Információkivonatolás szabad szövegekből szabályalapú és gépi tanulásos módszerekkel

Miháltz Márton[1], Schönhofen Péter[2]

[1] Pázmány Péter Katolikus Egyetem Információs Technológiai Kar
H-1083 Budapest, Práter utca 50/a
mmihaltz@gmail.com
[2] in4 Kft
1011 Budapest, Bem rakpart 26. III/2.
schonhofen@gmail.com

Kivonat: Bemutatunk háromféle megközelítést egy információkivonatoló rendszerre, melynek célja doménfüggő szöveges információk kinyerése nagy tételben angol nyelvű Wikipédia-szócikkekből. Az első megközelítés mély nyelvi elemzést és manuálisan létrehozott információkinyerő mintákat használ. Ennek kiterjesztése egy olyan módszer, mely képes annotált példamondatok segítségével ilyen mintákat automatikusan megtanulni. A harmadik módszer csupán szófaji egyértelműsítésre támaszkodik és felügyelt gépi tanulást alkalmaz. Mindhárom módszer esetében bemutatjuk azok kiértékelését és összehasonlítását, két különböző doménen (tanulmányi adatok, díjak elnyerése.)

1 Bevezetés

Szeretnénk bemutatni egy saját fejlesztésű információkivonatoló rendszert nagy mennyiségű, megbízható szöveges információ kinyerésére angol szövegekből, mely az iGlue projekt [1] – melynek célja személyek, földrajzi helyek, intézmények stb. adatainak egységesen kezelt, szemantikusan összekapcsolt adattárba gyűjtése – számára készült.

A rendszer bemeneti szövegállománya jelenleg a Wikipédia nyílt tartalmú webes enciklopédia [2] angol nyelvű szócikkeinek halmaza. A fejlesztéshez használt első domén a tanulmányi adatok területe volt. Az egyes személyekhez az alábbi attribútumokat szerettük volna kinyerni a róluk szóló Wikipédia-szócikkek szöveges részéből: oktatási intézmény neve, ahol a személy tanulmányokat folytatott; tanulmányok kezdete és vége (dátumok); fokozatszerzés dátuma; elért tudományos fokozat; tanulmányterület. Például:

In 1977*, he graduated magna cum laude from* Harvard University *with a* B.A. *in* mathematics *and* economics*.*

Intézmény neve: Harvard University
Tanulmányok kezdete: -
Tanulmányok vége: -

Fokozatszerzés dátuma: 1977
Elért fokozat: B.A.
Tanumányterület(ek): mathematics; economics

Az információkivonatoló rendszer működése mély nyelvi elemzésen és az ezeken definiált minták, valamint az egyes attribútumok névelemtípusainak felismerésén alapul. A kivonatoló minták az igei vonzatkereteken (tagmondatok főigéje és annak vonzatai, szabad határozói) alapulnak. A mintakészlet előállítására mind teljesen manuális, mind félig automatikus módszerekkel is kísérletet tettünk. Emellett bemutatunk egy kísérletet a feladat megoldására felügyelt gépi tanulással is.

A cikk következő részében ismertetjük az elsőként alkalmazott nyelvi elemzés és névelem-felismerés fontosabb részleteit, a felmerült problémákra adott megoldásainkat, valamint egy teljesen manuálisan létrehozott mintákkal működő rendszer kiértékelésének eredményeit. A 3. részben ismertetjük a felügyelt gépi tanulásos megközelítést és összevetjük a mintafelismerésen alapuló módszerrel. Végül a 4. részben bemutatunk egy kísérletet a mintaalapú megközelítés részben automatikussá tételére.

2 Mintaalapú információkivonatolás

2.1 Korpuszépítés

Az információkivonatolás forrásául a Wikipédia-szócikkeit választottuk, mivel ezek nagy mennyiségben állnak rendelkezésre, viszonylag egységes, géppel jól feldolgozható enciklopédikus stílust követnek, valamint a nyílt közösségi fejlesztői megközelítés miatt tartalmilag elfogadható pontosság jellemzi őket.

A korpusz alapja a statikus Wikipédia dump [4] 2008 júniusi verziója volt, mely összesen mintegy 2.4 millió szócikket tartalmaz. Ezek között heurisztikákkal korábban sikerült beazonosítani 90.000, nagy valószínűséggel személyekről szóló szócikket, ez képezte a feldolgozás bemenetét. A HTML-oldalak szöveges tartalmát (nyers szöveget tartalmazó bekezdések) elkülönítettük a formázástól, külön megtartva olyan metainformációkat, mint az oldal címe és különböző címváltozatai (egy adott oldalra utaló átirányító oldalak (redirection page) követésével), kategóriacímkéi, a szövegben lévő hiperlinkek stb.

2.2 Nyelvi elemzés

A nyers szöveget a LingPipe mondatszegmentáló eszközével [3] bontottuk mondatokra, majd ezt követte a nyelvi elemzés az *Enju parser* szintaktikai elemzővel [5]. Az Enju egy gyors, valószínűségi HPSG-nyelvtannal működő angol parser, mely képes predikátum-argumentum szerkezetek és frázisstruktúrák azonosítására. A következő lépésben az Enju által létrehozott elemzési szerkezetek eredményeiben azonosítottuk az igei szerkezeteket, majd az utolsó lépésben ezeken működött az esemé-

nyeken (igei szerkezeteken) alapuló információkinyerő modul. Az alábbiakban az utóbbi két modulról lesz bővebben szó.

Az Enju parser kimenete a frázisstruktúra-viszonyokat XML-hierarchiában, míg a predikátum-argumentum viszonyokat és egyéb jellemzőket (pl. morfológiai információk, aspektus, igenem stb.) jegyszerkezetek formájában adja meg az egyes mondatokra.

A feldolgozás során először azonosítottuk a mondatot alkotó VP-k közül azokat, melyek az információkivonatolás számára releváns információkat tartalmaznak (mellérendelt tagmondatok, vonatkozó mellékmondatok, bizonyos határozói mellékmondatok (pl. „miután", „mielőtt")). A tagadott, vagy nem állító módban álló főigéjű VP-ket kihagytuk.

A következő lépésben az egyes VP-ket alkotó összetevőket azonosítottuk: főige (és partikulája), alany, direkt tárgy és indirekt tárgy, valamint a vonzat vagy módosítói szerepet betöltő prepozíciós frázisok.

Az NP-kben csak a fejjel bezárólag vettük figyelembe a tokeneket, illetve a fej után következő appozíciókat és birtokos szerkezeteket. Az NP-k elejéről elhagytuk a determinánsokat, birtokos névmásokat, prepozíciókat stb.

Ha a főige vonzata igei volt, akkor a beágyazott igét és annak vonzatait/határozóit is azonosítottuk.

Minden összetevőben azonosítottuk az azt alkotó tokenek felszíni alakját, lexikai alakját, szófajkódját, valamint mondatbeli pozícióját.

A koordinált összetevőket szétbontottuk és előállítottuk a többi összetevővel való összes kombinációjukat.

Példa:

Input mondat:

After receiving a Bachelor's Degree in mathematics and physics at the University of Michigan, he went on to obtain a Ph.D. in electrical engineering at Harvard in 1998.

Output elemzési szerkezetek (egyszerűsített):

((Verb, "receive"), (Subj, "he"), (Obj, "Bachelor's Degree"), (PP-in, "mathematics"))
((Verb, "receive"), (Subj, "he"), (Obj, "Bachelor's Degree"), (PP-in, "physics"))
((Verb, "go on"), (Subj, „he"), (Verb2, „obtain"), (Obj2, "Ph.D."), (PP-in2, "electrical engineering"), (PP-at2, "Harvard"), (PP-in2, "1998"))

2.3 Névelem-felismerés

Az információkinyerő minták a nyelvi elemzésben azonosított igei szerkezetekre, mint eseményekre, illetve ezek összetevőire, mint „szereplőkre" alapulnak. Az egyes mintákban az eseménykeret különböző szereplőire (oktatási intézmény, elért tudományos fokozat, tanulmányterület, végzés dátuma stb.) a szintaktikai tulajdonságokon felül szemantikai megszorításokat is tettünk. Így például egy lehetséges szabály az alábbi mintának felelhet meg:

```
Subj(PERSON) + V('attain') + Obj(DEGREE) + PP-in(SCHOOL)
+ PP-in(DATE)
```

Vagyis megköveteljük, hogy a VP feje az „attain” ige legyen, az alanyi szerepű igevonzat SZEMÉLY típusú névelem legyen, a tárgy egy TUDOMÁNYOS FOKOZAT típusú NP stb.

A szemantikai megszorítások (névelemtípusok) ellenőrzésére reguláris kifejezéseket és/vagy lokális lexikonokat használtunk fel. A lexikonok minél kimerítőbb összeállításához számos online információforrást és weboldal anyagát felhasználtuk (WordNet, Wikipédia, CrunchBase, univ.cc stb.) Így pl. a lehetséges tanulmányterületek listája mintegy 2.100, az oktatási intézmények listája 34 ezer tételt tartalmazott.

2.4 Mintaillesztés

Az információkinyerő modullal csak azokat az igei szerkezeteket dolgoztuk fel, amelyekben valamelyik meghatározott igevonzat/módosító azonos volt vagy a címszóban megjelenő személynévvel, annak valamilyen névváltozatával, vagy egy (hím- vagy nőnemű) személyes névmás volt, ezzel valószínűsítve azt, hogy a kinyert információ a kérdéses személyre vonatkozik.

A kérdéses eseményszereplőket a főigétől függően kb. 20 összetett szabály (minta) azonosította. A minták hivatkoznak a nyelvi elemzés által azonosított összetevőkre, valamint használják a felismerhető szemantikai kategóriákat (névelemtípusokat).

A minták fejlesztéséhez és folyamatos, iteratív validációjához készítettünk egy fejlesztői korpuszt, melyben humán annotátorok 200 db, véletlenszerűen kiválasztott személy Wikipédia-szócikkében azonosították a releváns tanulmányi attribútumokat. A minták és a mintafelismerés fejlesztéséhez ezen a halmazon végeztünk folyamatosan pontosság- és fedésméréseket, illetve elemeztük a negatív találatokat.

2.5 Problémák

A munka során számos olyan probléma merült fel, melynek során az Enju parser hibás elemzésein kellet korrekciót végezni.

Az első problémát a prepozíciós frázisok illesztési problémája jelentette (PP-attachment problem), a parser ugyanis inkonzekvens módon ugyanolyan típusú PP-ket különböző esetekben különböző összetevőkhöz kapcsolt. Emiatt a VP-kben a PP-ket rendezetlen listaként kezeltük, és speciális szabályokkal vettük őket figyelembe. Így például az időhatározókat (dátum típusú NP-k 'in' vagy 'on' prepozícióval) a mondatbeli pozíciójukat figyelembe vevő szabályokkal azonosítottuk.

Egy másik, igen gyakori problémát a névelemek (named entityk) határainak hibás felismerése okozta. Ennek orvosolására igyekeztünk minél több névelemet az elemzés előtt, a szegmentált nyers szövegen felismertetni és speciális karakterekkel egyetlen input tokenné összevonni, hogy a parser ezután egyetlen (főnévi) entitásként kezelje őket. A névelemek elő-felismerésének legegyszerűbb eszköze az eredeti szövegben nagy kezdőbetűket tartalmazó, wikipédiás hiperlinkkel ellátott szövegdarabok (anchor textek) azonosítása volt, mivel ezek nagy valószínűséggel tulajdonnévi enti-

tásoknak felelnek meg. Szintén felismertük és összevontuk azokat a többszavas névkifejezéseket, melyek többszavas, nagy kezdőbetűs tokeneket tartalmazó Wikipédia-oldal-címekkel voltak azonosak.

Hasonló probléma volt, hogy az elemző koordinációként értelmezett bizonyos, vesszőt tartalmazó névelemtípusokat, például dátumokat, vagy az angolban gyakori intézménynév-vessző-földrajzi összetételű tulajdonneveket (pl. University of California, Berkeley.) Az előbbiek felismerésére reguláris kifejezéseket, az utóbbiakhoz reguláris kifejezéseket és névlistákat használtunk (34 ezer oktatási intézménynév, 2,3 millió földrajzi név).

Egy további, gyakori problémát a többelemű NP-felsorolások hibás, néha koordinációként, néha appozícióként való elemzése jelentett, ezt az Enju kimenetének feldolgozása során külön szabályokkal kellett korrigálni.

2.6 Kiértékelés

A rendszer kiértékeléséhez az annotátorokkal készíttettünk egy újabb, 100 szócikkből álló annotált kiértékelő halmazt. Ezen a mintán kiszámítottuk a tanulmányok doménen működő, kézzel fejlesztett mintákon alapuló információkivonatoló rendszer pontosságát és fedését a kinyert attribútumokra nézve. Pontosságon a rendszer által helyesen megadott értékek és a rendszer által megadott értékek arányát, fedésen a rendszer által helyesen megadott és a referenciaértékek arányát értjük. A pontosság 94,22%, a fedés 60,33% volt ezen a mintán (F-mérték = 73,55%)

3 Információkivonatolás felügyelt gépi tanulással

A tanulmányok domén esetében a rendszer teljesítményének növelésére kísérletet tettünk a szabályalapú megközelítés ötvözésére felügyelt gépi tanulásos módszerekkel. A tanításhoz a Wikipédia-kategóriacímkék felhasználásával, valamint kézi annotációval generáltunk mintegy 200 tanítópéldát, azonban a szabályalapú módszerhez képest csak kevesebb attribútumot tudtunk azonosítani (intézmény neve, tudományos fokozat, fokozatszerzés dátuma).

A példákat csupán mondatszegmentálásnak, tokenizálásnak és szófaji egyértelműsítésnek vetettük alá. A tanulóalgoritmus a maximum entropy módszert használta [6], a felhasznált feature-ök a kérdéses elemet megelőző és az azt követő n-gramok (n=1,2,3 és n=1,2), illetve az azt megelőző legközelebbi ige töve voltak.

A kiértékelő halmaz segítségével elvégeztük a 3 attribútum gépi tanulásos módszerrel történő felismerésének külön-külön kiértékelését (pontosság és fedés), majd egyenként megvizsgáltuk, hogy a szabályalapú módszer kimenetének metszetével (pontosság várható növekedése) vagy uniójával (fedés várható növekedése) érünk-e el jobb eredményeket (1. táblázat.) A legjobb eredményeket az intézménynév és a tudományos fokozat attribútumok esetében, a két módszer eredményeinek uniójával kaptuk. Az intézményneveknél a szabályalapú módszerhez képest a kombinált módszer a fedésen 9,15%-os növekedést %), a pontosságon 1.25%-os csökkenést (91.01%) eredményezett, míg a tudományos fokozatoknál a fedés 18,28%-os növe-

kedést (80,88%), a pontosság 0,24%-os csökkenést (94,01%) mutatott. A hibrid módszerrel így sikerült a teljes rendszer fedését szignifikánsan növelni, miközben a pontosságot is sikerült a kritikusnak ítélt 90%-os küszöb felett tartani.

1. táblázat: A szabályalapú és a gépi tanulásos módszerek, valamint ezek uniójának és metszetének pontossága és fedése az egyes tanulmányi attribútumok felismerésében.

	Intézménynév		Fokozatszerzés dátuma		Tudományos fokozat	
	P	R	P	R	P	R
Szabály-alapú	92,25%	67,29%	100%	54,69%	94,25%	62,60%
Gépi tanulásos	90,81%	40,63%	84,51%	46,88%	91,93%	43,51%
Unió	91,01%	**76,44%**	89,71%	75%	94,01%	**80,88%**
Metszet	100%	10,50%	100%	4,69%	100%	2,29%

4 Mintaáltalánosítás

Az információkinyerő rendszer fejlesztésének következő szakaszában kísérletet tettünk egy olyan változat kifejlesztésére, mely képes annotált példamondatokból jórészt automatikus módon, információkinyerő mintákat önállóan tanulni. A cél egy olyan általános metódus kifejlesztése volt, mely annotált példákból kiindulva, a szükséges humán munkaerő-ráfordítást minimalizálva adaptálható egy-egy újabb IE-doménre akár egyetlen munkanap alatt is. A humán annotátor feladata csupán a rendszer által megtanult minták ellenőrzése, kiegészítése, illetve az esetlegesen előforduló negatív minták felismerése és megjelölése lenne.

4.1 Tanítópéldák előállítása

Annotált tanítópéldák előállításához felhasználtuk a Yago projekt [7] eredményeit, mely a teljes angol nyelvű Wikipédia-szócikkállomány strukturáltan rendelkezésre álló (tehát nem a szabad szöveges részekbe eső, hanem a keretes részekbe (info boxok) tartozó, kategóriacímkékben megjelenő) információit dolgozta fel és szervezte szemantikai hálózatba.

A Yago tudásanyagának egy része 2-argumentumú relációk formájában áll rendelkezésre. A relációkban álló párok a Wikipédia-szócikkekben jellemzett entitások (pl. személyek, intézmények stb.) Az entitások mind WordNet-synsetek, mind Wikipédia-kategóriaosztályok alá vannak rendelve. Feltételezve, hogy bizonyos redundancia várható a Wikipédia-szócikkek strukturált és strukturálatlan részei között, az entitások neveit a Wikipédia-szócikkek szövegében visszakeresve automatikusan előállíthatunk annotált tanítópéldákat egy-egy Yago-relációhoz.

A mintaáltalánosító eszköz fejlesztéséhez a díjadás domént használtuk fel (ki milyen díjat, elismerést, kitüntetést stb. nyert), mivel összehasonlítási alapként ehhez is