XI. Erdélyi Tudományos Diákköri Konferencia

Angol nyelvű szintaktikus szövegelemzők kiterjesztése szemantikus és valószínűségi elemekkel

Szerző:
Szilágyi Péter

Babeş-Bolyai Tudományegyetem
Matematika és Informatika Kar
Informatika szak, 3. évfolyam

Témavezető:
Dr. Csató Lehel

Babeş-Bolyai Tudományegyetem
Matematika és Informatika Kar
Programozási Nyelvek és Módszerek Tanszék

Kolozsvár
2008. május 23-24.

Kivonat

Jelen pillanatban a számítógépek természetes szövegekből információt kizárólag ismerős sablonok alapján tudnak kinyerni, vagyis az óriási adathalmaznak csupán a felületét érintik. Több információ kinyerésére már nem elegendőek a sekély elemzések, szükség van módszerekre amelyek a szöveget mélyen tudják elemezni és megállapítani benne a szemantikus összefüggéseket. A dolgozat bemutat egy lehetséges megközelítést arra, hogy hogyan lehetne egy szövegnek a már létező szintaktikus elemzők által létrehozott mondattani értelmezését számítógépek számára kezelhetőbb formára hozni. E cél elérésére három módszer lesz felhasználva: a szavak fontosabb kategóriákba való besorolása szemantikus szabályok alapján; a mondattani értelmezésből kiindulva a kategorizált szavak közti relációk felépítése szintaktikus szabályok alapján; illetve a többértelmű relációk tisztázása egy maximum entrópia modell alapján.

1. Bevezető

A számítógépeknek és az nek köszönhetően a jelent az információ korszakának nevezik. Ebben a korszakban az adatok főként emberektől emberek fele áradnak, minimális szabályszerűséggel és nagy komplexitással (kötetlen szöveg, hanganyag, képanyag). Egy felmérés alapján [4] a 2002 év leforgása alatt 33 terrabyte egyedi nyomtatott publikáció jelent meg, amit már önmagában emberileg teljes mértékben lehetetlenség feldolgozni, és ez az szám azóta csak növekedett. Érthető tehát, hogy egyre inkább csábító a rengeteg adat számítógépekkel való feldolgozása, viszont ez jelen pillanatban még komoly gondot jelent, mivel a természetes szövegek komplexitása és szabálytalansága teljes mértékben használhatatlanná tesz minden eddigi adatfeldolgozási módszert. Ez a dolgozat ennek a nehéz feladatnak egy kis részfeladatára próbál megoldást nyújtani.

A dolgozatban angol nyelvű szövegeknek a feldolgozását kíséreltük meg. Több okból kifolyólag esett épp erre a választás, de talán a legnyomósabbak közülük azok az angol nyelv taglaltsága, ami nagyon megkönnyíti a szövegek feldolgozását; a rengeteg már meglévő kutatás és publikáció ami egy szilárd alapot képez kiindulópontként; valamint a tény, hogy az internetes és számítógépes világban az angol nyelv a domináns, így sokkal több alkalmazása lehetséges.

A szintaktikus elemzők fognak a kutatás kiindulási pontjaként szolgálni, amikről lesz egy rövid bemutatófejezet. Majd egy saját fejlesztésű folyamat kerül tárgyalásra, ami segítségével a szintaktikus elemzők által létrehozott elemzési fákat át lehet alakítani egy számítógépek által sokkal kezelhetőbb formára. Ez az átalakítási folyamat segítségével feltárhatóak egy mondatban rejlő szemantikus entitások, valamint ezek egymás közti relációi, ami bármilyen szövegfeldolgozási feladat alapjaként szolgálhat. Ennek az eljárásnak az első két fázisa teljes mértékben saját

munka, a harmadik alapgondolatául pedig a [5, 6] cikkek szolgáltak. A dolgozat egy kiértékelés után felvázolja egy pár lehetséges alkalmazását a folyamatnak és végezetül pedig összefoglalja az leírtakat.

2. Szövegelemzők

A szövegelemzés az a folyamat, amikor egy adott bemeneti szöveget valamilyen lingvisztikai adatstruktúrává alakítunk át [1]. Maga a szövegelemzési fogalomnak elég tág értelme van, nagyon sok fajta speciálisabb elemzési folyamat létezik, ezért szükséges megkülönböztetni, hogy pontosan melyikről is van szó. Ilyenek például a morfológikus-, szintaktikus-, szemantikus- illetve társalgási szövegelemzők. Ez a dolgozat csak a szintaktikus elemzőkkel foglalkozik, ezért ezentúl szövegelemző illetve csak simán elemző alatt is szintaktikai szövegelemzőt kell érteni.

2.1 Szintaktikus elemzők

Nyelvészetben a szintaxis az egy nyelv szabályainak a gyűjteményét jelenti, vagyis definiálja, hogy a nyelv szavait miképpen lehet egymáshoz illeszteni, hogy értelmes mondatokat nyerjünk ki belőlük. Számítástechnikai környezetben ezeket a nyelvi szabályokat viszont formálisan kell specifikálni, ehhez környezetfüggetlen nyelvtanokat szoktak felhasználni. Így tehát nagyon nagy vonalakban, a szintaktikus elemzés az a folyamat, amikor egy nyelvnek egy mondatát megpróbálják visszaalakítani annak alkotóelemeire (szavak, szókapcsolatok, tagmondatok, mondatok) az előre definiált, formális szintaktikus szabályok alapján, de ugyanakkor megpróbálják az alkotóelemek közötti összefüggéseket, hierarchiát is feltárni.

Az egységesítés érdekében, a *Penn Treebank*[1] projekt leszögezte, hogy angol szövegelemzés során pontosan milyen mondattani entitást hogyan is kell jelölni, és ehhez tartja magát manapság minden szintaktikus szövegelemző. A tovább lépés előtt szükség van bár egy pár – a fent említett – fontosabb jelölésnek az ismertetésére (a teljesség igénye nélkül), amikre szükség lehet

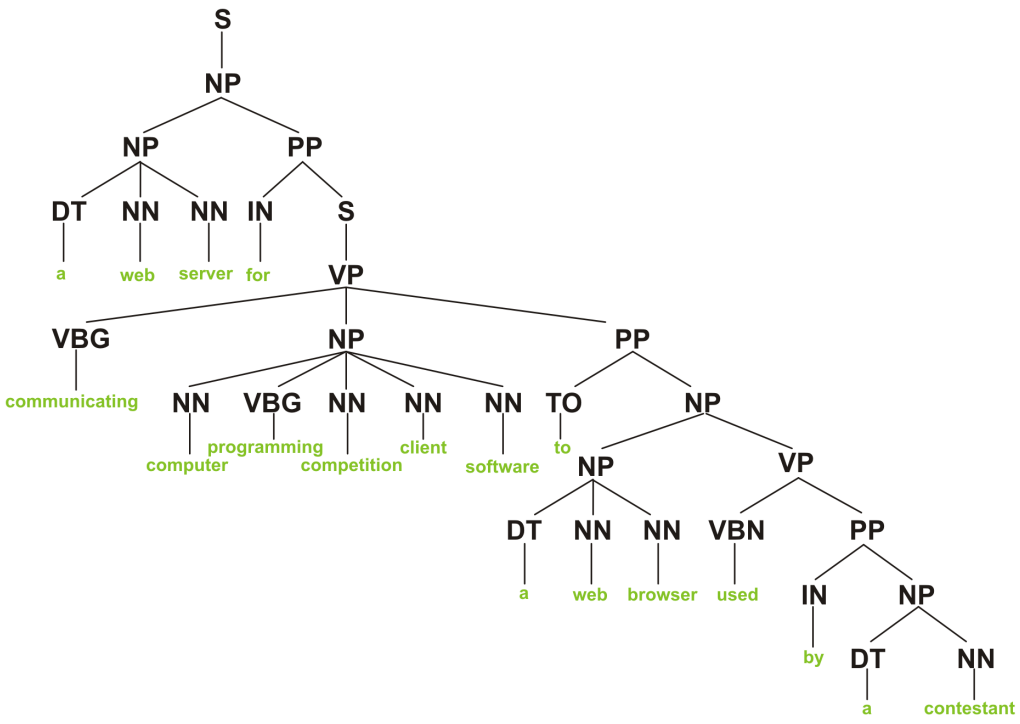
[1] http://www.cis.upenn.edu/~treebank/

a dolgozat megértése érdekében (1. táblázat).

Jelölés	Angol elnevezés	Magyar elnevezés
S	Declarative sentence	Kijelentő mondat
NP	Noun phrase	Főnévi szókapcsolat
PP	Preposition phrase	Prepozíciójú szókapcsolat
VP	Verb phrase	Igei szókapcsolat
CC	Coordinating conjunction	Mellérendelő kötőszó
DT	Determiner	Névelő
IN	Subordinating conjunction	Alárendelő kötőszó
NN	Noun, singular	Egyesszámú főnév
VB	Verb, base form	Alapformájú ige

1. táblázat. A *Penn Treebank* projekt jelölései közül egy pár fontosabb

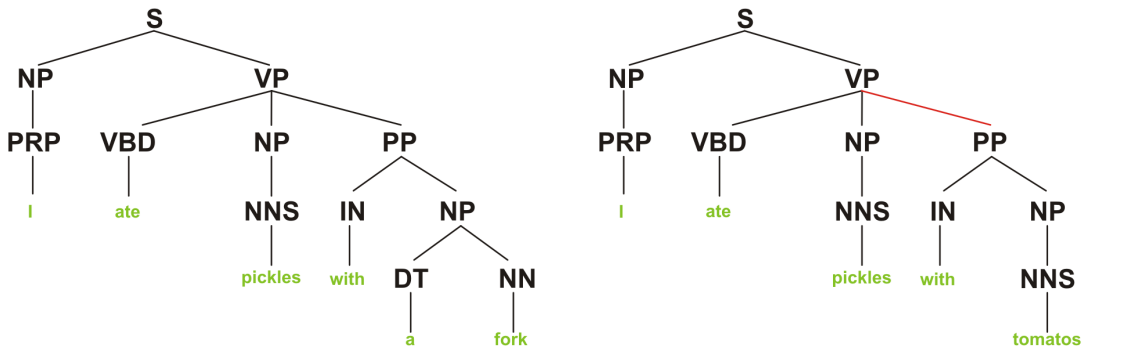
Továbbá a szemléltetés kedvéért, az 1. ábrán látható, hogy hogyan is néz ki egy lehetséges szintaktikai értelmezése egy bemeneti példa mondatnak az előbb említett jelölést használván, amiket szintaktikai elemzésfáknak illetve csak egyszerűen szintaxisfáknak is fogunk nevezni.



1. ábra. Egy példa mondat elemzése az *OpenNLP* program segítségével: *'a web server for communicating computer programming competition client software to a web browser used by a contestant'*

Mint előbb említve volt, működésileg a szintaktikus elemzők alapjául a környezetfüggetlen nyelvtanok szolgálnak, a nyelv szintaxis szabályainak egy leszűkített, formalizált változatai. Ennek az egyszerű modellnek viszont komoly korlátozásai vannak, éspedig az, hogy a nyelvek komplexitásából adódóan, illetve a nyelvtanok által megengedett végtelen kombinációk miatt egy adott mondatot jóformán korlátlan féle képp lehet értelmezni. Az elemző számára tehát problémát jelent, hogy a nyelvtan által megengedett számtalan értelmezésből melyik is a helyes.

Az egyértelműsítés problémájára a megoldás a valószínűségek bevezetése. Valószínűségi szövegelemzőknek tehát azokat az elemzőket nevezzük, amelyeknél a nyelvtanuk által definiált szabályokhoz hozzá vannak rendelve bizonyos valószínűségi értékek. Így az elemző pontosan ki tudja választani a temérdek lehetőség közül a neki betanított legvalószínűbbet. A szabályokhoz rendelt valószínűségeket automatikusan, előre megjelölt korpuszok alapján számolják, egyszerűen megvizsgálván az adott szabály alkalmazásának előfordulási számát a szöveggyűjteményben. Habár ez a megoldás sokkal jobb az előzőnél, mivel kizárólag szintaktikán alapul, ezért nem tud megkülönböztetni olyan mondatokat, amelyek szintaktikailag ekvivalensek és csupán a szavak szemantikai értelme teszi őket mondattanilag különbözővé (2. ábra).



(a) Jó elemzés

(b) Rossz elemzés, mert a '*with tomatoes*' az '*ate*' igét módosítja, pedig az ige tárgyát ('*pickles*') kellene

2. ábra. Szintaktikailag ekvivalens mondatok

Sok módszer létezik, amely az elemzés pontosságán próbál tovább javítani, viszont ezek bemutatása már a dolgozat hatáskörén kívül esne.

A szintaktikus szövegelemzők működésének illetve lehetséges hibáinak a bemutatására szükség volt, mivel a dolgozat ezeket az elkészített elemzési fákat fogja tovább átalakítani, számítógépek számára sokkal kezelhetőbb, értelmezhetőbb formátumra. Viszont a kiindulási elemzés hibáin a továbbiakban bemutatott feldolgozási lépések nem tudnak javítani, ezért a végeredmény helyessége nagymértékben függ a felhasznált szintaktikus szövegelemző pontosságától.

3. Feldolgozási lépések

A számítógépek számára a szintaktikus elemzésfák egy óriási lépést jelentettek a szövegek pontosabb megértése, illetve a belőlük való több információ kinyerése irányába. Ennek ellenére nem lehet egy programtól elvárni, hogy közvetlenül ezekkel a mondattani fákkal próbáljon dolgozni, ezekből próbáljon információt kibányászni, hiszen ezek még mindig magukban hordozzák a mondatok nyelvi megfogalmazását, amelyek a programot csak bonyolítják, ugyanakkor semmiféle többlet-információval nem rendelkeznek.

Az elkövetkezendőkben három lépés lesz bemutatva, amelyek segítségével a számítógépek számára szinte kezelhetetlen szintaktikus elemzésfákat egy tömörebb formára lehet alakítani. Ezek már nem fogják a szöveg eredeti megfogalmazására vonatkozó információkat tartalmazni, csupán a szövegben rejlő entitásokat, valamint azok közötti relációkat. Így a programok nem kell a rengeteg fölösleges információn átrágják magukat, hanem már eleve egy sokkal kompaktabb formában lévő, sokkal információdúsabb adathalmazt értelmezhetnek.

3.1 Jelölésrendszer

A természetes nyelvfeldolgozás általában entitások és a köztük fennálló relációk, összefüggések megállapításával foglalkozik, ezért a legkézenfekvőbb ábrázolásmód gráfok segítségével szokott megvalósulni. A tömörség kedvéért *Lisp* listákkal is szoktak dolgozni, de a nehezen való átláthatóságuk miatt e dolgozatban a gráfos megoldás lesz alkalmazva.

Az itt bemutatott módszerek alkalmazása során a feldolgozandó gráfban különböző képp lesznek bizonyos csúcspontok jelölve, viszont, hogy ne kelljen minden új bevezetett fogalomnál jelöléseket is tárgyalni, ezért ezek megtalalálhatóak összesítve a függelék 1. táblázatában.

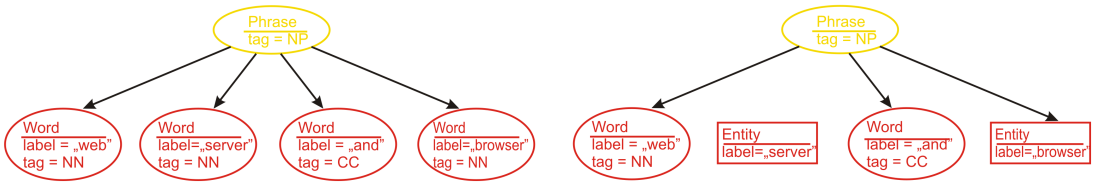
3.2 Szemantikus és metaelemek kiválogatása

A szintaxisfákat egyszerűsítő folyamat talán legfontosabb – ugyanakkor egyszerűbb – része a szemantikai elemek kiválogatása a szintaktikai elemzésből. Ez a kiválogatás három fajta szemantikus elemet határoz meg: entitásokat, amik előfordulnak a mondatban (angolul: *entity*); entitások közötti cselekvéseket (angolul: *action*); illetve az entitásokat megváltoztató módosítókat (angolul: *modifier*). Az imént felsoroltakon kívül még két fajta metaelem is meg lesz határozva, amik a pontosabb feldolgozást fogják elősegíteni: referenciák (angolul: *reference*) fogják megjelölni azokat az entitásokat amelyek nem azon a ponton vannak definiálva, hanem visszautalások egy már korábban említett dologra; illetve kötéseket (angolul: *binding*) amik többértelmű relációknak felelnek meg és pontos értelmezésük bonyolultabb eljárást igényel.

3.2.1 Entitások kiválasztása

Az entitások képezik egy mondatnak az alanyait illetve tárgyait, tehát szófailag a főnevek lesznek az entitás jelöltek, amik a szintaxisfában alapértelmezetten a főnévi szókapcsolatokon belül fordulnak elő. A feladat ebben a lépésben tehát az, hogy az elemző által főneveknek nyilvánított elemek közül kiválasszuk az entitásokat. Ehhez két észrevétel fontos: egy entitás előtt előforduló főnevek nem entitások, hanem a már kiválasztott entitás módosítói (ezekről később lesz szó), tehát egy főnévi szókapcsolatban mindig az utolsó főnév lesz az képviselt entitás. A második észrevétel, hogy egy főnévi szókapcsolat a szintaxisfában akkor és csakis akkor tartalmaz két entitást, ha ezek mellérendelő viszonyban vannak és jelen van egy megfelelő elválasztó szó vagy írásjel (’*és*’, ’*vagy*’, vessző). Ezt a két megfigyelést szem előtt tartva az entitások kiválasztása egyszerűen a főnévi szókapcsolat felbontása entitás csoportokra a mellérendelő viszonyok alapján, majd mindegyik csoport utolsó főnevének átalakítása entitás elemmé.

A 3. ábrán látható egy szemléltető példa entitás kiválasztásra, ami egy főnévi szókapcsolaton belül válogatja ki, hogy mi számít entitásnak és mi nem. Érdemes megfigyelni, hogy ez az egyszerű példa magába foglalja az összes említett megfigyelést.

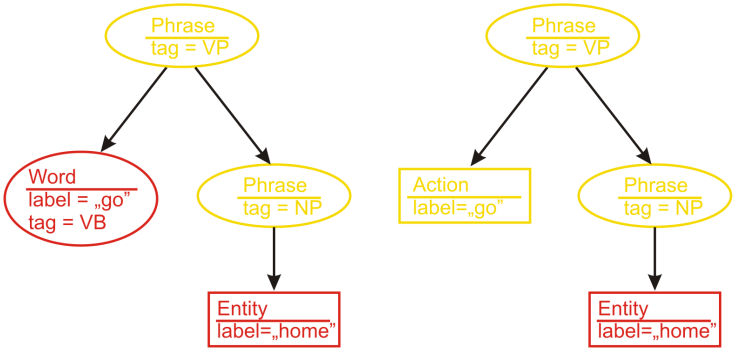


(a) Kiválasztás előtt (b) Kiválasztás után

3. ábra. Entitások kiválasztása a '*web server and browser*' szókapcsolatból

3.2.2 Cselekvések kiválasztása

Az entitások közti cselekvéseket szófailag az igék határozzák meg, ezek pediglen a szintaxisfában az igei szókapcsolatokban jelennek meg. Mivel ezekbe az igei csoportokba a szintaktikus elemzők mindig csak egy igét helyeznek el, ezért a kiválasztás mondhatni banális, egyszerűen a megfelelő csoportokban tartózkodó igéket át kell alakítani cselekvés elemekre. Nagyon fontos viszont megjegyezni, hogy mivel az elemzők valószínűségeken alapszanak, előfordulhat, hogy egy szóhoz igei szófajt rendelnek viszont nem teszik igei szókapcsolatba, ami nagyjából ekvivalens azzal, hogy a az a bizonyos szó nem ige az aktuális szintaxisfában. Szemléltetésképpen a 4. ábrán megtalálható egy leegyszerűsített példa.



(a) Kiválasztás előtt (b) Kiválasztás után

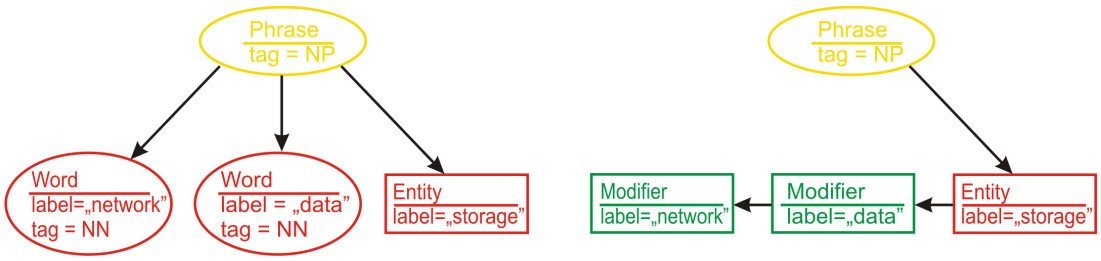
4. ábra. Cselekvések kiválasztása a '*go home*' szókapcsolatból

3.2.3 Módosítók kiválasztása

Szövegek elemzése során általános jelenség, hogy az entitások nagyon ritkán állnak önmagukban, legtöbbször más szavak is csatolódnak hozzájuk, amelyek valamilyen módon leszűkítik a

konkrét értelmét az entitásnak (például '*web böngésző*'), de ugyanakkor akár több ilyen elem is megjelenhet (például '*hálózati adat-tároló*').

Az entitáskiválasztás tárgyalásánál említve volt már, hogy nem minden főnév számít entitásnak, kizárólag az entitás csoport utolsó főneve. A többi főnév ellenben a csoportban megjelenő entitás értelmét egy specifikusabbra változtatja, ha több van, akkor mindegyik még tovább specializálja azt. Ezeknek a kiválasztása tehát egyszerűen annyiból áll, hogy az entitások meghatározása után mindegyik előtte szereplő főnevet be lehet sorolni, mint módosítója az adott entitásnak. Példa erre a többszörös leszűkítésre a 5. ábrán látható.



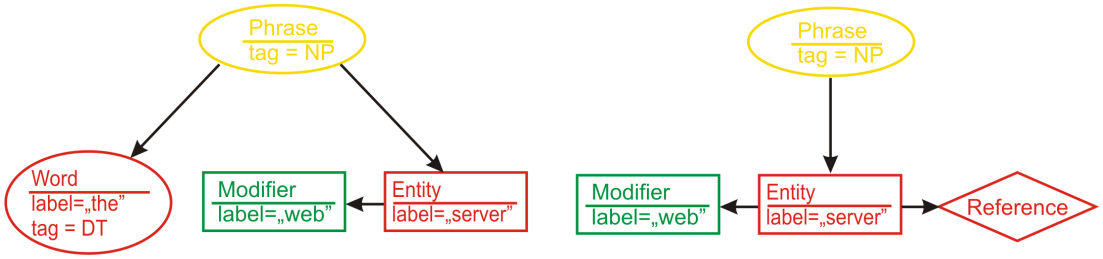
(a) Kiválasztás előtt (b) Kiválasztás után

5. ábra. Entitásmódosítók kiválasztása a '*network data storage*' szókapcsolatból

3.2.4 Referenciák kiválasztása

Bármilyen természetes szövegben egy entitást csak egyszer definiálnak, az összes többi esetben ahol szükség volna ugyanarra a bizonyos entitásra egyszerűen visszautalnak rá. Nagyon egyszerű módszerektől kezdve nagyon bonyolultakig minden féle visszautalás lehetséges, viszont mivel ez már önmagában egy különálló kutatási terület, ezért jelen dolgozatban az összes közül csak egy lesz tárgyalva.

Az entitás visszautalások közül talán a legegyszerűbb, amikor egy határozott névelővel van megvalósítva a hivatkozás. (például '*az alma*' vagy angolul '*the apple*'). Az utalás teljes kulcsa a '*the*' szavacskában rejlik, mert az angol nyelvben ez az egyetlen határozott névelő. Ebből adódóan a referenciák kiválogatása végett meg kell vizsgálni az összes főnévi szókapcsolatot, és amelyik tartalmazza a '*the*' névelőt arról biztosan lehet tudni, hogy utalás egy már korábban említett dologra. Egy ilyen kiválasztásra példa megtalálható a 6. ábrán.



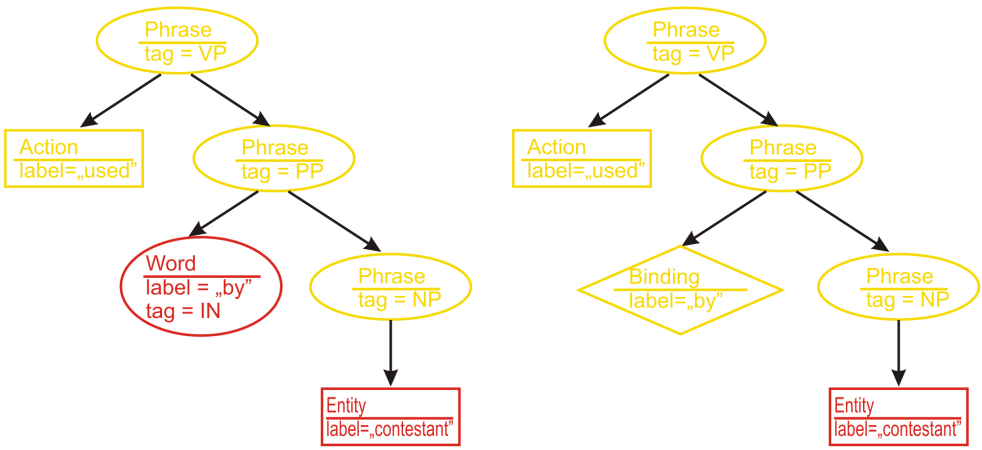
(a) Kiválasztás előtt (b) Kiválasztás után

6. ábra. Referenciák kiválasztása a ’*the web server*’ szókapcsolatból

3.2.5 Kötések kiválasztása

Számítógépek számára a szövegfeldolgozás során az egyik legnehezebb feladat az a prepozíciók helyes értelmezése. Erre a megoldás a későbbiekben lesz bemutatva, viszont ahhoz, hogy tudjuk őket értelmezni, előbb szükség van azokat is meghatározni a szintaxisfa alapján.

Mivel a lehetséges prepozíciók számossága igen kicsi, ezért a szintaktikai elemzők mindig tökéletes pontossággal meg tudják határozni őket. Így számunkra nagyon egyszerű a feladat, csupán meg kell keresni a szintaxisfában a megfelelő címkével ellátott szó elemeket és lecserélni őket kötés metaelemekre (megkeresni a pontos értelmüket ezen a ponton túlságosan komplikált lenne, ezért itt csak metaelemre cseréljük őket, majd csak később a végleges relációra). A 7. ábra bemutat egy példát erre az eljárásra.



(a) Kiválasztás előtt (b) Kiválasztás után

7. ábra. Kötések kiválasztása a ’*used by a contestant*’ szókapcsolatból

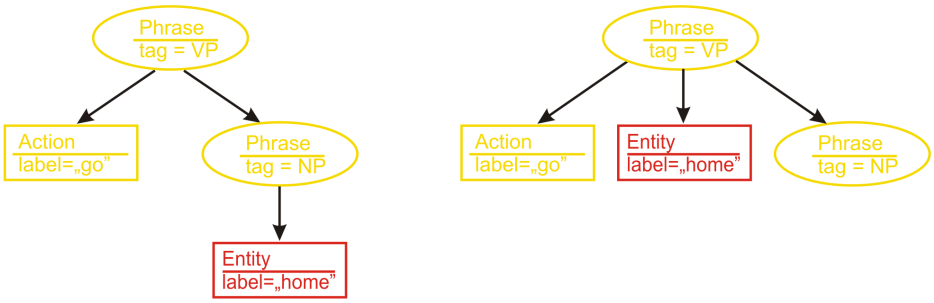
3.3 Szintaxisfa összevonása

Az előző részben bevezetett módszerek segítségével – jó esetben – sikerült a szintaktikai elemző által nyújtott – megjelölt – szintaxisfában minden szó elemet lecserélni egy speciálisabb, kategorizált elemre (entitás, cselekvés, módosító, referencia illetve kötés). A feldolgozandó fa így az előbb kiválasztott elemeken kívül már csak mondat (angolul '*sentence*'), tagmondat (angolul '*clause*') és szókapcsolat (angolul '*phrase*') csúcspontokat tartalmaz. Ezek az elemek még mindig meggátolják a programokat, hogy magával az információval foglalkozzanak, tehát a továbbiakban ezeknek a kiküszöböléséről lesz szó.

Mint az eddigiekből kiderülhetett, minden fajta szókapcsolatnak megvan a saját domináns elemtípusa, amelyik (vagy amelyikek) helyét képviseli (főnévi szókapcsolatok entitásokat, igei szókapcsolatok cselekvéseket valamint prepozíciójú szókapcsolatok kötéseket képviselnek). Ez a megfigyelés azzal egyenértékű, hogy bármilyen szókapcsolat helyettesíthető az általa képviselt elemekkel, vagyis ha egy domináns elemet a szókapcsolattal egy szintre hozzuk a fában, akkor az új fa ekvivalens marad a régivel.

Az egyetlen dolog, amire vigyázni kell az előbbi elgondolásban, hogy egy szókapcsolat tartalmazhat nem domináns elemeket is. Ezek a másodlagos elemek a fában nem egy felsőbb szinthez csatolódnak, hanem az aktuális szinten lévő elsődleges elemekhez, azoknak az értelmét egészítik ki, avagy azokhoz rendelnek plusz relációkat. Így ahhoz, hogy a főelemek felfele mozdítása valóban ekvivalens fát eredményezzen, a másodlagos elemeket először át kell csatolni a domináns elemekhez.

A két fenti gondolatmenet alapján az összevonási módszer három egyszerű műveletre redukálódik: (1) Másodlagos elemek átcsatolása a domináns elemekhez, (2) A domináns elemek felvitele az őket tartalmazó szókapcsolattal egy szintre (3) Az üresen maradt szókapcsolat elemek törlése. A folyamat maga pedig egyszerűen ezeknek a műveleteknek az egymás utáni alkalmazása, amíg van amire. Egy példa e három műveletre megtalálható a 8. ábrán.



(a) Összevonás előtt

(b) Domináns elem felvitele

(c) Üres szókapcsolat elem törlése

(d) Másodlagos elem átkötése

8. ábra. A ’*go home*’ összetett szókapcsolat szintaxisfájának összevonása

3.4 Relációk értelmének tisztázása

Egy nyelv mondatának a szintaktikai felépítéséből nagyon sok információt ki lehet nyerni a benne rejlő entitásokról illetve ezek közötti relációkról. Bizonyos relációk viszont nem a mondattani szerkezetből adódnak, hanem kötőszavak, prepozíciók értelme illetve ezek kontextusa biztosítja a megfelelő viszonyt. Ez egy elég nagy problémát von maga után, mivel [3] alapján az oxfordi angol értelmező szótár (*New Oxford Dictionary of English*) összesen 373 prepozíciót, és ezekhez 847 hozzárendelt értelmezést tartalmaz. Ezekből 218 többszavas szókapcsolat (például ’*by means of*’) amik egyetlen értelemmel rendelkeznek, így a megmaradt 155 prepozícióra összesen 629 értelem jut, vagyis átlagosan 4,06. Ugyanakkor a leggyakoribb prepozíciók között létezik olyan is (’*on*’), amelynek 25 különböző értelme van.

A fenti értékekből világosan látszik, hogy egy számítógépes program önmagában nincs ahonnan eldöntse, hogy mi a helyes értelme egy ilyen relációnak, mivel nincs ahogy értelmezze a reláció kontextusát amiből adódik a megfelelő értelem. A probléma megoldása végett valószínűségi modellek irányába kell fordulni, vagyis egy olyan modellre van szükség ami a mondatban rejlő – számítógépek által kinyerhető – információ alapján egy jó becslést tudna adni, hogy milyen relációról van szó. A [5, 6] javaslataiból kiindulva egy maximum entrópia modell

segítségével próbálkoztunk eredményeket elérni.

Ugyanakkor, ha valószínűségekről van szó, akkor nyilván kell egy elegendően nagy szöveg korpusz, ami alapján fel lehet állítani a modellt és meghatározni a valószínűségi értékeket. A cél elérése érdekében a dolgozat a *The Preposition Project*[2] által a *SemEval-2007*[3] műhelyprogramra nyújtott adatbázist fogja felhasználni, amely 24668 megjelölt mondatot, a 34 leggyakoribb angol prepozíciót illetve ezeknek 247 értelmét tartalmazza.

3.4.1 Maximum entrópia modell

Nagyon kötetlenül fogalmazva, a maximum entrópia modellek arra használatosak, hogy bizonyos kikötések alapján felállítsanak egy valószínűségi változót, ami a lehető legegyenletesebben írja le egy megfigyelt sztochasztikus folyamat kimenetét. A maximum entrópia módszernek az a feladata, hogy egy nagy adathalmazban látott bemeneti adatok – úgynevezett jellemzők (angolul '*features*') – illetve kimeneti adatok alapján próbálja levezetni, hogy a folyamat milyen szabályoknak, kikötéseknek próbál megfelelni, vagyis, hogy mik a mi mesterséges modellünk paraméterei. Így a meghatározott paraméterek segítségével egy valószínűségi modellhez jutunk, amit majd a későbbiekben fel lehet használni ismeretlen jellemző-kombinációk lekezelése érdekében.

Ahhoz, hogy maximum entrópia modellt lehessen felállítani a prepozíciós relációk értelmének megállapítására, szükség van egy bemeneti-kimeneti adathalmazra, ahol a bemeneti adatok jellemzők, illetve a kimeneti adatok pedig reláció értelmek (például cél reláció). Az előbb már említett *The Preposition Project* adathalmazát fel lehetne használni erre a célra, viszont abban a kimeneti reláció-értelmekhez bemeneti adatként mondatok vannak hozzárendelve. Tehát a feladat az, hogy a bemeneti mondatokat át kell alakítani jellemző halmazokra, amik alapján majd el lehet készíteni a modellt.

Ilyen feladatok esetében a standard megoldás abból szokott állni, hogy először elkészítenek egy hatalmas jellemző adatbázist (milliós nagyságrendben), ami magába foglal minden féle fajta jellemzőt amit ki lehet nyerni a feldolgozandó szöveggyűjteményből. Ez után speciális módszerek, algoritmusok segítségével megpróbálják ezt az óriási adathalmazt leredukálni tízezres

[2]http://www.clres.com/prepositions.html
[3]http://nlp.cs.swarthmore.edu/semeval/

nagyságrendre, ami már kizárólag azokat a jellemzőket tartalmazza, amik a legjobban befolyásolták a végeredményt az eredeti szöveggyűjteményben. Végezetül pedig ezek a kiválasztott jellemzők alapján elkészítik a végső maximum entrópia modellt amit majd fel fognak használni a későbbiekben.

Ezeket a standard módszereket jellemző kiválasztásnak nevezik – angolul '*feature selection*' – és önmagukban is teljes kutatási témakörnek felelnek meg. Mivel ez a megoldás nagyon kívül esne a dolgozat hatókörén, ezért itt a jellemző kiválasztás nem automatikus módszerekkel történt, hanem kézileg volt megvalósítva, kipróbálván különböző jellemző kategóriákat illetve kombinációjukat és kiértékelvén, hogy melyik milyen pontosságot eredményez.

A kiértékelés végett a szöveg korpusz két részre van osztva: egy 16557 mondatból álló kiképző halmazra, ami alapján készül el a modell; illetve egy 8111 mondatból álló teszt halmazra, amely segítségével a modell pontosságát lehet leellenőrizni. A kimeneti adatoknak két változatuk van, durva értékelés céljából az egyik változat csak 148 értelmet rendel hozzá a prepozíciókhoz, finom értékelés céljából mind a 247 értelem fel van használva. A különböző kipróbált jellemző kategóriák kiértékelése a 2. táblázatban látható, illetve a legjobb közülük összehasonlítva a *SemEval-2007* során elért eredményekkel [2] megtalálható a 3. táblázatban.

Kiválasztott jellemzők	**Durva értékelés**	**Finom értékelés**
Prepozíció önmagában	49,07%	39,21%
Prepozíció és a mondatban megjelenő szófajok halmaza	44,60%	34,49%
Prepozíció és a mondatban megjelenő szavak hamaza	52,11%	42,41%
Prepozíció és a mondatban megjelenő szótövek halmaza	55,00%	45,37%
Prepozíció és a mondatban megjelenő szótövek halmaza, mindegyik megjelölve, hogy a prepozíció előtt vagy után fordul elő	68,66%	59,89%

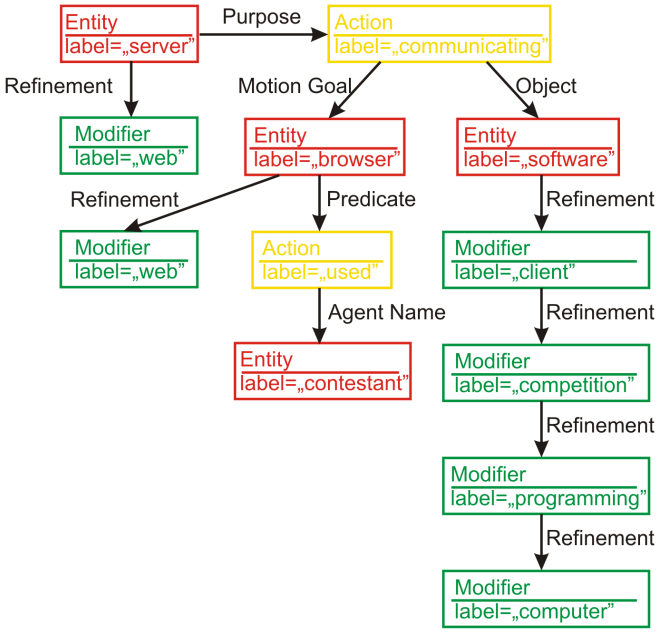
2. táblázat. Különböző jellemző kategóriák összehasonlítása

Résztvevő csapatok	Durva értékelés	Finom értékelés
University of Melbourne	75,50%	69,30%
Az itt leírt módszer	**68,66%**	**59,89%**
Koç University	64,20%	54,07%
Instituto Trentino di Cultura, IRST	61,00%	49,60%

3. táblázat. Összehasonlítás a *SemEval-2007* eredményeivel

4. Értékelés

A dolgozatban bemutatott módszerek a szintaktikai elemzők által – egy szöveg alapján – létrehozott szintaxisfát alakították tovább egy kompaktabb formára. Ez az új forma sokkal kedvezőbb számítógépes programok számára, hiszen ezekből a gráfokból kiindulva a programozók már eleve a megoldandó problémára tudnak koncentrálni, nem kell a nyelvi és nyelvtani elemekkel foglalkozniuk. Szemléltetésképpen, hogy a módszerek valóban életképesek, a dolgozat elején példaként felhozott szintaxisfa (1. ábra) feldolgozott változata megtalálható a 9. ábrán.



9. ábra. Egy példa mondat végső gráfja a bemutatott folyamat alkalmazása után: *'a web server for communicating computer programming competition client software to a web browser used by a contestant'*

Habár a bemutatott eljárás működőképes, tökéletesnek nem nevezhető. Az elkövetkezendő pár bekezdésben erről lesz bővebben szó, valamint az esetleges javítási lehetőségekről.

4.1 Pontosság

Az itt bemutatott eljárás a szintaxisfákat veszi alapul, ezért nem eredményezhet jobb megoldást mint amit a szintaktikus elemzők megengednek. Vagyis a módszerek nem képesek az eredeti fában lévő esetleges hibákat korrigálni, ahhoz már szemantikusan kéne értelmezni a teljes szöveget. Következményképpen az itteni folyamat képes összevonni bármilyen hibás szintaktikai elemzésfát, viszont a végeredmény tartalmazni fogja az összes olyan rossz értelmezést, amit az elemző eredetileg elkövetett.

A legtöbb szintaktikai elemző valószínűségeken alapszik, és ebből adódóan lehetőséget ad, hogy egy adott szövegnek ne csak az általa hitt legvalószínűbb elemzésfáját adja meg, hanem jóformán akárhányat. Egy pár plusz feltétel, ellenőrzés bevezetésével elég sok egyértelmű félreértelmezést el lehetne kapni, valamint az előző megfigyelés alapján megvan arra a lehetőség, hogy a hibás fát figyelmen kívül hagyjuk, és a következő legvalószínűbbet próbáljuk feldolgozni.

4.2 További tennivalók

A három leírt feldolgozási lépés közül kettő kézileg szerkesztett szabályokon alapszik. Kézileg viszont nagyon nehéz és hosszadalmas minden lehetőséget lekezelni, így mindig maradnak bizonyos hiányosságok a szabályrendszerben, ami miatt egy-egy speciális szerkezetet vagy lekezeletlen új adattípust nem fog tudni feldolgozni a program. Habár a bemutatott módszerek bonyolultabb és hosszadalmasabb szöveget is fel tudnak dolgozni, egy jelenlegi hiányosságuk, hogy számokkal, értékekkel nem foglalkoznak még.

A problémára két megoldás is lehetséges. Az egyszerűbb, az kézileg betömni a szabályrendszerben a hiányosságokat. Az előnye, hogy egy megbízható és pontos feldolgozásra lehet számítani, a hátránya viszont hogy kézileg minden lehetőséget lekezelni lehetetlen, és nem is

érdemes egy adott ponton túl. A nehezebb, az felállítani egy valószínűségi modellt, ami garantálná, hogy nem lesznek meglepetések, ami lekezeletlen típusokat illeti, viszont ehhez szükség van egy megfelelően megjelölt szöveg korpuszra, amiknek az előállítása pénz- és időigényes, a beszerzése nehézkes, valamint tudtom szerint ennek a dolgozatnak megfelelő nem létezik még.

5. Alkalmazás

A természetes szövegfeldolgozás egy nagyon népszerű kutatási témakör, mert rengeteg érdekes alkalmazása van és ugyanakkor csak most kezdenek igazán a számítógépek elegendően erősek lenni, hogy meg tudjanak felelni az elvárásoknak. Éppen ezért a természetes interfészek kutatása eléggé fellendülőben van.

Eddig a kereső motorok (*Google*™, *Yahoo!*™, *MSN*™, stb.) kizárólag kulcsszavak alapján működtek, viszont manapság egyre inkább kezdenek új keresők megjelenni (*Ask*™, *Powerset*, stb.), amik próbálkoznak természetes lekérések alapján keresni. Tehát a felhasználó felteszi a kérdését, amire a kereső motor majd megpróbálja megkeresni a választ. Ilyen keresők esetében viszont nagyon fontos, hogy minél pontosabban értelmezzék mind a felhasználó kérését, mind pedig az indexelendő adathalmazt, valamint a benne rejlő információt és összefüggéseket, aminek a segítségére szolgálhat az itt bemutatott folyamat.

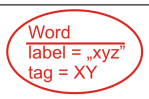
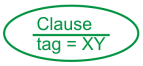
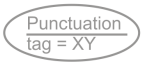
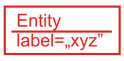
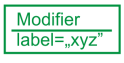
Érdemes megfigyelni, hogy a bemutatott lépések mind invertálhatóak, ami egy nagyon érdekes alkalmazását teszi lehetővé a folyamatnak, jobban mondva a folyamat fordítottjának: számítógépek által összeállított entitások és a közöttük lévő relációk visszaalakítása felhasználók számára természetes szöveggé.

6. Összefoglalás

A témába való rövid bevezetés után a dolgozat betekintést nyújtott a szintaktikus elemzők világába illetve ezek működési elveibe. Ezeket követően sor került egy egyszerűsítési folyamat leírására, amely segítségével a bonyolult szintaxisfákat egy sokkal tömörebb formára lehetett alakítani. E cél elérése érdekében két teljesen saját fejlesztésű feldolgozási fázis volt bemutatva (*szemantikus és meta elemek kiválogatása* illetve *szintaxisfák összevonása*), valamint egy harmadik fázis a többértelműség tisztázására, amely alapötletét a [5, 6] szolgáltatták. Egy rövid értékelés során meggyőződhettünk róla, hogy a kidolgozott eljárás valóban egy életképes megoldása a problémának, ugyanakkor még akad bőven tennivaló a tökéletesítés érdekében. Végezetül egy pár új gyakorlati alkalmazás került bemutatásra, amiből körvonalazódott, hogy egy jelenleg kibontakozó kutatási területről van szó, aminek komoly jövője lehet a számítástechnika világában.

A. FÜGGELÉK

Jelölések és ábrázolások

Ábrázolás	Magyarázat
	Szintaktikus elemző által meghatározott szó elem, maga a szó (*label*) illetve annak szófaja (*tag*)
	Szintaktikus elemző által meghatározott szókapcsolat elem illetve a típus (*tag*), hogy milyen szófaj helyét tölti be
	Szintaktikus elemző által meghatározott tagmondat elem illetve annak típusa (*tag*)
	Szintaktikus elemző által meghatározott mondat elem
	Szintaktikus elemző által meghatározott írásjel elem illetve annak típusa (*tag*)
	Jelölés, hogy a gráf csomópont, amihez kapcsolódik az csak egy utalás egy már előbb definiált entitásra
	Többértelmű reláció, magába foglalván az eredeti szavat (*label*) amiből származik a többértelműség
	A dolgozat által meghatározott szemantikus entitás, illetve elnevezése (*label*)
	A dolgozat által meghatározott entitások közötti cselekvés, illetve elnevezése (*label*)
	A dolgozat által meghatározott entitás- vagy cselekvésmódosító, illetve elnevezése (*label*)

1. táblázat. A dolgozat által felhasznált szintaktikus- (ellipszis), meta- (rombusz) és szemantikus (téglalap) elemek fontosság szerint rangsorolva (piros, sárga, zöld, kék és szürke)

Irodalomjegyzék

[1] Daniel Jurafsky, James H. Martin: *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition, Second edition*. 2008.

[2] Ken Litkowski, Orin Hargraves *SemEval-2007 Task 06: Word-Sense Disambiguation of Prepositions*. In proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 24–29. Prague, June 2007.

[3] Kenneth C. Litkowski: *Digraph Analysis of Dictionary Preposition Definitions*. In proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions, pages 9-16. Philadelphia, July 2002.

[4] Peter Lyman, Hal R. Varian, Kirsten Swearingen, Peter Charles, Nathan Good, Laheem Lamar Jordan, Joyojeet Pal: *How Much Information?*, October 2003.

[5] Patrick Ye, Timothy Baldwin: *Preposition Sense Disambiguation Using Rich Semantic Features*. In proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 241–244. Prague, June 2007.

[6] Patrick Ye, Timothy Baldwin: *Verb Sense Disambiguation Using Selectional Preferences Extracted with a State-of-the-art Semantic Role Labeler*. In proceedings of the 2006 Australasian Language Technology Workshop (ALTW2006), pages 139–148. Sydney, November 2006.