# Bűnügyi névelem-felismerés

Molnár Gábor József[1], Kojedzinszky Tamás[1], Farkas Richárd[2]

[1]Szegedi Tudományegyetem, Informatikai Tanszékcsoport
6720 Szeged, Árpád tér 2.
gjmolnar@inf.u-szeged.hu, Kojedzinszky.Tamas@stud.u-szeged.hu
[2]MTA-SZTE, Mesterséges Intelligencia Tanszéki Kutatócsoport
Szeged, Tisza Lajos krt. 103. III. lépcsőház
rfarkas@inf.u-szeged.hu

**Kivonat:** Ebben a munkában bemutatjuk a Szervezett Bűnözés Elleni Koordinációs Központ és a Szegedi Tudományegyetem közös projektjében elkészült magyar nyelvű névelem-felismerő rendszert. A feladat bűnügyi dokumentumok szövegeiben található fontosabb szereplők felismerése, azaz előre definiált kategóriákba tartozó kifejezések azonosítása és azok megfelelő osztályba sorolása volt. A feladat és megoldásának érdekességei, hogy egyrészt egy sokosztályos klasszifikációt kellett megoldani, ahol az osztálycímkéken egy rendezés van definiálva, másrészt a szekvenciajelölés kétszintű: az első szinten jelölt tokenek további osztályokba sorolandók, valamint mivel a rendszer pontossága kiemelt szempont volt, ezért megvizsgáltuk kézi szabályok integrálási lehetőségeit is.

## 1 Bevezetés

Ebben a munkában egy komplex bűnügyi névelem-felismerő rendszert mutatunk be, amely a Szervezett Bűnözés Elleni Koordinációs Központ és a Szegedi Tudományegyetem közös projektjében készült el. A rendszer alapvetően a korábban Szegeden kidolgozott névelem-felismerő keretrendszer [1] egy kiterjesztett változata.

A kiterjesztésre a feladat három specialitása miatt volt szükség.

- A feladat egyik érdekessége, hogy a szokványos négy névelemosztály (földrajzi hely, személynév, szervezetnév, egyéb tulajdonnév) helyett 13 szemantikai osztályt kellett megkülönböztetnünk.

- Mivel a felismert említésekben azonosítani kellett azok alkotóelemeit is (például személyneveken belül vezetéknevet és keresztnevet), kétszintű predikciós megközelítéseket implementáltunk és teszteltünk empirikusan. Ezen módszerek közül kettő egy-egy teljes modellt épít mindkét szintre, míg a harmadik módszer abban tér el a korábbiaktól, hogy a második szint minden egyes osztályára külön gépi tanuló modellt igényel.

- Habár a névelemosztályok többsége nem ismerhető fel jól reguláris kifejezések, listák illesztésével, ezek nagyban hozzá tudnak járulni egy statisztikai rendszer pontosságához. A szabályalapú módszereket többféleképpen kombináltuk statisztikai megközelítésekkel (feltételes valószínűségi mezőket alkalmaztunk), összesen három névelem-felismerő megközelítést vizsgáltunk.

Kiemeljük, hogy az egyes osztályok közti többértelműség igen magas volt, azaz a tulajdonnevek több osztályba is besorolhatóak. Ilyen esetek tipikusan a helységnevekkel kapcsolatosan fordulnak elő legtöbbször. A valós életben sok vezetéknévvel találkozhatunk, amelyek egyben településnevek is. Ilyen esetekben természetesen nem helyként szeretnénk klasszifikálni a kifejezést, hanem személynévként, de számos példa fordulhat elő helynevek és szervezetek (pl. *Szegedi Tudományegyetem*), vagy szervezetek és személynevek esetén is (pl. *Vörösmarty Mihály Általános Iskola*). Ezek a példák azért okoznak problémát, mert egy egyszerű listaillesztő rendszer képes felismerni helyként a *Szegedi* szót és szervezetként a *Szegedi Tudományegyetemet* is. Az általunk alkalmazott gépi tanulási módszer ezt kiküszöböli, a szövegkörnyezet alapján dönti el, hogy melyik jelölés alkalmasabb.

A dokumentumokban szereplő tokenek (szóalakok) képezték az osztályozás alapegységét, és rájuk épültek a gépi tanulási modellek. A tanító adatbázis kialakítása során a szövegek a whitespace és minden látható speciális karakter mentén lettek szavakra bontva.

Az egyes tokeneket bináris jellemzővektorok reprezentálták (részletes leírásukat [1] tartalmazza). A vektorban szereplő nullák és egyesek azt jelzik, hogy az adott tokenre teljesül-e az adott jellemző. A jellemzők megadására egy paraméterfájlon keresztül nyújt lehetőséget a rendszer[1]. Ebben a fájlban található minden olyan információ, amely a tanulási folyamat által használt jellemzőkhöz kapcsolódik.

A különböző rendszerek kombinációit empirikus módon értékeltük ki és vetettük össze. Az így kapott eredmények jól mutatják, hogy a probléma nehézsége ellenére a kapott jelölések jól közelítik az emberi annotációt.

## 2 Kétszintű címkézés

A kétszintű osztályozás miatt természetesen adódik az a lehetőség, hogy minden szintre egy külön modellt építsünk. A fő kizáró oka annak, hogy egy modellel hozzunk döntést az első és a másodrendű kifejezésekre is, az, hogy a két szint szoros kapcsolatban áll egymással.

Ahhoz, hogy képesek legyünk a szintezettségnek megfelelően címkézni, minden szinten különböző modellekre van szükség. Azonban a hierarchia miatt nem elegendő egyszerűen elfogadni az egyes modellek predikcióit, hanem szükséges még a szintekhez tartozó modellek kombinálása is.

[1] A paraméterezhető tulajdonnév-felismerő rendszer a Creative Commons licenc alatt elérhető: www.inf.u-szeged.hu/rgai/NER

A probléma megoldására egyetlen másodrendű modell helyett külön másodrendű modell készült az egy szülőhöz tartozó másodrendű jelölésekhez. Például egy-egy modell készült külön az előtag, vezetéknév, keresztnév osztályokhoz, amelyek a személynév első szintű jelöléshez tartoznak. Ez azt jelenti, hogy a tanító adatbázisból kigyűjtöttük a másodrendű címkével annotált kifejezéseket, úgy, hogy külön tanító adatbázis épült minden azonos elsőrendű szülővel rendelkező jelöléshez. Ezzel több új tanítóhalmaz alakult ki, amelyeknek száma megegyezett azon elsőrendű jelölések számával, amelyeknek léteznek leszármazottai (olyan első szintű osztályok is voltak, amelyekhez nem tartozott másodrendű leszármazott). Ezután az eredeti tanító adatbázist használtuk elsőrendű modell építésére, az újabb tanítóhalmazok segítségével pedig több kisebb modell készült. Ezek a kis modellek gyorsan felépültek, hiszen a tanító algoritmusnak nem kellett foglalkoznia azokkal a tokenekkel, amelyek egyik névelemosztályba sem estek (a névelem-kategóriákba eső kifejezések száma csak a töredéke azon tokeneknek, amelyek egyik tulajdonnévosztályba sem tartoznak), és képes volt kizárólag egy elsőrendű jelölés leszármazottaira fókuszálni.

## 3 Szabályalapú jelölés

A szabályalapú jelölések majdnem 20 címkére vonatkoznak (összesen több mint 50 első, illetve másodrendű osztály van), és egy-egy címkére több reguláris kifejezés is van definiálva. Az egyes felismerő kifejezések megírása különösen problémás volt, hiszen a nyelv változatossága miatt – akár az olyan szabványosnak vélt egyedek, mint a telefonszám is – több vagy összetett szabályok megírására volt szükség. Előnyük, hogy ha egy szövegrészre jelölést tesznek, az nagy valószínűséggel helyes is, azonban csak kevés egyedet fednek le.

Az ún. „Egymást követő" megközelítések a szabályalapú és a gépi tanulási módszereket külön-külön futtatják le, meghatározott sorrendben egymás jelöléseire adott megkötésekkel.

Az „RB + CRF" jelölés lényege, hogy elsőként a reguláris kifejezések illesztése történt a nyers szövegre, majd ezt követte a gépi jelölés (gépi tanuló modellként a Conditional Random Fields /CRF/-et használtuk [2]). Mivel a szabályok nagy pontossággal jó kategóriába sorolják a kifejezéseket, ezért abban az esetben, amikor a gépi tanuló modellnek is volt egy alternatív jelölése arra a kifejezésre, amelyre már a reguláris kifejezés illesztett, azt nem vettük figyelembe, és a szabály annotációját tekintettük érvényesnek.

A „CRF+RB" módszer esetén először a gépi tanuló modell jelölt, és csak utána következett a szabályalapú módszer. A modell nagy valószínűséggel jelöléseket végzett a szövegnek azon részein is, amelyekre egyértelmű szabályokat adtunk.

A „Bővített jellemzők" esetén (ez a módszer bizonyult a leghatékonyabbnak) a tanítás előtt a szabályalapú jelölő által készített jelölések bekerülnek a tanító adatbázisba a megfelelő tokenek mellé extra jellemzőként, így a statisztikai tanulóalgoritmus ezekkel a tulajdonságokkal egészítheti ki a jellemzőkészletét, majd a predikció során is lefutó szabályfelismerő növeli a felismerés pontosságát.

## 4 Eredmények

A tanuló adatbázis és a tesztfájlok kialakítására a rendelkezésre bocsátott 200 dokumentumból volt lehetőség. Mivel az összes dokumentum adatvédelmi okok miatt szigorú anonimizálási folyamaton esett át (pl. eredetileg egy nevet a „Vvvvv Kkkkk” karaktersorozatra cseréltek), így az anonimizált halmazon tanult modell és az ezen mért teszteredmények nem mutatnak pontos eredményeket.

A valós adatokon készített modell (a Szervezett Bűnözés Elleni Koordinációs Központban házon belül) és az azon végzett tesztelés összességében valamivel elmarad az anonimizált adatokon végzettétől, azonban vannak névelemosztályok, melyek esetében javultak az elért eredmények.

Az alábbi táblázatban látható az előző fejezetben bemutatott három különböző szabályalapú és gépi tanulási módszer kombinálására szolgáló módszer eredménye (néhány osztálycímkékre vonatkozóan és a végső rendszer pontosságára).

1. táblázat: Eredmények.

| | RB + CRF | CRF + RB | Bővített jellemzők |
|---|---|---|---|
| Hely | 78,12 | 78,5 | 79,72 |
| Irányítószám | 97,92 | 98,92 | 100 |
| Város | 85,25 | 84,43 | 94,71 |
| Kerület | 95,05 | 88,42 | 100 |
| Utca | 80,39 | 78,38 | 95,95 |
| Házszám | 81,25 | 77,73 | 98,29 |
| Személynév | 96,02 | 96,46 | 97,11 |
| Előtag | 89,06 | 89,6 | 88,72 |
| Vezetéknév | 95,19 | 95,19 | 98,2 |
| Keresztnév | 96,81 | 96,69 | 99,2 |
| végső F-mérték: | **88,28** | **87,84** | **91,33** |

Jól látható, hogy az „Egymást követő” módszerek közötti különbség elhanyagolható, viszont a „Bővített jellemzők” módszere szignifikánsan jobb eredményt produkált.

## Bibliográfia

1. Farkas R., Szarvas Gy.: Nyelvfüggetlen tulajdonnév-felismerő rendszer, és alkalmazása különböző domainekre. In:, Alexin Z., Csendes D. (szerk.): IV. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2006) 22–31
2. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of ICML (2001)