. (2013): Magyarlanc 2.0: szintaktikai elemzés és felgyorsított szófaji egyértelműsítés. In: Tanács A. és Vincze V. (szerk.), IX. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged 368–374