Subecz Zoltán - Nagyné Csák Éva

Igei események detektálása magyar szövegekben

Subecz, Zoltán - Nagyné Csák, Éva: Verbal Event Detection in Hungarian Texts

The task of event detection is to identify mentions of events in texts. For the purposes of this task, an event mention is any expression denoting an event or state that can be assigned to a particular point, or interval, in time. In texts most event mentions correspond to verbs, and most verbs introduce events. However this is not always the case. Events can be introduced by noun phrases, and some verbs fail to introduce events. Events sometimes are expressed with multiword expressions. Multiword expressions (MWEs) are lexical units that consist of more than one orthographical word, i.e. a lexical unit that contains spaces. The task is the detection of events in texts. In this article we paid attention with verbs and infinitives. We wrote Java program for this task. We introduce in this article the principles and results of this application.

Keywords: Event detection, Information extraction, Data Mining, Text mining, Programming

Összefoglaló

Az események detektálásának a feladata az esemény-előfordulások azonosítása a szövegekben. Esemény előfordulásnak tekintünk minden olyan kifejezést, ami olyan eseményt vagy állapotot jelöl, amit egy adott időponthoz, vagy intervallumhoz tudunk kapcsolni. 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. Az igék közül pedig nem tekintünk minden szót eseménynek. Vannak olyan események, amelyeket két szóval fejezünk ki (pl. döntést hoz). És ez a két szó, a magyar nyelv szabad szórendje miatt, gyakran nem egymás mellett található. A feladat a szövegekben megtalálható események detektálása. A cikkben az igei és főnévi igenévi egy- és többszavas események detektálásával foglalkozunk. A feladatra Java nyelven írtunk programot. A cikkben bemutatjuk a program készítési elveit, lépéseit és eredményeit.

Kulcsszavak: Eseménydetektálás, Információkinyerés, Adatbányászat, Szövegbányászat, Programozás

Bevezetés

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. Ennek a kutatásnak a fő területe az eseménydetektálás. Olyan program készítése, ami egy adott szóról, vagy kifejezésről eldönti, hogy esemény, vagy nem. Majd látjuk, hogy ez gyakran nem könnyű feladat. A többértelműség miatt általában a szó szövegkörnyezetét és a mondatban betöltött szerepét is vizsgálni kell.

1. Események, és azok detektálása

Eseményeket több fajta szófajhoz tartozó szó hordozhat. Ezek közül a leggyakoribbak az igék

és a főnévi igenevek. Például fut, olvas, úszik. Nem csak a megtörtént eseményeket keressük. Attól, hogy az eseményt tagadjuk vagy jövő időben/felszólító/feltételes módban stb. szerepel, attól még esemény marad. A következő szavak is csak módosítják az eseményt valamilyen modális vagy aspektuális szempontból: kell, muszáj, szabad, akar, fog, szokott, tud, kíván, óhajt, mer, szándékozik.

Az igék mellett más szófajok is kifejezhetnek eseményeket. Például a főnevek, melléknévi igenevek és határozói igenevek. Példák ezekre a csoportokra: főnévi események: írás, olvasás, kertészkedés, szállító, képviselő; melléknévi igenévi események: lángoló, épülő; határozói igenévi események: futva, idézve, hazaérkezve. Mindegyik csoportnál találkozunk többértelműséggel. 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 forró melléknévi igenév esemény a következő mondatban: Láttam a tűzhelyen forró vizet. Viszont a következő mondatban nem esemény: Idd meg a tejet, de vigyázz, mert még forró. Az eseménydetektálás nehezebb feladata az ilyen többértelmű szavaknál eldönteni, hogy eseményt jelöl az adott szó, vagy sem. Ehhez a szövegkörnyezet vizsgálatára is szükség van.

Ebben a cikkben csak az igei és főnévi igenévi események elemzésével és detektálásával foglalkozunk. Ezen belül két fő feladatot kell megoldani. A több szóból álló kifejezések felismerése; és azon igék meghatározása, amelyek események.

2. Igei események

Az igék nagy része eseményt jelöl, de van néhány, amelyik ettől kivétel. Ide tartoznak a segédigék is. A segédige olyan viszonyszó, mely az ige grammatikai jelentéseit hordozza. A segédige szám- és személyjelentést, idő- és modális jelentést, ritkábban aspektuális és pragmatikai jelentést tartalmaz. A segédigék az utána következő főnévi igenévvel alkotnak egy egységet. Önállóan nem tekintjük eseménynek, hanem az utána álló főnévi igenevet tekintjük annak. Például Tud úszni. esemény: úszni; segédige: tud. Példák nem esemény igékre: van, marad, múlik, fog, volna, szokott, akar, bír, kell, kezd, kíván, lehet, mer, óhajt, szabad, szándékozik, szeretne, szokott, talál, tetszik, tud, látszik.

Az egyik megoldandó feladat a többértelműség kezelése. Példa többértelműségre: Tegnap találtam egy kulcsot. Itt a talál szó esemény. Azt találtam mondani, hogy... Itt a talál szó nem esemény. Ebben a mondatban a hozzá tartozó mond szó az esemény. Másik példa többértelműségre: Otthon maradt. marad: itt nem esemény. Lemaradt a vonatról. marad: itt esemény. Látjuk, hogy az igekötő is megváltoztathatja az ige eseményjellegét.

A program feladata itt az igékre és a főnévi igenevekre a szövegkörnyezet alapján eldönteni, hogy esemény, vagy nem. Ehhez a programot a bemeneti szövegek egy részén tanítottuk, egy másik részén pedig kiértékeltük a program működését. A program a bemeneti mintákat és szabályokat felismeri, megtanulja, és a kiértékelésnél azokat alkalmazza az ismeretlen szövegeken.

3. A több szóból álló kifejezések

Vannak olyan igei események, amelyek nem csak egy szóból állnak. Ilyenek például a félig kompozicionális szerkezetek (FX-ek). Az FX-ek egyik típusa a főnévből és igéből álló többszavas kifejezések (VerbFX). (Például döntést hoz, fény derül, győzelmet arat, kereslet támad) Ezekben a kifejezés szemantikai tartalmát nagyrészt a főnév hordozza, ugyanakkor az ige vállal főszerepet a szerkezet szintaxisának kialakításában. Mivel jelentésük nem teljesen kompozicionális, a szerkezet részeinek egyenkénti lefordítása nem (vagy csak nagyon ritkán) eredményezi a szerkezet idegen nyelvű megfelelőjét. A szerkezet nyelvi elemzésekor célszerű a főnevet és az igét egy komplex egységként kezelni. [Vincze 2009:390] Vannak olyan FX-ek amelyek kettőnél több szóból állnak, ezért számítástechnikai felismerésük nem könnyű. (Például döntést fog hozni; Hozzátok meg a szükséges döntéseket.) De vannak

olyan FX-ek is, amelyek között 8-10 másik szó található. A többszavas események felismerésére már született egy alkalmazás magyar nyelvre [Nagy 2013:47-58]. A többszavas kifejezések detektálásánál a feladatunkhoz igazítva felhasználtuk ennek az alkalmazásnak az elveit is.

A program először kiválogatja a többszavas kifejezéseket, majd az igéknél és a főnévi igeneveknél detektálja az eseményeket.

4.1 Korpuszok, annotálás

Az alkalmazásunk teszteléséhez a Szeged Korpusz egy olyan változatát használtuk fel, amelyikben annotálva vannak a többszavas kifejezések. [Vincze 2009:390-393] Az annotált korpuszban a tanításhoz be vannak jelölve a többszavas kifejezések. 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.
Az igei események detektálásához is ezt az 5010 mondatot használtuk fel. Ezeket a mondatokat az eseménykinyerés szempontjából magunk annotáltuk. Minden igéhez és főnévi igenévhez bejelöltük, hogy eseményt jelöl-e, vagy nem. Ezeket a bejelöléseket használjuk fel a program tanításánál és a kiértékelésnél.

4.2 Statisztikai adatok

Mondatok száma a Tanító korpuszon: 4510
Mondatok száma a Kiértékelő korpuszon: 505
Mondatok száma összesen: 5015

A Tanító korpusz adatai:
Tokenek száma: 90205
FX kifejezések száma: 488
  Ebből Esemény - FX-ek száma: 361
Igék száma (főnévi igenév nélkül): 8491
  Ebből Esemény - Igék száma: 4894

A Kiértékelő korpusz adatai:
Tokenek száma= 10086
FX kifejezések száma: 54
  Ebből Esemény - FX-ek száma= 45
Igék száma (főnévi igenév nélkül)= 954
  Ebből Esemény - Igék száma= 593

4.3 Felhasznált programcsomagok

A feladatokat osztályozásra vezettük vissza. Az osztályozáshoz a Weka programcsomagnak a J48-ad döntési fa elemzőjét használtuk fel. (Weka: Data Mining Software in Java) [Weka 2013]. A Weka adatbányászati feladatokhoz készített gépi tanuló algoritmusok gyűjteménye.

A feladathoz felhasználtuk még a Magyarlanc 2.0 programcsomagot. [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.

5. A TÖBB SZÓBÓL ÁLLÓ KIFEJEZÉSEK DETEKTÁLÁSA

Minden potenciális kifejezésről el kell dönteni, hogy FX kifejezés, vagy sem (bináris osztályozás). A bemeneti mondatokat először a Magyarlanc programmal dolgoztattuk fel. Ez mondatonként csoportosítva minden szóhoz egy elemzést készít. Minden szóhoz megadja a szó lemmáját, szófaját, morfológiai jegyeit (SubPOS: típus; Mood: mód; Tense: idő; Per: személy; Num: szám; Def: határozottság). A program függőségi fa alapú szintaktikai elemzőt használ, ami egy elemző fába rendezi a mondat szavait a közöttük lévő kapcsolatok alapján. Így szavanként az előbb felsorolt információk mellé meghatározza, hogy melyik szóhoz kapcsolódik a mondatban szintaktikailag, és mi a kapcsolódás típusa. [Subecz 2013:203-205 ,Zsibrita 2013:368-374]

A következő lépésben kiválogattuk azon szócsoportok, részmondatok halmazát, amelyekben az FX-ek nagy része jó fedéssel megtalálható (FX jelöltek). Erre két módszert vizsgáltunk meg. Az első módszernél szófaji sorrendek alapján kerestünk FX jelölteket. Ezeknél az FX-jelölteknél a kifejezés egyik végén N(főnév), a másikon V(ige) áll. Kigyűjtöttük, hogy az FX kifejezések szófajai hogyan követik egymást a Magyarlanc programmodul eredményét felhasználva. A leggyakoribbak a [N, V] alakú FX-ek voltak. Például: igénybe vesz. A második leggyakoribbak a [V, N] alakúak voltak. Például: kerül sor. A vizsgálatban ezzel a leggyakoribb két FX mintázattal foglakoztunk [N, V] és [V, N]. Ezt a két mintázatot Reguláris kifejezések segítségével kerestettük meg a tanító korpuszban. A reguláris kifejezés egy olyan, bizonyos szintaktikai szabályok szerint leírt karaktersorozat (String), amivel meghatározható karaktersorozatok egy halmaza. Az ilyen kifejezés valamilyen minta szerinti szöveg keresésére (esetleg cseréjére) használható.

A másik módszert a szintaktikai elemző segítségével készítettük el. Az FX kifejezések igei(V) és főnévi(V) szavai között nagy fedéssel a következő kapcsolatok állnak a Magyarlanc elemzésénél: SUBJ – alany, OBJ - tárgy, OBL – egyéb névszói. Ilyen szintaktikai kapcsolatban lévő főnévi-igei párok meghatározásával [N…V] és [V…N] alakú FX jelölteket kapunk.

Ezen FX-jelölteknek csak egy kis része lesz a valódi FX. Hiszen ha egy mondatban szerepel a döntés és a hoz szó, akár egymástól távol, nem jelenti azt, hogy ezek egy összetartozó kifejezés szavai.

6. Jellemzőtér kiválasztása a több szóból álló kifejezésekhez

A következő lépésben a tanítást készítettük elő. Kiválasztottunk olyan jellemzőket (feature), amelyek segítenek majd az FX jelöltek osztályozásában. Olyan tulajdonságokat kerestünk, amelyek jól jellemezhetik az FX kifejezéseket. A következő jellemzőket választottuk ki:

- Már rendelkezésre állt egy lista (FX-lista1), amelyik tartalmaz gyakori FX kifejezéseket (579 kifejezés). Ezt a listát kiegészítettük a tanító halmazon lévő annotált FX–ekkel, manuális ellenőrzés után (FX-lista2). A lista minden kifejezését úgy állítottuk össze, hogy [N V] formában tartalmazza az FX-et. Az N szónál a lemmát tároltuk el, a V szónál pedig az abszolút lemmát. Például az álomba merült kifejezés esetén az álom merül szópárt tároltuk az FX-listában. A jellemzőben az tároltuk el, hogy az FX-jelölt szerepel-e valamelyik listában. Külön megvizsgáltuk, azt hogy milyen eredményt ad, ha csak az első listát használjuk fel, és ha mind a kettőt.
- Egy másik összegyűjtött lista tartalmazza a leggyakoribb igéket, amelyek előfordulnak FX kifejezéseknél: ad, fog, folytat, hoz, jut, kap, kerül, nyújt, tart, tesz, végez, vesz, köt, ér, lép, áll, játszik. A következő jellemzőben azt tároltuk, hogy az FX jelölt igéje szerepel-e ebben a listában, és ha igen, akkor melyik ezek közül.
- A következő jellemzőben megadtuk az FX jelölthöz tartozó morfológiai jegyeket (SubPos, Mood, Cas, Num, PerP), amelyek közül többről a Magyarlanc bemutatásánál már volt szó.
- Egy jellemzőben megadtuk, hogy az FX jelölt első szava ige, vagy főnév.
- A főnévi szónál jellemzőként megadtuk, hogy a főnév szótöve igei alapú-e. Ez azért előnyös, mert az FX kifejezések főnévi tagja gyakran igéből származik. (használatba vesz, megoldást talál)
- Frekvenciainformációk. A tanító korpusz alapján megszámoltuk, hogy egy FX-jelölt N és V szava van egy mondatban, akkor az milyen gyakran, és milyen arányban valódi FX kifejezés.

- Végül megadtuk jellemzőként, hogy a megismert három szintaktikai osztály (SUBJ, OBJ, OBL) melyike áll fenn az FX jelöltnél.

Ezek után elvégeztük a tanítást és a kiértékelést a Weka programcsomag J48-as döntési fa alapú tanító algoritmusával.

7.1 Igei események detektálása és a jellemzőtér

Itt a program feladata az igék és főnévi igenevek esetén eldönteni, hogy esemény, vagy nem.

Ehhez a feladathoz külön osztályozót készítettünk. Az osztályozáshoz minden igéhez és főnévi igenévhez a következő jellemzőket válogattuk ki:

- Morfológiai jegyek (SubPos, Mood, Cas, Num, PerP), mint az FX-modulnál
- Felszíni jellemzők. Szóhossz lemmahossz, valamint a szó sorszáma a mondaton belül.
- Lexikai jellemzők. Az adott szó létige, vagy segédige-e?
- Mivel egy szónak az eseményjellegét meghatározza az is hogy előtte, vagy utána áll-e létige vagy segédige, ezért ezt a négy jellemzőt is felvettük.
- Frekvenciainformációk. A tanító korpuszon minden igére megszámoltattuk, hogy milyen arányban esemény, vagy nem. Mivel az igekötő is megváltoztathatja egy ige eseményjellegét, ezért minden igekötő-ige párhoz megszámoltattuk, hogy milyen arányban esemény, vagy nem esemény.
- 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.
- Szintaktikai jellemzők. A Magyarlanc program megadta a szavak közötti szintaktikai kapcsolatokat és azok típusát. Ezen jellemzőcsoportban eltároltuk, hogy az adott ige alá milyen kapcsolattal tartoznak szavak. Például, alany, tárgy stb.

7.2 Szabály alapú módszerrel való kiegészítés

A jogi szövegeknél sok olyan ige található, amelyek más szövegkörnyezetben események, de itt nem azok. Például: A törvény kimondja, hogy…. A paragrafus rögzíti, hogy…. A kimondja és a rögzíti szavak más környezetben események, de itt nem események. Ezért felvettünk szabályokat az ilyen típusú alany+ige típusú esetekre, amelyeket a detektálásnál alkalmaztunk. Például: ha alany=törvény és ige=kimondja, akkor kimondja ≠ esemény.

8. Eredmények az 5000 mondatos korpuszon

A kiértékelésnél a megszokott értékeket 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);
F-mérték: F=2*P*R/(P+R)

8.1 Az FX-modul eredményei

Itt a következő vizsgálatokat végeztük el:
A eset: A vizsgálatban csak a kiinduló FX-listát használtuk fel (FX-lista1), és nem használtuk fel a frekvenciainformációkat.
B eset: A kiinduló FX-lista1 mellett felhasználtuk a frekvenciainformációkat is.
C eset: Felhasználtuk a tanítókorpusz alapján készített másik FX-listát is (FX-lista2). (Szótárillesztés)
A mérési eredményeket az 1. táblázat tartalmazza:
Az adatokon látszik, hogy a frekvenciainformációk használata jelentősen javította az FX-modul eredményét (+14%). A szótárillesztéses módszer ezen az eredményen is jelentősen javított (+17%).

1. táblázat: Eredmények - több szóból álló kifejezések (%)

	Pontosság	Fedés	F-mérték
A, FX-lista1-el frekvenciainformációk nélkül	90,48	41,30	56,72
B, FX-lista1-el frekvenciainformációkkal	96,43	58,70	72,97
C, Szótárillesztéssel: FX-lista2-vel	93,18	89,13	91,11

Forrás: saját elemzés

8.2 Az Eseménydetektáló modul eredményei

Itt a következő vizsgálatokat végeztük el:
A eset: Csak igék vizsgálata
B eset: Igék és főnévi igenevek vizsgálata
C eset: Kiegészítés szabály alapú módszerrel

A mérési eredményeket az 2. táblázat tartalmazza:
A 95% feletti eredmények jónak mondhatóak ezen a területen.

2. táblázat: Eredmények - Események detektálása (%)

	Pontosság	Fedés	F-mérték
A, Csak igék vizsgálata	94,22	96,29	95,25
B, Igék és főnévi igenevek	94,85	96,23	95,53
C, Kiegészítés szabály alapú módszerrel	95,67	96,23	95,95

Forrás: saját elemzés

8.2.2 Baseline modell

Az Eseménydetektáló feladathoz készítettünk egy egyszerű Baseline modellt is, azért hogy az eredményünket értékelni tudjuk. A modell minden igét eseménynek tekint. Ez az egyszerű modell a következő eredményt éri el az eseménydetektálásnál (%):
Pontosság: 67,95 Fedés:100 F-mérték: 80,92

Látjuk, hogy a módszerünk sokkal jobb eredményt ért el, mint a Baseline modell, azt 15%-al haladta meg.

Összegzés

Munkánkban az igei események detektálásával foglalkoztunk. Az események detektálásának a feladata az esemény-előfordulások azonosítása a szövegekben. A feladatot két részre osztottuk. Az első részben a szövegekben azonosítottuk a többszavas főnévi + igei kifejezéseket. A második részben pedig az igék és főnévi igenevek közül kiválogattuk az eseményeket. Az események kiválogatását a többértelműség nehezíti meg, ezért az eseményjelöltek szövegkörnyezetét is vizsgálni kell. A feladatokra szövegbányászati módszereket alkalmaztunk, és Java nyelven készítettünk hozzá alkalmazást. Mindkét modulnál az alkalmazások jó eredményeket adtak.

Felhasznált irodalom

[1.] Nagy T. István, Vincze Veronika, Zsibrita János [2013]: Félig kompozicionális szerkezetek automatikus felismerése doménadaptációs technikák segítségével a Szeged Korpuszon. IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 47-58. p.
[2.] Subecz Zoltán, Nagyné Csák Éva [2013]: Események detektálása természetes nyelvű szövegekben. MAFIOK 2013-Matematikát, fizikát és informatikát oktatók XXXVII. országos Konferenciája Miskolc, 201-208. p.
[3.] Vincze Veronika [2009]: Félig kompozicionális főnév + ige szerkezetek a Szeged Korpuszban. VI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 390-393. p.
[4.] Weka [2013] Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka/
[5.] Zsibrita János, Vincze Veronika, Farkas Richárd [2013]: Magyarlanc 2.0: szintaktikai elemzés és felgyorsított szófaji egyértelműsítés. IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 368-374. p.