Igei események detektálása és osztályozása magyar nyelvű szövegekben

Subecz Zoltán, Nagyné Csák Éva

Szolnoki Főiskola
5000 Szolnok, Tiszaligeti sétány 14.
{subecz,csak}@szolf.hu

Kivonat: Jelen tanulmányunkban bemutatjuk megközelítésünket, amely igei eseményeket képes detektálni és osztályozni magyar szövegeken. Első lépésben azonosítottuk a többszavas főnévi+igei kifejezéseket. Majd detektáltuk az eseményeket, a detektált eseményeket pedig osztályokba soroltuk. A feladatok mindegyikéhez gazdag jellemzőkészleten alapuló bináris osztályozót használtunk. Az osztályozót kiegészítettük szabályalapú módszerekkel is. Módszerünket a Szeged Korpusz öt különböző doménjén is megvizsgáltuk, és hasonlósági gráfok segítségével elemeztük a részkorpuszok kapcsolatát.

1 Bevezetés

Munkánkban természetes szövegekben előforduló **események detektálásával és osztályozásával** foglalkoztunk. Az események **detektálásának** a feladata az esemény-előfordulások azonosítása a szövegekben, az **osztályozással** pedig a megtalált eseményeket előre meghatározott kategóriákba rendeljük. Esemény-előfordulásnak tekintünk minden olyan kifejezést, ami olyan eseményt vagy állapotot jelöl, amit egy adott időponthoz, vagy intervallumhoz tudunk kapcsolni.

Noha az igéken kívül lehetnek események más szófajú szavak is (pl. főnevek, igenevek stb.), a szövegekben a *legtöbb esemény igékhez kapcsolódik*, ezért jelen munkánkban az **igei eseményekkel** foglalkoztunk. Az igék közül azonban *nem mindegyik* tekinthető eseményjelölőnek (például: van, volt, lesz, marad, segédigék), így ezek kiszűrésére külön figyelmet kell fordítani. Vannak olyan események, amelyeket két szóval fejezünk ki (pl. döntést hoz), ezek szintén külön kezelést igényelnek. Több munka is foglalkozott már részletesen a *többszavas igei kifejezésekkel* [8, 6, 7], ezek eredményeit felhasználtuk.

A feladat a szövegekben megtalálható események detektálása és osztályozása. Munkánkban elsősorban az **igei egy- és többszavas eseményekkel** foglalkozunk. A rendszer bemenete egy tokenszinten címkézett tanító korpusz. A feladatot három részre osztottuk. A szövegekben először a **több szavas főnév + igei kifejezéseket** válogattuk ki, majd a maradék igékből **detektáltuk az eseményeket**. A megtalált eseményeket ez után **osztályoztuk**. A feladat megoldásához *statisztikai és szabály alapú módszereket* is alkalmaztunk.

2 Kapcsolódó munkák

Sok kutatás foglalkozik az események detektálásával. A legtöbb munkában csak adott eseményekkel foglalkoznak (például üzleti), vagy még azon belül is csak kiemelt eseményekkel (például cégfelvásárlás). Jelen munkánkban **minden igei esemény** detektálásával és osztályozásával foglalkoztunk. Néhány kutatás foglalkozott angol nyelvre igei események detektálásával és osztályozásával. A legtöbb munkában az igei mellett más szófajhoz tartozó eseményeket is megvizsgáltak.

Bethard [1] statisztikai jellemzők alapján detektált eseményeket. Figyelembe vett többszavas kifejezéseket is. A következő jellemzőcsoportokat használta fel az osztályozónál: az adott szó, trigramok a szó elején, végén, morfológiai jellemzők, szófaj, szintaktikai jellemzők, időbeliség kifejezése, tagadási jellemző, WordNet hiperním jellemző. Nem csak a vizsgált szóra, hanem a környező néhány szóra is kigyűjtötték ezeket a jellemzőket. Detektálásra a modell 88,3-os F-mértéket ért el, osztályozásra 70,7-ot.

Llorens és társai [3] CRF modellt alkalmazott szemantikai szabályok felismerésével események detektálásához és osztályozásához. Morfológiai, szintaktikai, szemantikai jellemzőket használtak fel az osztályozáshoz. Egyes jellemzőket nem csak az adott szóhoz, hanem néhány szavas környezetükhöz is kigyűjtöttek. Detektálásra a modell 91,33-os F-mértéket ért el, osztályozásra 73,51-ot.

Marsic [4] csak igei eseményekkel foglalkozott, azok detektálásával és osztályozásával. Statisztikai módszereket használt a feladathoz. Morfológiai és szintaktikai jellemzőket használt fel. Detektálásra a modell 86,49-os F-mértéket ért el.

Bittar [2] francia nyelvű szövegekhez végzett eseménydetektálást. Detektálásra a modell 88,8-os F-mértéket ért el.

Az általunk megvalósított megközelítés *gépi tanuló módszer* alapján detektálja és osztályozza az eseményeket, amit *szabály alapú módszerrel* is kiegészítettünk. A feladathoz *gazdag jellemzőteret* használtunk fel. A detektálás előtt kinyertük a *többszavas kifejezéseket*. A **detektálásnál** 93,85-os, két **osztályozásnál** pedig 85,93 és 66,06-os F-mértéket értünk el.

3 A Korpusz, programok

Alkalmazásunkban a **Szeged Korpusz** egy olyan változatát használtuk fel, amelyikben annotálva vannak a többszavas kifejezések [8]. A korpusznak egy részét használtuk fel, ami *5010 mondatot* tartalmaz a következő területekről: *üzleti rövidhírek, szépirodalom, jogi szövegek, újsághírek, fogalmazás*. Tanításhoz véletlenszerűen kiválasztottuk a korpusz 90%-át, kiértékelésre pedig a maradék 10%-ot.

A detektálásához is ezt az 5010 mondatot használtuk fel. A mondatokat nyelvész segítségével *annotáltuk* a detektáláshoz és az osztályozáshoz is. Az annotátorok közötti egyetértés a detektálásnál 87%-os volt, az osztályozásnál 81%.

Az osztályozáshoz a *Weka*[1] *programcsomagnak* a C4.5 döntési fa algoritmust implementáló J48 tanuló algoritmust alkalmaztuk. A magyar nyelvű szövegek feldolgozásához a *Magyarlanc 2.0* [9] csomagot használtuk.

4 Többszavas kifejezések detektálása

A faladat első részeként detektáltuk a szövegekben a *többszavas kifejezéseket*. Az alkalmazásunkban felhasználtuk az [6]-os publikációban bemutatott alkalmazás elveit. Erről a modulról részletesen írtunk az [7]-es publikációban is, így itt most csak a lényegét foglaljuk össze.

Az 5010 mondatot tartalmazó korpuszunk 100291 db tokent, és ezen belül 542 többszavas kifejezést tartalmazott.

Az alkalmazásban a következő *alapjellemzőket* használtuk fel az osztályozásnál: felszíni jellemzők, lexikai jellemzők, morfológiai jellemzők, szintaktikai jellemzők. Az 5010 mondatos korpuszon az alkalmazásunk ezekkel a jellemzőkkel a következő eredményeket érte el: Pontosság=90,48 Fedés=41,30 *F-mérték=56,72*

A mérésünket még kiegészítettük két jellemzővel. Az első esetben *frekvenciainformációkat* vettünk fel. Minden főnév + ige többszavas kifejezéshez (lemma + abszolút lemma párhoz) meghatároztuk, hogy milyen arányban volt a tanító korpuszon esemény. A tanításnál és a kiértékelésnél felhasználtuk ezt az arányt is, mint jellemzőt. Ezzel a kiegészítéssel a következő eredményt értük el: Pontosság=96,43 Fedés=58,70 *F-mérték=72,97*. Ez a jellemző jelentősen javította az eredményt.

A másik esetben ezt még kiegészítettük a következő jellemzővel. Az alapjellemzők között a lexikai jellemzőknél volt egy *lista*, amiben tárolásra kerültek gyakori többszavas kifejezések [6]. Ezt a listát kiegészítettük a tanító korpuszból vett újabb többszavas kifejezésekkel. Az újabb jellemző akkor kapott igaz értéket, ha a többszavas kifejezés-jelölt szerepelt ebben a listában (szótárillesztés). Ezzel a kiegészítéssel a következő eredményt értük el: Pontosság=93,18 Fedés=89,13 *F-mérték=91,11*. Ez a jellemző is jelentősen javította az eredményt.

5 Igei események detektálása

Ebben a modulban az igei és főnévi igenévi eseményeket *detektáltuk*. A feladatot bináris *osztályozásra* vezettük vissza, amit *szabály alapú módszerrel* is kiegészítettünk. Ehhez a modulhoz külön osztályozót készítettünk. Az osztályozásnál eseményjelölteknek az igéket és a főnévi igeneveket válogattuk ki.

Az 5010 mondatunk 9445 igét tartalmazott, amiből 5487 volt eseményt jelölő.

[1] Weka [2013] Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka/

5.1 Jellemzőkészlet

Az eseményjelöltekhez a következő jellemzőket gyűjtöttük ki:

- **Felszíni jellemzők-1: Bigramok, trigramok, fourgramok**: A vizsgált szavak elején és végén lévő 2-es, 3-as, 4-es betűcsoportok. A jellemzők közé felvettük, hogy egy adott szó milyen betűcsoporttal kezdődik és végződik.
- **Felszíni jellemzők-1**: Szóhossz lemmahossz, valamint a szó sorszáma a mondaton belül.
- **Lexikai jellemzők**: Az adott szó létige, vagy segédige-e? Egy-egy listába kigyűjtöttük a létigéket és a segédigéket. Jellemzőként megadtuk, hogy az adott szó szerepel-e valamelyik listában.

Mivel egy szónak az eseményjellegét meghatározhatja az is hogy előtte, vagy utána áll-e létige vagy segédige, ezért ezt a négy bináris jellemzőt is felvettük.

- **Morfológiai jellemzők**: Mivel a magyar nyelv igen gazdag morfológiával rendelkezik, ezért számos morfológiaalapú jellemzőt definiáltunk. Jellemzőként definiáltuk az eseményjelöltek MSD-kódját felhasználva a következő morfológiai jegyeket: típus (SubPos), mód (Mood), eset (Cas), idő (Tense), személy (PerP), szám (Num), határozottság (Def).

Jellemzőként felvettük még az igekötőt és az adott szó, valamint az előtte és az utána álló szó szófaját.

- **Szintaktikai jellemzők**: Megadtuk, hogy az adott eseményjelölthöz milyen szintaktikai kapcsolattal tartoznak szavak. (például alany, tárgy, …). Kiemelt figyelmet szenteltünk ezek közül a PRED kapcsolatnak, mert nem esemény igéknél gyakran ilyen kapcsolata van az igének. Ezért ezt is definiáltuk a jellemzők között.
- **Szemantikai jellemzők**: Ehhez a Magyar WordNet-et [5] használtuk fel. Először olyan osztályozót készítettünk, amelyikbe jellemzőként felvettük, hogy a vizsgált szónak mik a *hipernimái*. A tanítással az osztályozó kiválogatta a döntési fába azokat a *synseteket*, amelyek alá jellemzően események tartoznak. Ezeket a kiválogatott synseteket használtuk fel a fő feladathoz. Egy listában felvettük ezeket, majd jellemzőként megadtuk, hogy az adott eseményjelölt szerepel-e valamelyik ilyen synset *hiponimái* között.

Ha csak a WordNet jellemzőt alkalmaztuk önállóan, akkor bár nem a legjobb, de 91,4-es F-mértéket értünk el.

A gépi tanuló módszerünket kiegészítettük szabály alapú módszerrel is. A jogi korpuszon sok olyan kifejezés volt, amelyekben az ige más szövegekben általában eseményt jelöl, de ebben a szövegkörnyezetben nem. Például: *A törvény **kimondja**, hogy…* Az okirat **meghatározza**, hogy… Ezekhez az esetekhez definiáltunk szabályokat. Például: ha alany=törvény és ige=kimondja, akkor kimondja ≠ esemény.

A kiértékelés során a pontosság (P), fedés (R) és F-mérték (F) metrikákat használtuk. Több fajta mérést is végeztünk. Megvizsgáltuk az alkalmazást az öt korpuszon együtt *tízszeres keresztvalidációval*, valamint külön-külön a *részkorpuszokon* is teszteltük a működését. Porlasztásos méréssel vizsgáltuk meg az egyes *jellemző csoportok*

jelentőségét az adott feladathoz. *Domainek közötti keresztméréseket* is alkalmaztunk, mely során a forráskorpuszon tanított modellt értékeltük ki a célkorpuszon.

Két baseline megoldást vizsgáltunk. Az egyikben minden igét és főnévi igenevet eseménynek tekintettünk. A másikban csak azokat az igéket és főnévi igeneveket tekintettük eseménynek, amelyek nem létigék és nem segédigék.

5.2 Eredmények – Detektálás

Az első **baseline** megoldásunk 80,92-es F-mértéket ért el, a másik pedig 85,32-öt.

Teljes jellemzőkészlettel véletlen felosztással a következő eredményeket értük el: Pontosság=93,68, Fedés=94,63, F-mérték=94,15. Tízszeres **keresztvalidációval** 93,85-ös F-mértéket kaptunk. Ha csak az igéket vizsgáljuk, akkor 93,05-ös F-mértéket értünk el. Ha *elhagytuk a szabály alapú módszert*, akkor csak 93,72-es F-mértéket értünk el.

Megvizsgáltuk, hogy az egyes **jellemzőcsoportok** hogyan befolyásolják a gépi tanulórendszer eredményeit. Ehhez *porlasztásos mérést* végeztünk, amelynek az eredményei az 1. táblázatban találhatóak. Ekkor a teljes jellemzőkészletből elhagytuk az egyes jellemzőcsoportokat, majd a maradék jellemzőkre támaszkodva tanítottunk. Az eredmények alapján a leghasznosabbnak a szemantikai, a lexikai és a szintaktikai jellemzők bizonyultak.

1. táblázat: Az egyes jellemzőosztályok

Jellemző	Pontosság	Fedés	F-mérték	Eltérés
Felszíni jellemzők-1: Bi-,tri-, fourgramok	92,44	95,79	94,08	-0,07
Egyéb felszíni jellemzők	92,47	96,23	94,31	+0,16
Lexikai jellemzők	91,73	94,92	93,30	-0,85
Morfológiai jellemzők	92,71	95,94	94,29	+0,14
Szintaktikai jellemzők	92,54	95,36	93,92	-0,23
Szemantikai jellemzők	91,54	94,19	92,85	-1,3

Kiegészítő mérések a Felszíni jellemzők-1 nélküli esetre.

A jellemzőkészletet kiegészítettük **szózsák** jellemzőkkel. Először felvettük a jellemzők közé az adott eseményjelölt szintaktikai alárendeltjeinek lemmájának halmazát. Ezzel 93,49-es F-mértéket értünk el. Utána ehhez hasonlóan a jellemzőkészletet az eseményjelölthöz tartozó szavak és a kapcsolat típusa halmazzal bővítettük. Ezzel 93,81-es F-mértéket értünk el. Látjuk, hogy ezek a kiegészítések nem javítottak a vizsgált eredményen.

Következő mérésként **frekvenciainformációkat** vettünk fel. A tanító halmaz alapján kigyűjtöttük, hogy az egyes igék lemmája milyen arányban esemény. Ezt az arányt is felvettük a jellemzők közé. Ez javított az eredményen: 95,82-es F-mértéket kaptunk. Mivel az igekötő is megváltoztathatja egy ige eseményjellegét, ezért a kö-

vetkező esetben nem a lemmához, hanem az igekötő+lemma párhoz vettünk fel az előzőhöz hasonló arányt. Ezzel még jobb eredményt értünk el: F-mérték= F:95,95.

Korpuszonként is megvizsgáltuk az alkalmazás működését. Ennek eredményei a 2. táblázatban láthatóak. Legjobban az Üzleti rövidhírek és az Újsághírek doménen teljesített a modell, leggyengébben pedig a Jogi doménen.

2. táblázat: Eredmények az egyes részkorpuszokon

Korpusz	**Pontos-ság**	**Fedés**	**F-mérték**
Fogalmazás	94,84	98,00	*96,39*
Jogi	92,11	86,42	*89,17*
Szépirodalom	96,03	96,03	*96,03*
Üzleti rövidhírek	97,14	97,84	*97,49*
Újsághírek	96,73	98,01	*97,37*

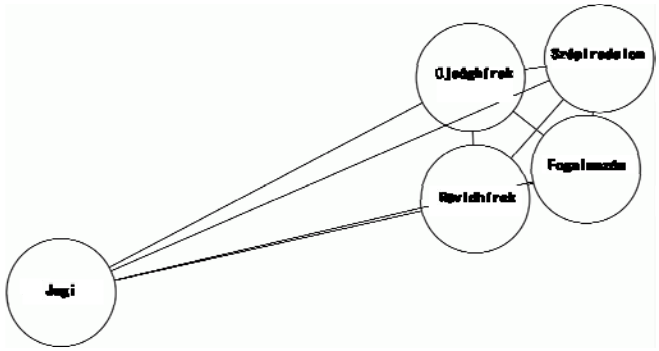
A **domainek közötti keresztméréseknél** a forráskorpuszon tanított modellt értékeltük ki a célkorpuszon. Ennek eredményét a 3. táblázatban láthatjuk. A fogalmazás korpuszon az újsághírek doménen tanított modell teljesített a legjobban 95,42-es F-mértéket elérve. A jogi korpuszon szintén az újsághírek doménen tanított modell teljesített a legjobban 83,11-es F-mértéket elérve. A szépirodalom korpuszon a fogalmazás doménen tanított modell teljesített a legjobban 94,73-es F-mértéket elérve. Az üzleti rövidhírek korpuszon az újsághírek doménen tanított modell teljesített a legjobban 95,71-es F-mértéket elérve. Az újsághírek korpuszon a fogalmazás doménen tanított modell teljesített a legjobban 94,80-es F-mértéket elérve.

3. táblázat: Keresztmérések eredményei az egyes részkorpuszokon

Korpusz	**Pontos-ság**	**Fedés**	**F-mérték**
Fogalmazás	97,49	98,91	*98,20*
Jogi	68,56	99,09	*81,04*
Szépirodalom	92,04	97,58	*94,73*
Üzleti rövidhírek	92,73	98,04	*95,32*
Újsághírek	91,26	98,62	*94,80*
Jogi	95,75	93,88	*94,81*
Fogalmazás	81,70	72,67	*76,92*
Szépirodalom	88,19	72,34	*79,48*
Üzleti rövidhírek	94,72	76,47	*84,62*
Újsághírek	90,74	71,90	*80,23*
Szépirodalom	97,52	98,24	*97,88*
Fogalmazás	94,68	95,64	*95,16*
Jogi	67,05	97,79	*79,56*
Üzleti rövidhírek	92,91	96,16	*94,51*
Újsághírek	91,38	96,22	*93,74*
Üzleti rövidhírek	98,00	99,09	*98,54*
Fogalmazás	92,63	95,86	*94,21*

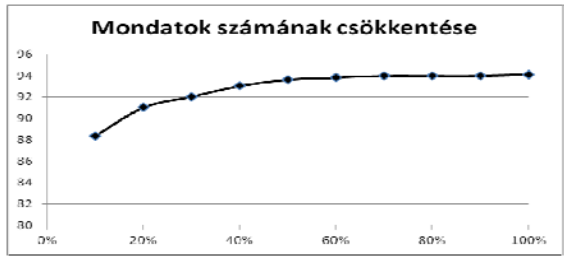
Jogi	69,83	96,74	*81,11*
Szépirodalom	91,26	96,48	*93,79*
Újsághírek	91,29	95,86	*93,52*
Újsághírek	95,06	99,27	*97,12*
Fogalmazás	93,33	97,60	*95,42*
Jogi	72,13	98,05	*83,11*
Szépirodalom	90,83	98,09	*94,32*
Üzleti rövidhírek	93,48	98,04	*95,71*

A keresztmérések eredményei alapján az egyes **domének közti hasonlóságokat** megjelenítettük egy *irányítatlan súlyozott gráf* segítségével. (1. ábra) Az ábrán látható, hogy a jogi korpusz a legkevésbé hasonló a többihez e szempontok alapján.



1. ábra: Doménhasonlósági gráf a keresztmérések eredményei alapján

A következő mérésben **csökkentettük a mondatok** számát. Az F-mértékekre kapott eredmény a 2. ábrán látható. A mondatok számát csökkentve romlik az eredmény.



2. ábra: Mondatok számának csökkentése

6 Igei események osztályozása

Az igei események detektálása után **osztályoztuk** azokat. Az osztályozást több szempont szerint is elvégeztük. Az igék alapkategóriáit vizsgáltuk meg: cselekvés, törté-

nés, létezés, állapot. Ezek közül az eseményeknél a **cselekvés és a történésnek** van fő szerepe, így ezt a két kategóriát emeltük ki. Az 5010 mondaton belül 3905 cselekvés és 1582 történés típusú esemény volt.

Ugyanazt a **jellemzőkészletet** használtuk fel, mint a detektálásnál. Mind a két osztályozásnál, a szemantikai jellemzőknél, a **WordNet**et felhasználva először osztályozóval olyan *synseteket* kerestünk, amelyek *hiponimái* között jellemzően az adott osztály szavai szerepelnek. Ezeket a *synseteket* egy listában felvéve jellemzőként, definiáltuk, hogy az adott szó szerepel e valamelyik ilyen synset *hiponimái* között.

Ha csak a WordNet jellemzőt alkalmaztuk önállóan, a cselekvés vizsgálatnál 87,26, a történés vizsgálatnál 73,31-es F-értéket értünk el.

Mind a két vizsgálathoz készítettünk 1-1 *baseline* megoldást.

6.1 Eredmények – Osztályozás

A **baseline** modellünk minden eseményt cselekvésnek tekintett. Ezzel 78,70-as F-mértéket ért el.

Teljes jellemzőkészlettel véletlen felosztással a következő eredményt értük el az F-mértékre: Cselekvés: 85,93; Történés: 66,06

Tízszeres **keresztvalidációval** kaptuk: Cselekvés: 84,9; Történés: 65,34

Megvizsgáltuk, hogy az egyes **jellemzőcsoportok** hogyan befolyásolják a gépi tanulórendszer eredményeit. Ehhez *porlasztásos mérést* végeztünk, amelynek az eredményei az 4. táblázatban találhatóak, osztályozásonként külön sorban, a következő sorrendben: Cselekvés (Cs); Történés (T). Ekkor a teljes jellemzőkészletből elhagytuk az egyes jellemzőcsoportokat, majd a maradék jellemzőkre támaszkodva tanítottunk. A cselekvés és a történés osztályoknál a szemantikai jellemzők voltak a legmeghatározóbbak.

4. táblázat: Az egyes jellemzőosztályok

Jellemző	Pontosság	Fedés	F-mérték	Eltérés
Felszíni jellemzők-1:	Cs: 82,64	89,49	85,93	*0*
Bi-,tri-, fourgramok	T: 74,48	59,34	66,06	*0*
Egyéb felszíni jellemzők	Cs: 80,67	85,91	83,21	*-2,72*
	T: 78,40	53,85	63,84	*-2,22*
Lexikai jellemzők	Cs: 81,61	88,37	84,85	*-1,08*
	T: 69,80	57,14	62,84	*-3,22*
Morfológiai jellemzők	Cs: 81,01	89,71	85,14	*-0,79*
	T: 87,85	51,65	65,05	*-1,01*
Szintaktikai jellemzők	Cs: 82,39	87,92	85,06	*-0,87*
	T: 77,78	61,54	68,71	*+2,65*
Szemantikai jellemzők	Cs: 78,71	81,88	80,26	*-5,67*
	T: 62,58	53,30	57,57	*-8,49*

Kiegészítő mérések a Felszíni jellemzők-1 nélküli esetre.
Itt is elvégeztük azokat a kiegészítő méréseket, mint a detektálásnál.

A jellemzőkészletet kiegészítettük **szózsák** jellemzőkkel. Először felvettük a jellemzők közé az adott szóhoz szintaktikailag tartozó szavak lemmájának halmazát. Ezzel a következő F-mértékeket értük el:

Cselekvés: 83,35; Történés: 66,07

Utána ehhez hasonlóan a jellemzőkészletet a vizsgált szóhoz tartozó szavak és a kapcsolat típusa halmazzal bővítettük. Ezzel a következő eredményeket értük el:

Cselekvés: 85,12; Történés: 66,67

Ez az utóbbi kiegészítés javított az eredményeken.

Következő mérésként **frekvenciainformációkat** vettünk fel. A tanító halmaz alapján kigyűjtöttük, hogy az egyes igék lemmája milyen arányban tarozik a vizsgált osztályba. Ezt az arányt is felvettük a jellemzők közé. Ez mindegyik osztálynál javított az eredményeken. A következő F-mértékeket kaptuk:

Cselekvés: 86,98; Történés: 76,70

Itt is megvizsgáltuk, hogy ha nem csak a szavakhoz tároljuk el az arányt, hanem az igekötő+lemma párhoz is, akkor az hogyan befolyásolja az eredményt. Ez volt ahol javított az F-mértéken:

Cselekvés: 88,20; Történés: 75,00

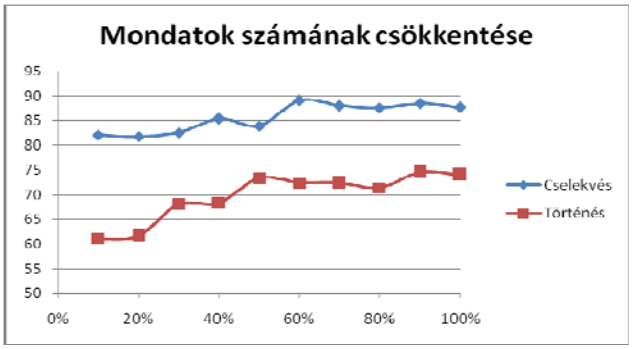
Korpuszonként is megvizsgáltuk az alkalmazás működését. Ennek eredményei az 5. táblázatban láthatóak. A cselekvéseket osztályozó modell a jogi korpuszon a történéseket osztályozó az üzleti rövidhírek korpuszon teljesített a legjobban.

5. táblázat: Eredmények az egyes részkorpuszokon

Korpusz	**Pontosság**	**Fedés**	**F-mérték**
Fogalmazás	Cs: 83,81 T: 57,14	85,44 32,43	*84,62* *41,38*
Jogi	Cs: 90,57 T: 77,78	87,27 87,50	*88,89* *82,35*
Szépirodalom	Cs: 79,44 T: 68,75	85,86 64,71	*82,52* *66,67*
Üzleti rövidhírek	Cs: 91,43 T: 88,33	85,33 94,64	*88,28* *91,38*
Újsághírek	Cs: 83,33 T: 70,00	82,52 61,76	*82,93* *65,63*

Domainek közötti keresztméréseket itt is végeztünk. A forráskorpuszon tanított modellt értékeltük ki a célkorpuszon. Legjobb eredményt a cselekvések osztályozásánál értük el, a szépirodalom doménen tanított modellel a fogalmazás korpuszon 85,5-os F-mértékkel. A leggyengébb eredményt pedig a történések osztályozásánál a fogalmazás doménen tanított modellel a szépirodalom korpuszon 53,91-os F-mértékkel.

A következő mérésben csökkentettük a mondatok számát. Az F-mértékekre kapott eredmények a 3. ábrán láthatóak. A **mondatok számát csökkentve** mindkét osztályozásnál romlottak az eredmények.



3. ábra: Mondatok számának csökkentése – osztályozás

7 Összegzés

Munkánkban bemutattunk gazdag jellemzőtéren alapuló gépi tanuló megközelítésünket, amely automatikusan képes magyar nyelvű szövegekben igei eseményeket azonosítani és azokat osztályozni. A problémát három lépésben oldottuk meg. Először detektáltuk a többszavas főnévi+igei kifejezéseket. Majd detektáltuk az igei és főnévi igenévi eseményeket, és osztályoztuk azokat. Módszerünket a Szeged Korpusz öt doménjén próbáltuk ki tízszeres keresztvalidációval. A modellünk jellemzőkészletét teszteltük porlasztásos módszerrel. A módszert teszteltük a doméneken egyesével, majd tanítva az egyiken a többi doménen pedig kiértékelve. Az egyes domének közötti hasonlóság kifejezésére hasonlósági gráfokat is megadtunk. A detektálásra az alapjellemzőkkel és 93,85-os F-mértéket, a két szempont szerinti osztályba sorolásra pedig 85,93 és 66,06-os F-mértéket értünk el. Kiegészítő mérésekkel javítottuk ezeket az értékeket. Ezek jó eredménynek számítanak a bemutatott előző munkákkal összehasonlítva.

Hivatkozások

1. Bethard, S.J.: Finding Event, Temporal and Causal Structure in Text: A Machine Learning Approach PhD thesis, University of Colorado (2002)
2. Bittar, A.: Annotation of Events and Temporal Expressions in French Texts, ACL-IJCNLP '09 Proceedings of the Third Linguistic Annotation Workshop (2009) 48–51
3. Llorens, H., Saquete, E., Navarro-Colorado, B.: TimeML Events Recognition and Classification: Learning CRF Models with Semantic Roles, COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics (2010) 725–733
4. Marsic, G.: Temporal processing of news: annotation of temporal expressions, verbal events and temporal relations. PhD thesis, University of Wolverhampton (2011)
5. Miháltz, M., Hatvani, Cs., Kuti, J., Szarvas, Gy., Csirik, J., Prószéky, G., Váradi, T.: Methods and Results of the Hungarian WordNet Project. In Tanács, A., Csendes, D.,

Vincze, V., Fellbaum, C., Vossen, P., eds.: Proceedings of the Fourth Global WordNet Conference (GWC 2008), Szeged, University of Szeged (2008) 311–320
6. Nagy T. I., Vincze V., Zsibrita J.: Félig kompozicionális szerkezetek automatikus felismerése doménadaptációs technikák segítségével a Szeged Korpuszon. IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2013) 47–58
7. Subecz Z., Nagyné Csák É.: Események detektálása természetes nyelvű szövegekben. Matematikát, fizikát és informatikát oktatók XXXVII. országos konferenciája, Miskolc (2013) 201–208
8. Vincze V.: Félig kompozicionális főnév + ige szerkezetek a Szeged Korpuszban. VI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2009) 390–393
9. Zsibrita J., Vincze V., Farkas R.: magyarlanc 2.0: szintaktikai elemzés és felgyorsított szófaji egyértelműsítés. IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2013) 368–374