Gradus Vol 9, No 2 (2022)
ISSN 2064-8014

FŐNÉVI ESEMÉNYEK AUTOMATIKUS DETEKTÁLÁSA MAGYAR NYELVŰ SZÖVEGEKEN

AUTOMATIC DETECTION OF NOMINAL EVENTS IN HUNGARIAN TEXTS

Subecz Zoltán [1*] https://orcid.org/0000-0001-5036-4679

[1] Informatika Tanszék, GAMF Műszaki és Informatikai Kar, Neumann János Egyetem, Magyarország
https://doi.org/10.47833/2022.2.CSC.003

Kulcsszavak:
információkinyerés
adatbányászat
szövegbányászat
gépi tanulás
eseménydetektálás

Keywords:
information extraction
data mining
text mining
machine learning
event detection

Cikktörténet:
Beérkezett 2021. február 10.
Átdolgozva 2022. október 31.
Elfogadva 2022. november 5.

Összefoglalás

A természetes szövegekből történő információkinyerés egyik fontos részterülete a névelemek azonosítása mellett az események detektálása. Szövegekben lévő események detektálása és analizálása fontos szerepet tölt be számos számítógépes nyelvészeti alkalmazásban, mint például a kivonatolás és a válaszkeresés. A szövegekben a legtöbb esemény igékhez kapcsolódik, és az igék általában eseményeket jelölnek. De az igéken kívül lehetnek események más szófajú szavak is pl. főnevek, igenevek stb. Munkánkban a szövegekben megtalálható főnévi események detektálásával foglalkoztunk. Jelen tanulmányunkban bemutatjuk gazdag jellemzőtéren alapuló gépi tanuló megközelítésünket, amely automatikusan képes magyar nyelvű szövegekben főnévi események detektálására.

Abstract

Besides named entity recognition, the detection of events in natural language is an important area of Information Extraction. The detection and analysis of events in natural language texts play an important role in several NLP (Natural language processing) applications such as summarization and question answering. Most events are denoted by verbs in texts and the verbs usually denote events. But other parts of speech (e.g. noun, participle) can also denote events. In our work we deal with the detection of nominal events. In this study we introduce a machine learning-based approach with a rich feature set that can automatically detect nominal events in Hungarian texts.

1. Bevezetés

A természetes szövegekből történő információkinyerés egyik fontos részterülete a névelemek azonosítása mellett az események detektálása [7]. Szövegekben lévő események detektálása és analizálása fontos szerepet tölt be számos számítógépes nyelvészeti alkalmazásban, mint például a kivonatolás és a válaszkeresés. A szövegekben lévő események felismerése, analizálása, és hogy hogyan viszonyulnak egymáshoz időben, fontos a szöveg tartalmának megismerésében.

Az esemény, ami történik egy adott helyen és időben. A szövegekben a legtöbb esemény igékhez kapcsolódik, és az igék általában eseményeket jelölnek. De az igéken kívül lehetnek

* Kapcsolattartó szerző. E-mail cím: subecz.zoltan@gamf.uni-neumann.hu

események más szófajú szavak is pl. főnevek, igenevek stb. Munkámban a szövegekben megtalálható főnévi események detektálásával foglalkoztam. Vannak olyan szavak (pl. írás), amelyek egyes mondatokban események, másokban pedig nem, ezért a szavak szövegkörnyezetét is elemezni kell. Jelen tanulmányomban bemutatom gazdag jellemzőtéren alapuló gépi tanuló megközelítésemet, amely automatikusan képes magyar nyelvű szövegekben főnévi események detektálására függőségi elemző és WordNet alkalmazásával. A rendszer bemenete egy token-szinten címkézett tanító korpusz. Modellem jelöltjei a mondatok főnevei voltak.

A feladatokhoz gazdag jellemzőkészleten alapuló osztályozót használtam. A jellemzők mellé kiegészítő módszereket is alkalmaztam, amelyek javították az eredményeket és a futási időt. Módszeremet a Szeged Korpusz öt különböző doménjén vizsgáltam meg.

Angol nyelvű szövegekre általában *konstituensfa alapú* szintaktikai elemzőt használnak az elő-feldolgozásnál az angol nyelv erősen konfiguratív tulajdonsága miatt, ahol is a legtöbb mondatszintű szintaktikai információt a szórenddel fejeznek ki. A konstituensfa (vagy szintaxis fa) egy mondatnak a különböző szintaktikai kategóriáinak fa reprezentációja és segít megérteni a mondat szintaktikai struktúráját.

Ezzel ellentétben a magyar nyelv gazdag morfológiával és szabad szórenddel rendelkezik. A *függőségi fákkal* dolgozó elemzők különösen jól használhatóak szabad szórendű nyelvek elemzésére, így a magyarra is. Ezek ugyanis könnyebben teszik lehetővé az egy-mással nem szomszédos, de összetartozó szavak összekapcsolását is. Ezért én a magyar nyelvű szövegeimre *függőségi fákkal dolgozó elemzőt* használtam. A szavak közötti szintaktikai kapcsolatok alapján a mondatok egy *függőségi fát* alkotnak. A fa legfelső eleme a *Root*. A fa *csomópontjaiban* vannak a mondat szavai, az ágak a szavak közötti szintaktikai kapcsolatokat reprezentálják. Ha a jelölt 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.

Megoldásomban a vizsgált szavak szemantikai jellemzéséhez felhasználtam a magyar *WordNet*-et [10]. Mivel egy szóalakhoz több jelentés is tartozhat a WordNet-ben, ezért az egyes jelentések között egyértelműsítést végeztem a *Lesk algoritmussal* [8].

Algoritmusaimat tesztadatbázisokon kiértékelve, versenyképes eredményeket érnek el az eddig bemutatott angol és más nyelvű eredményekkel összehasonlítva.

2. Kapcsolódó munkák

Az EVITA [13] volt az első esemény felismerő eszközök egyike. Az eseményeket nyelvészeti és statisztikai technikák kombinálásának segítségével ismeri fel. Nyelvészeti ismereteken alapuló szabályokat használ fő jellemzőként. A főnévi esemény felismeréshez WordNet osztályokat is használ, valamint a főnevek szemantikai egyértelműsítésére Bayes osztályozót alkalmaz.

Boguraev és társa [2] gépi tanuláson alapuló módszert mutattak be automatikus esemény-annotáláshoz. A feladatot osztályozásra visszavezetve, RRM (robust risk minimization) osztályozót alkalmaztak. Jellemzőkként lexikai, morfológiai és szintaktikai chunk típusokat használtak két- és háromelemű ablakokban vizsgálva.

Bethard és társa [1] esemény felismerésre fejlesztették a STEP rendszert. Szintaktikai és szemantikai jellemzőket alkalmaztak és az esemény felismerési feladatot osztályozásként oldották meg. Gazdag jellemző készletet építettek be: lexikai, morfológiai, szintaktikai függőségi és választott WordNet osztályokat. E jellemzőkre alapozva Support Vector Machine (SVM) modellt implementáltak.

Llorens és társa [9] bemutattak egy kiértékelést esemény felismerésre. Szemantikai szerepeket adtak meg jellemzőként és CRF (Conditional Random Field) modellt építettek események felismeréséhez.

Jeong és társa [6] függőségi elemzőt használtak, de csak a jelölt főnév és a közvetlenül ahhoz kapcsolódó ige közötti kapcsolatot vizsgálták. Kombinált jellemzőket építettek be, az ige és a hozzá tartozó kapcsolat-típus párokat alkalmazva. A WordNetet is használták, de jelentésegyértelműsítés nélkül. A MaxEnt osztályozási algoritmust a következő jellemzőkészlettel implementálták: felszíni, lexikai, szemantikai, függőségi alapú jellemzők. A jellemzőket a Kullback-Leibler divergencia mód-szerrel súlyozták.

Olasz szövegekre Caselli és társa [3] csak igéből képzett főnévi eseményekkel foglalkoztak, amihez a Weka döntési fa alapú osztályozót használták. Vizsgálták a jelölt argumentum struktúráját, az aspektuális módosítókat, a jelölt előtti és utáni 3-3 szó szófaját.

Spanyol szövegekre Peris és társa [11] szintén csak igéből képzett eseményekkel foglalkoztak. Osztályozásra a Weka döntési fa osztályozóját alkalmazták és külső főnévi lexikont használtak fel. Függőségi elemzőt alkalmaztak, de csak a jelölt főnév és a közvetlenül ahhoz kapcsolódó ige közötti kapcsolatot vizsgálták. Felhasználták a jelölt argumentum struktúráját is.

Német nyelvű szövegekre Gorzitze és társa [5] bootstrapping módszert alkalmaztak események felismerésre. Idővel kapcsolatos kifejezéseket és aspektuális igéket kerestek a jelölt közelében. A jelölt és a közvetlenül ahhoz kapcsolódó ige kapcsolatát vizsgálták és szabály alapú függőségi elemzőt használtak.

Magyar szövegekre Subecz [14] detektált eseményeket, de csak igei és főnévi igenévi eseményekkel foglalkozott. A következő jellemző készletet használta: felszíni, lexikai, morfológiai, szintaktikai, WordNet és frekvencia jellemzők. Ezen jellemzők mellett szabály alapú módszereket is alkalmazott.

3. Események, főnévi események

A szövegekben a legtöbb esemény igékhez kapcsolódik, és az igék általában eseményeket jelölnek. De az igéken kívül lehetnek események más szófajú szavak is például főnevek, igenevek. Munkámban a szövegekben megtalálható főnévi események detektálásával foglalkoztam. Példák főnévi eseményekre: futás, építés, írás, háború, ünnepség.

A főnévi eseményeknek két nagy csoportja van: igéből képzettek (deverbális) és nem igéből képzettek (nem deverbális). Példa igéből képzett eseményekre: futás, írás. Példa nem igékből képzett eseményre: háború. Az igéből képzett főnevek két fő fajtája az események és az eredmények, amelyeknél gyakori a kétértelműség is. Vannak olyan szavak (például írás), amelyek egyes mondatokban események, másokban pedig eredmények.

Például az *írás* főnév a következő mondatban esemény: *Az írás 5 órakor kezdődött*. Viszont a következő mondatban nem esemény, hanem eredmény: *Elolvastuk az írást*. A többértelműség miatt nem elég a szóalak vizsgálata, a szövegkörnyezetet is elemezni kell.

4. Környezet

Alkalmazásomban a Szeged Korpusz [4] egy részét használtam fel a következő területekről: *üzleti rövidhírek, szépirodalom-fogalmazás, számítógépes szövegek, újsághírek, jogi szövegek*. Tanításhoz és kiértékeléshez tízszeres keresztvalidációt alkalmaztam. A mondatokat két nyelvész annotálta, az annotátorok közötti egyetértés Kappa = 0,7 volt.

A feladatokat *bináris osztályozásra* vezettem vissza. Az osztályozáshoz a *Weka programcsomagnak* a J48-as döntési fa elemzőjét használtam fel. A döntési fákat széleskörűen alkalmazzák osztályozási problémákra az NLP-ben, egyik előnye, hogy az elkészített fa könnyen értelmezhető és ellenőrizhető.

A feladathoz alkalmaztam még a Magyarlanc 2.0 programcsomagot is [16]. A csomagot magyar szövegek mondatokra és szavakra bontására, a szavak morfológiai elemzésére, majd szófaji egyértelműsítésére, és mondatok függőségi nyelvtan szerinti szintaktikai elemzésére alkalmaztam. A Magyarlanc programcsomag is készít a szavakhoz morfológiai elemzést, de a HunMorph[15] elemzőcsomag sok esetben részletesebb elemzést ad, ezért ezt is felhasználtam. Így a feladathoz két morfológiai elemzőt is alkalmaztam.

Ahogy láttuk a kapcsolódó munkáknál, mások is használtak függőségi elemzést. De az elemzőfában mindenki csak a jelölt és a vele közvetlen kapcsolatban lévő szavakat vizsgálta. Én vizsgáltam a jelölt és a fában tőle távolabbi igék kapcsolatát is. Modellem jelöltjei a mondatok főnevei voltak. Számos esetben a jelölt alá egy részfa tartozott a függőségi fában.

A vizsgált főnevek szemantikai jellemzéséhez a magyar *WordNet*-et[10] alkalmaztam. A WordNet hiperním hierarchiájában található szemantikai kapcsolatokat használtam fel.

Statisztikai adatok: A vizsgált korpusz 10000 mondatot tartalmaz. Jelöltek száma (főnevek): 48388 db. Pozitív jelöltek száma (esemény főnevek): 7626 db. A jelölteket két fő részre osztottuk a hasonló tulajdonságok alapján. Az egyik csoportba az igéből képzett főnevek, a másikba a többi

főnév került. Igéből képzett jelöltek: 5325 db. Igéből képzett pozitív jelöltek: 4169 db. Nem igéből képzett jelöltek: 43063 db. Nem igéből képzett pozitív jelöltek: 3457 db.

5. Az osztályozás bemutatása

Az osztályozáshoz bináris osztályozót használtunk. A mondatok főnevei voltak a jelöltek. Ezek az elemzőfában egy-egy csomópontot jelentenek.

5.1. Jellemzőkészlet

A tanító és a kiértékelő halmazon a jelöltekhez jellemzőket vettem fel. Módszeremet gazdag jellemzőtérrel valósítottam meg. Az eseménydetektálással kapcsolatos feladatokban gyakran használt jellemzőket én is alkalmaztam. Ezeken kívül újakkal is kibővítettem a jellemzőkészletemet. Az új jellemzőket a magyar szövegek tulajdonságai alapján választottam ki. A jellemzőkhöz felhasználtam a függőségi elemzőfát és a magyar WordNet-et is. A fő jellemző csoportokat több részre bontottam, mivel a részek hatását külön-külön vizsgálom majd. Ezen csoportok közül a következők voltak az új, máshol ezen a területen nem látott jellemzőcsoportok: a *két Morfológiai elemző együtt, az Elemzőfa 1-2, Szózsák 1-3, WordNet jellemzők-2-4*.

Felszíni jellemzők: *Bigramok, trigramok*: A vizsgált szavak végén lévő 2-es, 3-as betűcsoportok. *PositionInSentence*: a jelölt hányadik szó a mondatban. *NagyBetuNemMondatElejen*: Azok a nagybetűs szavak, amelyek nem a mondat elején állnak legtöbbször névelemek, így ez a jellemző utalhat a nem-esemény jellegre.

Morfológiai jellemzők-1: Mivel a magyar nyelv igen gazdag morfológiával rendelkezik, ezért számos morfológia-alapú jellemzőt definiáltam. Ebben a csoportban a Magyarlanc morfológiai elemzőjét használtam fel. Jellemzőként definiáltam az eseményjelöltek MSD morfológiai kódját, felhasználva a következő morfológiai jegyeket: *típus(SubPos), mód(Mood), eset(Cas), idő(Tense), személy(PerP), szám(Num), határozottság(Def)*. További jellemzők: *Lemma*: a jelölt lemmája. *hasVerbRoot*: igéből képzett-e a jelölt. *SzofajElotte* és *SzofajUtana*: a jelölt előtti és utáni szó szófaja. *LegkozelebbiIgeMondatbanLemma*: a jelölthöz a mondatban legközelebb álló ige lemmája. *Igeto*: igéből képzett főnév esetén az alapige.

Morfológiai jellemzők-2: Ebben a csoportban a HunMorph morfológiai elemzőjét használtam fel. *IgetoVan*: van-e igető. *IgebolFonevKepzo*: Igéből képzett főneveknél a képző. *IgeToHunMorph*: igéből képzett főnév esetén az alapige.

Morfológiai jellemzők-3: A Magyarlanc és a HunMorph morfológiai elemző is a többjelentésű szavak esetén minden jelentéshez megadnak külön morfológiai elemzést. Ebben a csoportban mindkét elemző esetén, egyértelműsítés nélkül, megadtam a jelöltekhez minden jelentéshez tartozó ragokat, képzőket, jeleket.

Elemzőfa jellemzők-1: Ezeket a jellemzőket a függőségi elemzőfa alapján készítettem. *JeloltEdgeType*: A jelölt és az elemzőfában a felette levő szó közötti kapcsolat típusa. *JeloltEdgeTypeNE*: A jelölt NE (névelem) típussal kapcsolódik-e a felette levő szóhoz. Ez utal a jelölt nem esemény jellegére. *JeloltFelettLemmaFaban*: A jelölt feletti szó lemmája az elemzőfában. *JeloltFelettIgeLemmaFaban*: Az elemzőfában közvetlenül a jelölt felett lévő ige (ha van) lemmája. *KozvetlenSzintaktikaiKapcsolat*: Ha a jelölt fölött van közvetlenül ige az elemzőfában, akkor a kettő közötti szintaktikai kapcsolat típusa. *LegkozelebbiIgeFeletteFabanLemma, LegkozelebbiIgeFelette-TavolsagFaban*: Az elemzőfában jelölt feletti legközelebbi ige lemmája és annak távolsága a fában a jelölttől. *JeloltReszfaTokenekSzama*: Az elemzőfában a jelölt alá tartozó részfa elemeinek száma. *FeletteSzoEdgeType*: Az elemzőfában a jelölt feletti szó és az a feletti szó közötti kapcsolat típusa. A Magyarlanc elemzőnél az időhatározók, időbeliséget kifejező szavak az események felett helyezkednek el, ezért ezek jelenléte utalhat a jelölt eseményjellegére.

Elemzőfa jellemzők-2: Ha a jelölt nem közvetlenül kapcsolódik a felette levő igéhez az elemzőfában, akkor részletesen jellemeztem a jelölt és az ige közötti útvonalat. *SzófajÚtvonal*: Egymás után írtam a jelölt és az ige közötti csomópontok szófaját. Például: C↑S↑V↑C↑V↑V. *Lemmaútvonal*: Hasonlóan az előzőhöz a jelölt és az ige közötti lemmákat írtam egymás után. Például: napoztatás↑és↑törölgetés↑hajszárító↑megszárít. *SzintaktikaiKapcsolat-Útvonal*: A jelölt és az ige közötti útvonalon a szintaktikai kapcsolatok típusai egymás után. Például: OBL↑COORD↑SUBJ↑COORD↑CONJ↑.

Szózsák jellemzők-1: Szózsák modellt használtam fel szócsoportok jellemzésére. *ReszfaLemmakSzozsakAtlag*: A jelölt alatt lévő részfa szavainak lemmáit reprezentáltam szózsák modellel. A tanító halmazon minden lemmához kiszámítottam, hogy milyen arányban tartozott pozitív jelölt a részfájához. Majd minden jelölthöz kiszámítottam a részfáját alkotó lemmákhoz tartozó arányok átlagát. Nagy átlag arra utal, hogy a jelölt részfájában fontos szavak vannak az eseményjelleg szempontjából. *ReszfaLemmakSzozsakLegnagyobb*: Hasonló az előzőhöz, de itt minden jelöltnél a részfájához tartozó lemmák közül azt választottam ki, amelyikhez legnagyobb valószínűség tartozott. Nagy maximális valószínűség utal arra, hogy a jelölt részfájában van legalább egy olyan lemma, ami erősen fontos az eseményjelleg szempontjából. Ez a jellemző segít a részfa egy-egy fontos szavának felismerésében. *KozvetlenAlattaLemmakSzozsakAtlag* és *KozvetlenAlattaLemmakSzozsakLegnagyobb*: Az előzőkhöz hasonló, de itt nem a jelölt részfájához tartozó minden szót vizsgáltam, hanem csak a részfa azon szavait, amelyek szintaktikailag kapcsolódnak a jelölthöz az elemzőfában. *KozvetlenAlattaEdgeTypeSzozsakAtlag* és *KozvetlenAlattaEdgeTypeSzozsakLegnagyobb*: Az előzőhöz hasonlóan, itt a jelölt és a hozzá szintaktikailag kapcsolódó szavak közötti kapcsolat típusát vizsgáltam. *LemmaParseTreePathIgeigLemmakSzozsakAtlag* és *LemmaParseTreePathIgeig-LemmakSzozsakLegnagyobb*: Ezeknél a szózsákba a jelölt és az elemzőfában felette lévő ige közötti útvonalon található lemmák kerültek.

Szózsák jellemzők-2: Ezekhez a jellemzőkhöz a lemmák a mondatból és nem az elemzőfából lettek kigyűjtve a szózsákba. *MondatbanKornyezet-N-LemmakSzozsakAtlag* és *MondatbanKornyezet-N-LemmakSzozsakLegnagyobb*: A mondatban a jelölt N távolságú környezetét jellemeztem szózsák modellel, N=3 és N=5 esetekben.

WordNet jellemzők: Ezekhez a jellemzőkhöz felhasználtam a magyar WordNet [10] hiperním hierarchiájában található szemantikai kapcsolatokat. Mivel egy szóalakhoz több jelentés is tartozhat a WordNet-ben, ezért az egyes jelentések között egyértelműsítést végeztem a Lesk algoritmussal[8] A WordNetben a synsetekhez definíció és példamondatok tartoznak. Az algoritmus alapján, többjelentésű eseményjelölt esetén megszámoltam, hogy az eseményjelölt szintaktikai környezetében lévő szavak közül hány található meg az egyes WordNet jelentések definíciói és példamondatai között. Azt a jelentést választottam, amelyikkel a legtöbb volt közös szó.

WordNet jellemzők-1: *EsemenyszerusegekAlatt*: A magyar WordNetben van egy mesterséges synset, ami alá jellemzően események tartoznak. De vannak események ezen kívül is, és vannak ebben a csoportban olyanok is, amelyek nem igazi események. Ebben a jellemzőben megadtam, hogy a jelölt beletartozik-e ezen synset hiponím hierarchiájába.

WordNet jellemzők-2: *WordNetSzozsakAtlag* és *WordNetSzozsakLegnagyobb*: A szózsák jellemzőkhöz hasonlóan itt a szózsákba a WordNetben a jelölt hiperním hierarchiájába tartozó szavakat vettem fel. *WordNetSzozsakLegnagyobbSynset*: Megadtam a jelölt hiperním hierarchiájában lévő synset-ek közül azt, amelyik a legnagyobb arányban tartozik események hiperním hierarchiájába.

WordNet jellemzők-3: *WordNetHipernimSynsetekTanulobol* (bináris): Készítettem egy halmazt, amibe kigyűjtöttem a tanító halmazból az esemény jelöltek hiperním hierarchiájának synset-jeit. Majd minden jelölthöz megadtam, hogy a hiperním hierarchiájának synset-jei közül tartozik-e valamelyik ebbe a halmazba.

WordNet jellemzők-4: *WordNetLegjobbLemmakAlatt*: Kigyűjtöttem azokat a lemmákat, amelyek a tanító halmazon legnagyobb arányban voltak események. Majd a jelölteknél jelöltem, hogy a jelölt lemmája alatta van-e valamelyik ilyen kiemelt lemma hiponím hierarchiájának a WordNet-ben.

Szózsák jellemzők-3: Először a Szózsák jellemzők 1-2 csoportoknál bemutatott minden esethez itt kiválasztottam a legjobb elemeket a szózsákokból 1-1 halmazba. Azokat, amelyek legnagyobb arányban tartoztak eseményekhez. Majd a következő jellemzőkkel jelöltem, hogy az adott jelölthöz tartozó szózsák tartalmaz-e az adott halmaz elemei közül. *LegjobbWordNetSynsetek*: A jelölt hiperním hierarchiájába tartozó synsetek között van-e ami szerepel a LegjobbWordNetSynsetek halmazban. *LegjobbReszfaLemmak*: A jelölt részfáinak lemmái között van-e olyan lemma, ami szerepel a LegjobbRészfaLemmák halmazban. *LegjobbLemmakUtvonalIgeig*: A jelölt és az elemzőfában a legközelebbi ige közötti lemmák között van-e olyan lemma, ami szerepel a LegjobbÚtvonalLemmák halmazban.

LegjobbMondatbanKornyezet-N-Lemmak: A mondatban a jelölt N távolságú környezetében van-e olyan lemma, ami szerepel a LegjobbMondatbanKornyezet-N-Lemmak között. Ezt megnéztem N=3 és N=5 esetekre is.

Lista-jellemzők: *FeletteLemmaIdohatarozoListabol*: Listába kigyűjtöttem gyakori idővel kapcsolatos szavakat. (például: előtt, folyamán) Ezek a szavak alatt az elemzőfában gyakran események vannak. Jellemzőként jelöltem, hogy a jelölt felett van-e ilyen idővel kapcsolatos kifejezés. *FeletteIgeAspektualisListabol*: Listába kigyűjtöttem gyakori aspektuális igéket (például elkezd, folytatódik). Ezen igék alá tartozó főnevek gyakran események. Jelöltem, hogy a jelölt felett az elemzőfában van-e ilyen ige.

Kombinált jellemzők-2 eleműek: Ezeknél a jellemzőknél az előző jellemzők közül kombináltam össze kettőt. *JeloltFelettLemmaFaban+JeloltEdge-Type*: Egy szó eseményjellegét gyakran pontosabban jelzi, ha a felette levő lemmát és a kettőjük közötti kapcsolatot együtt vizsgáljuk, mintha csak külön-külön vizsgálnánk azokat. Hasonlóan együtt vizsgáltam a következőket:
JeloltFelettIgeLemmaFaban+JeloltEdgeTypeOBJ,
JeloltFelettIgeLemmaFaban+JeloltEdgeTypeSUBJ,
JeloltFelettLemmaFaban+LegjobbWordNetSynsetek,
JeloltFelettIgeLemmaFaban+ LegjobbWordNetSynsetek

Kombinált jellemzők - 3 eleműek: Az előző kételemű jellemzőkhöz hasonlóan itt három jellemzőt kombináltam össze: *JeloltFelettLemmaFaban+EdgeType+WordNetLegjobbSynset*, *JeloltFelettIgeLemma-Faban+JeloltEdgeType+WordNetLegjobbSynset*,

5.2. További alkalmazott módszerek

A következő módszerek újaknak tekinthetőek, mert ezeken a területeken nem láttam máshol az alkalmazásukat. Mindegyik hasznos volt az eredménye alapján, így más NLP feladatoknál is hasznosak lehetnek.

Statisztikai arány felhasználása az osztályozásnál. A jelöltekhez a jellemzőket *két módszer* alapján választottam ki. *Első módszernél* az előző részben bemutatott alapjellemzőket használtam fel. *Második módszernél* az alapjellemzők helyett a tanító adatokon számított statisztikai arányokat használtam fel. A tanító halmaz alapján megszámoltam minden jellemző esethez, hogy hány alkalommal fordult elő és ebből hányszor volt a jelölt pozitív. Ezek alapján kiszámítottam a hozzá tartozó pozitív-arányt. Például ha a *Lemma* jellemzőnél a *Lemma = írás* eset 5-ször fordult elő és ebből 3-szor volt pozitív eset, akkor hozzá a 0,6-es pozitív-arány tartozott. Ebben az esetben az osztályozónak a jelöltekhez nem az alapjellemzőt, hanem a hozzá tartozó arányt adtam meg. Az előző példánál *Lemma-arány* = 0,6. Ezzel *jelentősen csökkentettem az osztályozó vektorterének méretét* az első módszerhez képest és így a *futási időt* is. Ez a kidolgozási időszakban hasznos volt. A két esetet összehasonlítva azt tapasztaltam, hogy legtöbbször a *valószínűségi módszer* adta a legjobb eredményeket. És a futási idő is *70-80-szor gyorsabb*, mint az alapjellemzők használata esetén.

A jelöltek csoportokra bontása. Az osztályozó hasonló tulajdonságú adathalmazon könnyebben találja meg a szabályokat, mint olyan halmazon, ami sokféle adatot tartalmaz. Ezért érdemes a jelölteket kisebb, hasonló tulajdonságú csoportokra bontani (ezzel megkönnyíteni az osztályozó döntését). Majd a csoportok eredményeit a TP, TN, FP, FN eredmények alapján összegezni. Ennek megfelelően a jelöltjeinket két fő szempont szerint csoportosítottam. Első lépésként a jelölteket két csoportra bontottam: igéből képzett (deverbális) és nem igéből képzett (nem deverbális) főnevek. Hiszen e két csoport tagjai eltérően viselkednek. Az igéből képzett főnevek között sokkal nagyobb arányban vannak események. Másik csoportosítás a jelöltek lemmái alapján történt. Itt 3 alcsoportot képeztem. Első csoportba azok a lemmák kerültek, amelyek nagy arányban események voltak a tanító halmazon. A másik csoportba a többi jelölt lemmája a tanító halmazról. Harmadik csoportba a kiértékelő halmazon azon jelöltek lemmái, amelyek nem szerepeltek a tanító halmazon. Így összesen 2*3=6 csoportot képeztem, és mindegyikre külön-külön végeztem el az osztályozást.

Valószínűségi arányok felhasználása az osztályozás eredményeinek javítására. Azokban az esetekben, amikor gyenge eredményt kaptam, (általában a kis fedés miatt), akkor az osztályozás elvégzése után azon jelölteknél, amelyek a tanító halmazon nagy arányban voltak események, az

értékelést a kiértékelő halmazon pozitívra állítottam. Ennek során, ha egy jellemzőre kicsi volt a fedés a tanító halmazon, de a jellemző egy adott értékéhez a tanító halmazon nagy arányban tartoztak események, akkor a kiértékelésnél a jelöltet eseménynek jelöltem.
Majd az eredményeknél látni fogjuk, hogy ezek a kiegészítő módszerek jelentősen javították az eredményeinket és a futási időt.

5.3. Jellemző-esetek számának csökkentése

A *vektortér méretét csökkentettem* a következő módszerrel: csak azokat a *jellemző-előfordulásokat* vettem fel az osztályozáshoz, amelyek a tanító halmazon *legalább háromszor* szerepeltek. Ezzel *jelentősen csökkentettem a futási időt* és csak az osztályozás szempontjából jelentéktelen jellemző-előfordulásokat hagytam ki.

Az 5. fejezetben leírt jellemzőket és módszereket a rendszer automatikusan legyűjti és megvalósítja a bemeneti szövegeken.

6. Eredmények

A kiértékelés során a pontosság (P), fedés (R) és F-mérték (F) metrikákat használtam.

6.1. Baseline mérés

Modellem hatékonyságának vizsgálatához Baseline mérést végeztem. Ennek keretében a jelöltek közül az igei alapúakat vettem pozitív esetnek a többit pedig negatívnak. Ennek eredménye a következő volt: pontosság: 66,67, fedés: 47,57, F-mérték: 55,52. A további eredményeimen látni fogjuk, hogy *gépi tanulási módszerem jóval felülteljesítette a Baseline mérés eredményét*.

6.2. Modellem eredménye

Gépi tanulási módszerem a következő eredményt érte el a teljes korpuszon az adott jellemzőkészlettel és a kiegészítő módszerekkel: Pontosság: 79,25, Fedés:67,04, F-mérték: 71,94. A kiegészítő módszerek alkalmazása nélkül a következő eredményeket kaptam: Pontosság: 70,32, Fedés: 60,51, F-mérték: 65,03. Látható, hogy a kiegészítő módszerekkel jelentős javulást tudtam elérni. A javulás 80%-át a jelöltek csoportosítása adta, a kisebb részt az osztályozás utáni javításból származott. Ha csak az első szempont szerint csoportosítottam, akkor azt kaptam, hogy az igéből képzett főnevek esetén a modell sokkal jobb eredményt ért el (F-mérték: 84,62), mint a nem igékből képzett főneveknél (F-mérték: 39,52).

Modellemet megvizsgáltam az öt részkorpuszon is. Ezekre az 1. táblázatban látható F-mértékeket kaptam.

1. Táblázat. Eredmények a részkorpuszokon (%)

Részkorpusz	***F-mérték***
Szépirodalom-fogalmazás	**75,24**
Újsághírek	**76,31**
Üzleti rövidhírek	**75,12**
Számítógépes szövegek	**71,57**
Jogi szövegek	**68,74**

Legjobb eredményemet az *újsághírek* doménen, a legrosszabbat pedig a *jogi szövegeken* kaptam.

6.3. Eredmények porlasztásos méréssel

Megvizsgáltam, hogy az egyes jellemzőcsoportok hogyan befolyásolják a gépi tanuló-rendszer eredményeit. Ehhez porlasztásos mérést végeztem. Ekkor a teljes jellemző-készletből elhagytam az egyes jellemzőcsoportokat, majd a maradék jellemzőkre támaszkodva tanítottam. Ennek eredményei a 2. táblázatban találhatóak. Az adatok azt mutatják, hogy az adott jellemzőcsoportot

elhagyva hogyan változott az eredmény. A csökkenő (negatív) eredmény azt jelzi, hogy a vizsgált jellemzőcsoportnak pozitív hatása van az esemény felismerésben.

2. Táblázat. A porlasztásos mérés eredményei (%)

Elhagyott jellemzők	***Változás az F-mértékben***
Felszíni jellemzők	**-0,28**
Morfológiai jellemzők-1	**-2,51**
Morfológiai jellemzők-2	**-0,52**
Morfológiai jellemzők-3	**-2,01**
Elemzőfa jellemzők-1	**-1,92**
Elemzőfa jellemzők-2	**-0,52**
Szózsák jellemzők-1	**-1,34**
Szózsák jellemzők-2	**-2,42**
Szózsák jellemzők-3	**-0,57**
WordNet jellemzők-1	**-0,32**
WordNet jellemzők-2	**-6,51**
WordNet jellemzők-3	**-0,53**
WordNet jellemzők-4	**-0,2**
Lista jellemzők	**0.0**
Kombinált jellemzők - 2 eleműek	**-0,79**
Kombinált jellemzők - 3 eleműek	**+0,1**

Ha a hasonló jellemzőcsoportokat összevonjuk, akkor a következő eredményeket kapjuk az összevont csoportokra (3. táblázat):

3. Táblázat. A porlasztásos mérés eredményei - összevonással (%)

Elhagyott jellemzők	***Változás az F-mértékben***
Morfológiai jellemzők	**-1,63**
Szózsák jellemzők	**-4,0**
Elemzőfa jellemző	**-1,56**
WordNet jellemző	**-7,7**
Kombinált jellemzők	**-0,95**

A 2. és a 3. táblázat eredményein látszik, hogy majdnem minden jellemző-csoportnak pozitív hatása volt a modell teljesítményére. Legjobb hatása a WordNet és a Szózsák jellemzőknek volt, de sokat javítottak a Morfológiai és az Elemzőfa jellemzők is. Mindkét morfológiai elemző hatása pozitív volt. A WordNet jellemzők-2 részcsoportnak volt a legjobb hatása (6.51%). Ebben használtam együtt a WordNet-et a szózsák modellel. A Lista jellemzőknek nem volt hatása. Negatív hatása volt a 3 elemű kombinált jellemzőknek, de a 2 elemű kombinált jellemzők hasznosak voltak.

A modellem, amelynek eredményét a 6.2-es fejezetben ismertettem, már csak a pozitív hatású jellemzőket tartalmazta.

6.4. Az eredmények összehasonlítása a kapcsolódó munkákkal

Angol szövegekre Jeong és társa [6] 71,8%-os, Romeo és társai [12] 67%-os F-mértéket értek el. Olasz nyelvre Caselli [3] 69%-os, spanyol nyelvre Peris és társai [11] 59,6%-os F mértéket értek el. A kapcsolódó munkákkal összehasonlítva, eredményeim (F-mérték = 71,9%) jónak számítanak.

Összegzés

Munkámban bemutattam gazdag jellemzőtéren alapuló gépi tanuló megközelítésemet, amely automatikusan képes magyar nyelvű szövegekben főnévi eseményeket detektálni. Öt részterületet vizsgáltam meg, összesen 10 000 mondattal. Gazdag jellemzőtéren alapuló jellemzőkészletemben felszíni, morfológiai, függőségi elemzőfa, szózsák, Wordnet, lista és kombinált jellemzőket használtam fel. Ezen jellemzőcsoportok mellet kiegészítő módszereket is alkalmaztam, amelyek javították modellem hatékonyságát, valamint a futási időt. Algoritmusaimat tesztadatbázisokon kiértékelve, versenyképes eredményeket érnek el az eddig bemutatott angol és más nyelvű eredményekkel összehasonlítva.

Irodalomjegyzék

[1] Bethard, S., Martin, J.H.: Identification of event mentions and their semantic class. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pp. 146–154. Association for Computational Linguistics (2006) https://doi.org/10.3115/1610075.1610098

[2] Boguraev, B., Ando, R.K.: Effective use of Timebank for TimeML analysis. In: Schilder, F., Katz, G., Pustejovsky, J. (eds.) Annotating, Extracting and Reasoning about Time and Events. LNCS, vol. 4795, pp. 41–58. Springer, Heidelberg (2007) https://doi.org/10.1007/978-3-540-75989-8_4

[3] Caselli, T., Russo, I., Rubino, F.: Recognizing deverbal events in context. In: Proceedings of CICLing 2011, poster session. Springer (2011)

[4] Csendes, D., Csirik, J.,Gyimóthy, T.: The Szeged corpus: a POS tagged and syntactically annotated Hungarian natural language corpus. In: Sojka, P., Kopecek, I., Pala, K. (eds.) TSD 2004. LNCS (LNAI), vol. 3206, pp. 41–47. Springer, Heidelberg (2004) https://doi.org/10.1007/978-3-540-30120-2_6

[5] Gorzitze, S., Pado, S.: Corpus-based acquisition of German event- and object denoting nouns. In: Proceedings of KONVENS 2012 (Main Track: Poster Presentations), pp. 259–263 (2012)

[6] Jeong, Y., Myaeng, S.: Using syntactic dependencies and Wordnet classes for noun event recognition. In: The 2nd Workhop on Detection, Representation, and Exploitation of Events in the Semantic Web in Conjunction with the 11th International Semantic Web Conference, pp. 41–50 (2012)

[7] Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice-Hall, Upper Saddle River (2000)

[8] Lesk, M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation, pp. 24-26, New York, NY, USA. ACM. (1986) https://doi.org/10.1145/318723.318728

[9] Llorens, H., Saquete, E., Navarro-Colorado, B.: TimeML Events recognition and classification: learning CRF models with semantic roles. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 725–733. Association for Computational Linguistics (2010)

[20] 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), pp. 311–320. University of Szeged, Szeged (2008)

[31] Peris, A.,Taule, M.,Boleda, G.,Rodriguez,H.: ADN-classifier: automatically assigning denotation types to nominalizations. In: Proceedings of the Seventh LREC Conference, 19-21 May 2010, Valetta, Malta, pp. 1422–1428 (2010)

[12] Romeo, L.,Lebani, G.E.,Bel,N., Lenci, A.: Choosing which to use? A study of distributional models for nominal lexical semantic classification. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), pp. 4366–4373 (2014)

[43] Sauri, R., Knippen,R., Verhagen, M.,Pustejovsky, J.: Evita: a robust event recognizer for QA systems. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 700–707. Association for Computational Linguistics (2005) https://doi.org/10.3115/1220575.1220663

[54] Subecz, Z.: Detection and classification of events in Hungarian natural language texts. In: Sojka, P., Horak, A., Kopecek, I., Pala, K. (eds.) TSD 2014. LNCS (LNAI), vol. 8655, pp. 68–75. Springer, Heidelberg (2014) https://doi.org/10.1007/978-3-319-10816-2_9

[65] Tron, V., Kornai, A., Gyepesi, G., Németh, L., Halácsy, P., Varga, D. Hunmorph: Open source word analysis. In Proceedings of the Workshop on Software, Software '05, pp. 77–85, Stroudsburg, PA, USA. Association for Computational Linguistics. (2005) https://doi.org/10.3115/1626315.1626321

[76] Zsibrita, J., Vincze,V., Farkas,R.: magyarlanc: a toolkit for morphological and dependency parsing of Hungarian. In: Proceedings of RANLP 2013, pp. 763–771 (2013)

[17] Wei Xiang, Bang Wang: A Survey of Event Extraction From Text. Published in: IEEE Access (Volume: 7), 2019, pp. 173111 – 173137, Electronic ISSN: 2169-3536, DOI: 10.1109/ACCESS.2019.2956831