A természetesnyelv-feldolgozás fizikai és nyelvi határai

Mészáros Evelin

Clementine
1115 Budapest Hungary
emeszaros@clementine.hu

Kivonat: A természetesnyelv-feldolgozó rendszerek fejlettsége mára már elért egy olyan szintet, hogy lehetőségünk nyílik különböző szoftverek segítségével olyan rendszereket építeni, amelyek megpróbálják értelmezni a feldolgozandó szöveges információt. A rendelkezésünkre álló eszközök egyre növekvő tárháza, amelyhez a nyílt forráskódú programnyelvek térhódítása is hozzájárul, lehetővé teszi azt a helyzetet, hogy egy másik idegen nyelvre írt módszertant a saját anyanyelvünkön implementáljunk. Ennek azonban lehetnek nehézségei, melyek a feldolgozandó nyelv jellegzetességeitől függenek, erre viszont a legtöbb esetben a szerzők nem adnak útmutatást. Jelen tanulmány ilyen útmutatásokat kíván nyújtani, vagyis hogyan tudunk egy angol nyelvre speciálisan kifejlesztett módszertant átültetni magyar nyelvre?

Kulcsszavak: szemantikai hasonlóság, ontológia, NLP

1 Bevezetés

A természetesnyelv-feldolgozáson belül többféle alkalmazási területen lehet szükség két szöveg összehasonlítására. Számos forrás áll rendelkezésünkre, amelyben egy ilyen feladat megoldására javasolnak módszereket a szerzők: ezek közül megemlíthetjük a kulcsszókinyerést vagy akár a kivonatolást, de ezek a metódusok rövid szövegek összehasonlításánál kudarcot vallanak, hiszen nagyon kicsi az esély arra, hogy az adott rövid leírásban szerepel akár egy kulcsszó is. Ebben az esetben szükségünk van egy olyan módszertanra, amely túllép az egyszerű gyakoriságalapú megközelítésen, és figyelembe veszi az adott nyelv struktúráját és a benne szereplő szavak egymáshoz kapcsolódó viszonyait is.

Ahhoz, hogy a gyakoriságok segítségével jól tudjuk definiálni egy adott szó környezetét, hatalmas méretű témaspecifikus korpuszra van szükségünk az adott nyelven – ennek hiánya, illetve idő- és szaktudásigényes előállítása az oka a hagyományos értelemben vett szóvektorokkal való ábrázolás elvetésének is –, ezért lehet egy jó megközelítés a nyelv hálózatként való definiálása. A hálózati megközelítés a nyelvfeldolgozás és a statisztika területén is megjelent, és számos tanulmány mutat rá arra, hogy a jelentésbeli hasonlóságot egy egész rendszerhez viszonyítva érdemes meghatározni. Ehhez azonban szükség van egy forrásra, amely tartalmazza a szavak egymáshoz viszonyított

kapcsolatát is – erre lehet egy jó kiindulási pont az ún. Wordnet[1], amely számos nyelven elérhető többé-kevésbé kidolgozott formában.

A magyar nyelv helyzete nagyon egyedi – agglutináló nyelv révén a toldalékok azonosítása korántsem olyan triviális, mint egyéb (pl. angol) nyelvek esetén, ahol a toldalékok nem a szavakkal egybeírva, hanem többnyire azok előtt különállóan helyezkednek el. További nehézséget jelenthet a szabad szórend is szemben egyéb nyelvek (pl. angol) kötött szórendjével.

A nyílt forráskódú szoftverek terjedésével együtt az autodidakta nyelvészek és adatelemzők is számos opciót mutatnak arra vonatkozóan, hogy hogyan tudjuk két szövegrészről eldönteni, hogy hasonló jelentésűek-e vagy sem. Ezeket a módszereket széles körben alkalmazzák sokféle tudásbázisalapú alkalmazásnál, egyéb információkinyerő rendszerek esetén, érzelemdetektáláshoz vagy akár a biostatisztikában is (Slimani, 2013). A lehetőségek tárháza azt a látszatot keltheti a kutatóban, hogy ezeket a modelleket és forrásokat ugyanebben a formában lehetséges reprodukálni.

A tanulmányban azt szeretném bemutatni, hogy a számos nagy világnyelvekre (legfőképpen angolra) készült szemantikai hasonlóságalapú módszerek milyen szempontok figyelembevételével ültethetők át egy adott nyelvre. Az írás második fejezetében a szemantikai hasonlóság elméleti hátterét és számításának néhány alternatíváját fogom részletezni, majd a következő fejezetben a felhasznált módszertant fogom leírni. A harmadik fejezetben pedig iránymutatást szeretnék adni arra vonatkozóan, hogy milyen szempontokat érdemes figyelembe venni egy idegen nyelvű módszertan magyar nyelvre való átültetésekor.

2 Elméleti háttér

2.1 Szemantikai hasonlóság

A bevezetőben utaltam már rá, hogy számos területen szükségünk lehet arra, hogy két szövegről eldöntsük, jelentésükben hasonlítanak-e egymásra vagy sem. Például ha egy szöveg koherenciáját szeretnénk megállapítani (Lapata és Barzilay, 2005), vagy ha egy gépi fordítás eredményét szeretnénk automatikusan értékelni (Papineni és mtsai, 2002), de akkor is, ha egy szövegnek szeretnénk a lényeges tartalmát kinyerni automatizáltan (Salton és mtsai, 1997). A mesterséges intelligencia térhódításának korában azonban akár egy párbeszéd összeállításánál is felmerülhet az igény az alkalmazására, vagy kérdés-válasz párok esetében is hasznos lehet ez a tudás.

A szemantikai hasonlóság legegyszerűbb megközelítése szerint a hasonlósági mutatót aszerint számoljuk, hogy mennyi azonos szó szerepel a két összehasonlítandó szövegben. (Mihalcea és mtsai, 2006). Számos továbbfejlesztésre történt kísérlet további szempontok bevonásával mint például a szótövezés, stopszó-eltávolítás, szófaji egyértelműsítés, leghosszabb összeillő tagmondatpár megválasztása, illetve egyéb súlyozó és normalizáló tényezők figyelembe vétele (Salton és mtsai, 1997).

[1] https://wordnet.princeton.edu

A legtöbb esetben ún. word-to-word összehasonlítások esetén a szövegben szereplő szavakat vesszük elemzési egységnek és ezeket hasonlítjuk össze egymással, viszont sokszor szükséges lehet az is, hogy többtagú kifejezéseket is összetartozónak vegyünk.

Jelen tanulmány elméleti hátterét egy olyan írás (Li és mtsai, 2006) adja, amely korpusz statisztikák és szemantikai hálózatok együttes figyelembevételével számolja ki két rövid szöveg közötti hasonlóság mértékét. Egyetértek az említett szerzőkkel abban, miszerint három fő hátránya van a szemantikai hasonlóság kiszámításának, amelyek egyben alkalmazási nehézségnek is felfoghatók (Li és mtsai, 2006):

- egy mondat szavai csak nagyon sokdimenziós vektortérben ábrázolhatók, és ez alacsony modellteljesítménnyel társulhat
- a legtöbb módszertan a kutató intenzív beavatkozását igényli a szövegelőkészítés folyamatában
- egy létrehozott modell nem adaptálható könnyen egyéb témakörökre

A szerző (Li és mtsai, 2006) megfogalmazza azt az igényt, miszerint egy hatékony szövegösszehasonlító módszer csak a mondatok értelmére koncentráljon, képes legyen automatikusan bővülni a kutató kézi beavatkozása nélkül (vagy csak korlátozott kutatói beavatkozással), és könnyen adaptálható legyen egyéb témakörökre is.

Hangsúlyozom, hogy jelen tanulmányban bemutatott módszertan kifejezetten rövid szövegek összehasonlítására mutatott hatékony eredményeket mind angol mind pedig magyar nyelven, hosszú szövegek esetén egyéb módszerek is hatékonynak bizonyulhatnak, melyekre terjedelmi korlátok miatt itt nem térek ki.

A következő alfejezetben a szemantikai hasonlóság kiszámításának azt a módját fogom ismertetni, melyet a későbbiekben leírt módszertanhoz is alkalmazok.

2.2 Szövegből adat - a szöveg reprezentációja

Egy szöveg vagy mondat matematikai modellje egy vektor, mely a tisztított (stopszómentes, lemmatizált) mondatban szereplő szavak előfordulását vagy előfordulási gyakoriságát tartalmazza.

Két mondat az 1. táblázatban látható módon ábrázolható vektorként:

1. táblázat: Mondatok vektorreprezentációja

	pénzintézet	van	hitel	bank	rendelkezik	kölcsön
A Egy pénzintézetnél van hitele. [	1	1	1	0	0	0]
B Egy banknál rendelkezik kölcsönnel. [	0	0	0	1	1	1]

Két szöveg (ebben az esetben mondat) hasonlóságának, távolságának legelterjedtebb mértéke a koszinusz távolság, melyet az 1-es képletben szereplő módon lehet kiszámítani, és amelyben A és B a mondatokat leíró vektorok.

$$Similarity(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|} \quad \textbf{(1)}$$

Ha a két mondatban nincsenek közös szavak, akkor a hasonlóság értéke 0, azaz a két mondat ezen mérték szerint egyáltalán nem tekinthető hasonlónak. Az 1. táblázatban szereplő példa is jól illusztrálja, hogy habár a két mondat szókészletben különbözik ugyan, de tartalmilag szinte teljesen megegyezik. A legtöbb publikációban található

módszertan gyengesége meglátásom szerint pedig pont abban rejlik, hogy csak a formailag teljesen megegyező szavakat – elemzési egységeket - tudja azonosítani, így az adott nyelvrendszer hálózatként való értelmezése erre a problémára egy hatékony kezelési eljárás lehet, melyet a 3.1 fejezetben fogok részletesen bemutatni.

3 A módszertan bemutatása

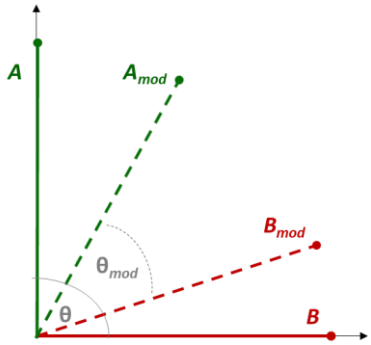
3.1 A felhasznált módszertan alapjai[2]

A módszertan kiválasztásánál fontos volt figyelembe venni azt is, hogy milyen egyéb módszerek lehetségesek az adott problematika megoldására. A kiinduló kérdés az volt, hogyan lehetne hatékonyan eldönteni, hogy két rövid terjedelmű magyar nyelvű mondat hasonlónak tekinthető-e. Tovább bonyolította a helyzetet az is, hogy nem egy általános témakörről lett volna szó, hanem egy konkrét szűkebb témáról, amelynek speciális nyelvezete van. A lehetőségek közül a divatos neurális hálók módszere nem tűnt alkalmazhatónak, hiszen a témának megfelelően annotált magyar nyelvű korpusz tudomásom szerint még nem áll rendelkezésre, továbbá az említett felügyelt tanulási algoritmus működésének ún. „fekete dobozként" fogja fel a tudományos szféra és a működésbe való kutatói beavatkozás túlságosan bonyolult. A terjedelmes dokumentumok esetén alkalmazott kulcsszókinyerés látszólag itt azért sem volt alkalmazható, mert a szövegek rövidsége miatt a szavak gyakorisága alapján nem lehetne kulcsszavakat azonosítani vagy azok lényegében egyenlőek lennének a mondat szókészletével, és ez folyamatos és aktív kutatói frissítését követelné a létrehozott modellnek. Korpusz hiányában a topik-modellezési eljárások sem látszottak hatékonynak.

Az alapul választott módszertan (Li és mtsai, 2006) erősségét az adja, hogy ötvözi a tudásbázis- és a korpuszalapú megközelítést. A módszertanban rendelkezésre álló tudásbázis egy hálózatként definiálható, ahol a hálózat pontjai az elemzési egységek – alapesetben szavak -, és hálózatelméleti alapfogalmakkal lehetséges leírni az egységek közötti kapcsolatokat.

A 2.2 fejezetben leírt vektorreprezentáció módosítása adja a módszer egyik sarkalatos pontját. A szöveghasonlóság módszertanához forrásként felhasznált cikk (Li és mtsai, 2006) javaslata szerint ezért a mondatokat leíró vektorok ne azt tartalmazzák, hogy egy adott szó előfordul-e a mondatban vagy sem (0/1), hanem azt, hogy az adott szóhoz milyen mértékben hasonló szó szerepel a mondatban ([0-1] közötti hasonlósági érték). Így a szókészlet alapján teljesen különböző mondatok vektorai „közelíthetők" egymáshoz, amit az 1. ábra szemléltet.

[2] A 3.1 fejezetben bemutatott módszertan az alábbi irodalom alapján készült: Yuhua és mtsai (2006)



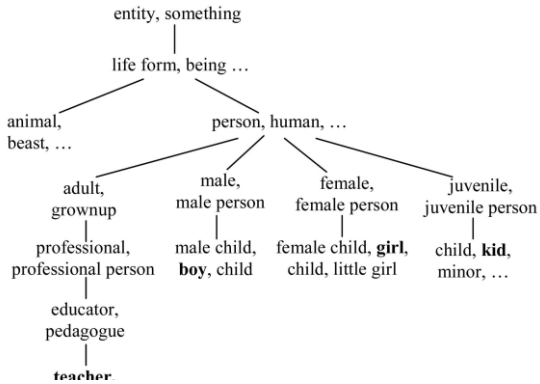
1. ábra: Az A és a B mondat vektorreprezentációjának ábrázolása

A hasonlósági értékekkel módosított vektor szemantikai vektornak nevezhető, amely az 1. táblázat példamondatai esetében a következőképpen néz ki:

2. táblázat: Arányokkal módosított szemantikai vektor

		pénzintézet	van	hitel	bank	rendelkezik	kölcsön	
A	Egy pénzintézetnél van hitele.	[1	1	1	0.9	1	1	]
B	Egy banknál rendelkezik kölcsönnel.	[0.9	0	0	1	1	1	]

Az eredeti vektor értékei módosultak azzal, hogy az értékek helyén nemcsak egyszerű bináris értékek (szerepel-e benne vagy sem), hanem arányszámok szerepelnek, amelyek a szavak egymáshoz viszonyított kapcsolatát tükrözik egy hierarchikus hálózatban – ún. ontológiában. Angol nyelven ugyan rendelkezésre áll egy nyilvánosan használható ontológia, WordNet (wordnet.princeton.edu), magyar nyelven azonban csak általános és jogi témában létezik nyilvánosan elérhető ontológia. További jelentős



2. ábra: Részlet az angol Wordnetből

erőforrásokat igénylő feladat ezért a kutatási kérdésnek megfelelő témaspecifikus hálózat létrehozása, de ezen hálózat létrehozásának folyamata nem képezi jelen tanulmány részét. Két szó hasonlóságát a 2-es képlet szerint írható le:

$$s(w_1, w_2) = f(l, d) = f_1(l) \cdot f_2(d) \quad \textbf{(2)}$$

A 2-es képlet szerint két szó hasonlósága függ attól, hogy mekkora a köztük levő legrövidebb út hossza az ontológiában (l ~ length), illetve attól is, hogy a két szó közös őse milyen mélyen helyezkedik el a hálózatban (d~depth). A 2. ábra szerint (melyen egy részlet látható az angol WordNet ontológiából) az angol *boy* és *girl* szavak között a legrövidebb út hossza 4 egység, míg a közös ősük (*person*) a hálózatban 2 mélységben található, ha az ontológia gyökerének az *entity* pontot vesszük. Li és mtsai (2006) szerint az ontológiából meghatározott úthossz és mélység a 3-as képletben szereplő függvények szerint számolható, ahol α és β a kutató által meghatározott értékek.

$$f_1(d) = \frac{e^{\beta d} - e^{-\beta d}}{e^{\beta d} + e^{-\beta d}} \quad \text{és} \quad f_2(l) = e^{-\alpha l} \quad \textbf{(3)}$$

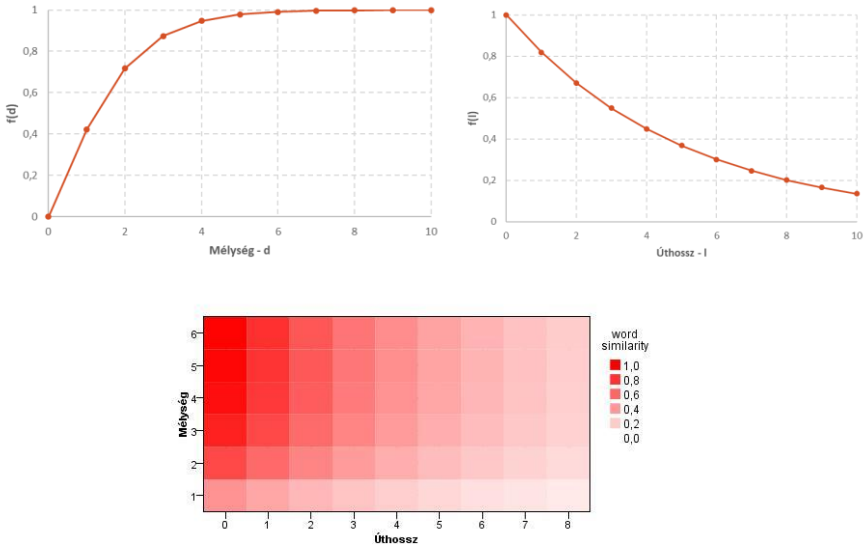
A módszertan magyar nyelvre való átültetése során az α és β paraméterek finomhangolása is megtörtént, de a tesztelés során ezeknek a paraméterek beállítása maradt α=0.3 és β=0.45 (Li és mtsai, 2006).

Két szó annál hasonlóbb egymáshoz, minél rövidebb úton juthatunk el egyik szótól a másikig, valamint minél mélyebben helyezkednek el a hálózatban, hiszen annál pontosabban lehet definiálni a szavak jelentését. A számításnál figyelembe vett úthossz és mélység függését a 3. ábra mutatja.

Két mondat hasonlóságát pedig a 3.1 fejezetben leírt módosítások alapján feltöltött vektorokkal számíthatjuk ki a koszinusz hasonlóság segítségével, melynek programkódrészlete a 4. fejezetben található.

3.2 Szórendi és nyelvtani hasonlóság

A mondatok hasonlóságának vizsgálatakor azonban nemcsak a mondatot alkotó szavak egyezősége, hasonlósága a fontos, hanem azok mondatban betöltött szerepe is. Ezért a szemantikai hasonlóságon túl egy másik hasonlósági mértéket is szükséges definiálni. A szövegösszehasonlítás módszertanához forrásként használt cikkben (Li és mtsai, 2006) a szerzők ezt egy szórendi hasonlósággal jellemzik. A szórendi hasonlóság esetében a mondatokat leíró szórendi vektor azt jelöli, hogy a két összehasonlítandó szövegrész szavai vagy elemzési egységei a mondatban hányadik helyen álló szóhoz hasonlítanak a legjobban. Hiszen nem mindegy, hogy „*Egy férfi látott egy vonatot*.” vagy „*Egy férfit láttunk a vonaton*.”. A 3. táblázatban látható, hogy ezen két mondat csupán a szórend figyelembevétele alapján teljesen ugyanaz magyar nyelven. Angol nyelven a szórend pontosan kifejezi a nyelvtani különbséget, de magyar nyelven a kötetlen szórend miatt a nyelvtani szerep figyelembevétele nélkül teljesen azonosnak számítana a két mondat, hiszen a bennük szereplő szavak szótövezett alakja azonos, azonban a mondat jelentése teljesen különböző.



3. ábra: A szóhasonlóság úthossz és mélység függése

3. táblázat: Példa egy szórendi vektorra

		férfi	lát	vonat	
A	Egy férfit láttunk a vonaton.	[1	2	3	]
B	Egy férfi látott egy vonatot.	[1	2	3	]

A szórendi vektorokból a szórendi hasonlóság a következőképp számítandó:

$$S_o(o_1, o_2) = 1 - \frac{o_1 - o_2}{o_1 + o_2} \quad \textbf{(4)}$$

A 4-es képletben o_1 és o_2 a mondatok szórendi vektorait jelölik. Két mondat hasonlósága végül a szemantikai és a szórendi hasonlóság súlyozott átlagaként számítható ki:

$$S(S_1, S_2) = \delta S_s + (1 - \delta) S_o \quad \textbf{(5)}$$

Az 5-ös képletben S_s a szemantikai hasonlóságot, S_o pedig a szórendi hasonlóságot jelöli, δ pedig egy arányt mutat, amelynek értéke 0 és 1 között lehet, Li és mtsai tanulmányukban a 0.85-ös arányt javasolják, és ez jelen modellben is megfelelőnek bizonyult, ezért ez nem került változtatásra.

A 3. táblázat alapján azt a következtetést vonhatjuk le, hogy a szórendi vektor módosításának van létjogosultsága, hiszen a magyar nyelv kötetlen szórendje nem feleltethető meg az angol nyelv kötött szórendjének.

Ezért a szemantikai hasonlóságon kívül kétféle hasonlóságot is lehetne definiálni: szórendi (Li és mtsai, 2006) és nyelvtani hasonlóságot.

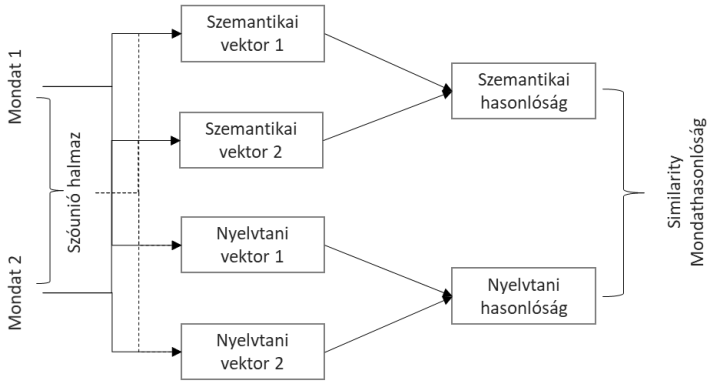
A nyelvtani hasonlóság kiszámításához egy nyelvtani vektor létrehozása látszott megfelelőnek, amely azt jelöli, hogy milyen nyelvtani szerepet töltenek be a két mondatban előforduló szavak. E szerint a magyar nyelvben a mondatbeli szerepeket kell rangsorolni az angol kötött szórendhez mérten, amelyben alapvetően az alanyt követi a tárgy, az állítmány majd ezt követően állnak a határozók. Ezzel a módosítással a 4. táblázatban szereplő vektor értékei a következőképpen módosulnak, és a 3. táblázatban szereplő két példamondat a nyelvtani vektor alapján már nem mondható azonosnak:

4. táblázat: Példa nyelvtani vektorra

		férfi	lát	vonat	
A	Egy férfit láttunk a vonaton.	[3	2	5	]
B	Egy férfi látott egy vonatot.	[1	2	3	]

A tesztelés során felmerült, hogy a szórendi vagy a nyelvtani hasonlóság figyelembevételével kapunk-e pontosabb eredményt, és a nyelvtani vektor használatával bizonyult jobbnak a modell – magyar nyelven.

Összefoglalva tehát két mondat hasonlósága két hasonlósági mérték – a szemantikai és a nyelvtani hasonlóság – súlyozott átlagaként számítható ki, ahogy ezt a 4. ábra is szemlélteti. A hasonlósági érték egy 0 és 1 közötti érték. Két mondat hasonlónak tekinthető, ha egy küszöbszint feletti a hasonlóság értéke. E küszöbszint meghatározása a tesztelés során történt meg, de fontos hangsúlyozni, hogy ez a küszöb minden kutatásnál eltérő lehet, és hasonlóan a módszertan alapjául szolgáló cikkhez (Li és mtsai,, 2006), ennek a küszöbnek az értéke itt sem kerül közlésre.



4. ábra: A szemantikai hasonlóság kiszámításának alapfolyamata

3.3 A módszer magyar nyelven való implementálása

A 3.2 fejezetben a módszertan alapjaiban történtek változtatások, melyek a magyar nyelv struktúrájához mérten alakultak ki. A módszer magyar nyelven való implementálásakor azonban számos további akadályba ütközhet a kutató. A módszertan alapjául

szolgáló cikk (Li és mtsai, 2006) szerzője azonban a tanulmányban semmilyen iránymutatást nem ad a módszer nyelvfüggő jellemzőit illetően, ezért a következőkben ezen kritikus pontokra fogok rátérni, amely nem kifejezetten matematikai vagy statisztikai, hanem inkább nyelvi eredetű.

Egy szövegelemzés szerves része a szövegtisztítás is, és ennek minősége döntő lehet a modell végső teljesítményét illetően.

A szövegösszehasonlító algoritmus Python programozási nyelven került implementálásra, amely esetén eddigi munkám során a magyar nyelven beépített szótövező algoritmusok egyike sem mutatott hatékony eredményeket (nltk package). A szótövezés során a Szegedi Tudományegyetem magyarlánc nevű JAVA alkalmazását használtam. Akadtak azonban olyan esetek, amiket a szótövezésen felül és azzal egyidőben kellett kezelni: a tagmondatra bontást, a tagadás kezelését, jelentésmódosító szavak helyzetét, melléknévi igenevek kezelését és a stopszavazást.

3.3.1 A tagmondatra bontás

A szövegek összehasonlításánál az összehasonlítandó szövegek hosszát is figyelembe kell venni, ezért célszerűnek tartható a szöveg hosszának normalizálása, ami jelen esetben a tagmondatra bontással tűnt kivitelezhetőnek.

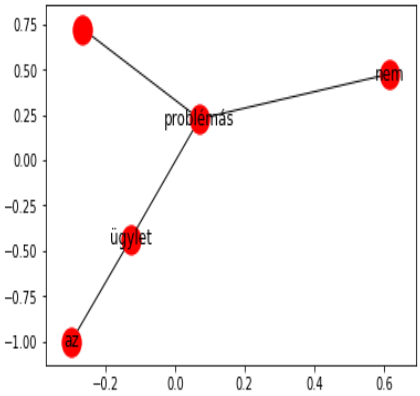
A tagmondatra bontást kétféleképpen közelítettem meg: egyrészről az írásjelek és a kötőszavak felől, másrészről a magyarlánc nyelvtani elemzéséből kiindulva. A magyarlánc alkalmazásban mellérendelő tagmondatok esetében tapasztalhatók hibák, de a nyelvtani elemzése ennek ellenére is hatékonynak bizonyult a tagmondatokra bontás elvégzésében.

Például a következő mondat esetében – *„Az ügyfél problémamentes, megbízható, ezért javasoljuk a pozitív elbírálást.”*. Az írásjelek és kötőszavak felől megközelítve ez a mondat három tagmondatból áll, viszont a magyarlánc alkalmazás szerint ez csupán két tagmondat, mivel az alanyi állítmányokat nem értelmezi külön tagmondatnak. Végül azonban mivel a felsorolásokat hatékonyan tudja értelmezni, ezért ezen szempontot erősebbnek érezve a nyelvi függőségi viszonyok szerinti részekre bontásra esett a választás.

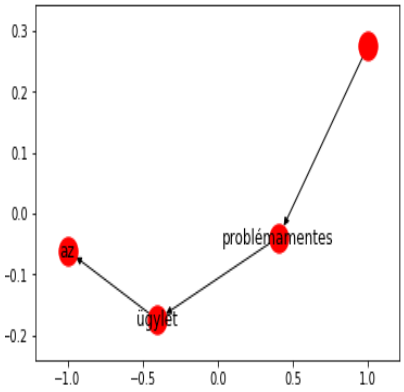
3.2.2. A tagadás kezelése

A módszertan alapjául szolgáló tanulmányban (Li és mtsai, 2006) a szerzők nem térnek ki arra, hogy mi a helyzet azzal, ha két mondat ellentétes jelentésű, hiszen a szemantikai hasonlóság értékkészletét 0 és 1 között határozzák meg. Azonban két mondat nagyon hasonló lehet szókészletében, de már csupán egy tagadás is megváltoztathatja a mondat értelmét. Például nem mindegy, hogy *„Egy banknál sem vezet számlát.”* vagy *„Egy banknál vezet számlát.”*. A tagadószavak kezelése viszonylag egyszerűnek tűnik, azonban többféleképpen is kifejezhetjük azt. Alapvetően a *„nem, nincs, sem, nincsen, nélkül, nélküli”* kifejezések egyszerűen kapcsolódhatnak az adott mondatrészhez, és ez a nyelvtani függőségekből jól látszik is, ezért ezt hozzá tudjuk kapcsolni az adott szóhoz (szavakhoz) automatikusan. Azonban a tagadás nemcsak különálló részként kapcsolódhat a szóhoz, hanem azzal egybeírva is. Példaként említhető az a két mondat, hogy *„Az*

ügylet problémamentes." és az *„Az ügylet nem problémás."*, melyekben a szavak egymáshoz viszonyított kapcsolatát az 5. és a 6. ábra mutatja. Az 5. ábrán látható, hogy a *problémás* szóhoz tartozik egy tagadószó *(nem)*, viszont a 6. ábra szerint a nyelvtani elemzés kimenetében semmi nem utal a tagadásra, pedig valójában a *problémamentes* is ugyanazt jelenti, mint az, hogy nem problémás.



5. ábra: Az ügylet nem problémás nyelvtani kapcsolatai



6. ábra: Az ügylet problémentes nyelvtani kapcsolatai

Azonban nem számítható egyszerűen tagadásnak az összes olyan melléknév, amely *„mentes"* szóval végződik, hiszen elképzelhető olyan eset, hogy például a *problémamentes* és a *jó* összehasonlításánál tévedünk azzal, ha tagadásnak vesszük a problémamentes szót. A tagadás tehát úgy értelmezhető, hogy a jelentés hasonló, de a kapcsolat iránya ellentétes, tehát -1-gyel való szorzás alkalmazható.

Összességében a tagadást több szinten érdemes kezelni, de itt is figyelembe kell venni a nyelvi sajátosságokat és a rendelkezésre álló nyelvi eszköztárt is.

3.2.3 Jelentésmódosított szavak kezelése

A szemantikai hasonlóság megállapításakor felmerül az a kérdés, hogy a bizonyos szófajokból képzett szavakat mennyire vegyük hasonlónak vagy azonosnak ahhoz a szóhoz, amiből képeztük. Jelen esetben például a melléknévi igenevek azzal az igével egyenrangúnak vehetők, amiből képeztük, vagy akár a főnevek -i képzős formái is azonosnak tekinthetők azzal a főnévvel, amiből létrejöttek. Ez a tulajdonság a magyarlánc a nyelvtani elemzésből látszik, ezért automatikusan ezeket szabályszerűen ki lehet szűrni, azonban vannak kivételek is, amelyek felvétele manuálisan történhet (például -i képzős melléknevek, amik főnévből képződtek, vagy melléknévi igenevek).

3.2.4 Stopszavazás

Minden szöveganalitikai elemzés során elképzelhető olyan helyzet, hogy bizonyos szavak nem relevánsak a kutatási kérdés szempontjából, és a tisztítási folyamat során ezeket el kell távolítanunk azért, hogy hatékonyabban tudjon működni a modell. Jelen esetben a vektorok hosszának növekedése csökkentené a hasonlóság mértékét, és nagyobb hiba is felmerülhetne, ezért a stopszavak meghatározását mindig nyelv- és kutatásfüggően kell meghatározni és nem szabad csak és kizárólag a beépített stopszólistákra hagyatkozni, hanem ez egy nagyon fontos szakértői döntés kell legyen.

4 Programkód részletek

#koszinusz távolság

```
def cosine_similarity(x,y):
    numerator = sum(a*b for a,b in zip(x,y))
    denominator = square_rooted(x)*square_rooted(y)
    return round(numerator/float(denominator),5)
```

#szemantikai hasonlóság

```
def semantic_similarity(word1, word2, G, a=0.2, b=0.45):
    if word1 == word2: #ha a két szó teljesen azonos
        return 1
    else:
        try:
            l=shortest_path_length(G, word1,
word2)["length"] #legrövidebb út hossza
            h=shortest_path_length(G, word1,
word2)["depth"] #közös ős mélysége a hálózatban
            return (math.exp(-1*a*l)*((math.exp(b*h)-
math.exp(-1*b*h))/(math.exp(b*h)+math.exp(-1*b*h))))
        except TypeError:
                return (np.nan)
```

5 Eredmények

A módszertan alapjául szolgáló tanulmány (Li és mtsai, 2006) magyar nyelven való implementálása nem volt kihívásoktól mentes. A 3. fejezetben leírt módosítások segítségével azonban egy hatékonyan működő szövegösszehasonlító rendszer jött létre. Két egymástól független teszten a 7. ábrán szereplő eredmények születtek a módszerek alkalmazásával (100-100 hasonló, és 150-150 nem hasonló mondatpár):

similar_dummy

similarity_dummy		SIMILAR	not_SIMIL...
not_similar	Count	37	139
	Column %	36.275	92.667
similar	Count	65	11
	Column %	63.725	7.333

similar_dummy

similarity_dummy		SIMILAR	not_SIMIL...
not_similar	Count	40	136
	Column %	40.000	90.667
similar	Count	60	14
	Column %	60.000	9.333

7. ábra: Modell teljesítménye

Az értékelésnél az első- és másodfajú hibát tekintve súlyosabb hibának számított az, hogy ha egy mondatpárról a modell azt mondta, hogy hasonlóak, miközben jelentésükben teljesen különbözőek voltak. A modell szerinti álpozitív értékek sokkal súlyosabb hibának számítottak, mint az álnegatívak, de összességében a 60 %-os találati arány és a maximum 10 %-os másodfajú hiba még megfelelő teljesítménynek mondható.

6 Konklúzió

Összességében elmondható tehát, hogy a nyílt forráskódú programnyelvek terjedésével egyre több ötlet és lehetőség kerül nyilvánosságra a természetesnyelv-feldolgozás területén, amelynek az implementálása sokszor szükségessé válhat egyéb idegen nyelven. A kutatási kérdésünknek megfelelő eszköztár és módszertan megtalálása nem egyszerű feladat, és jelen tanulmány azt hivatott bemutatni, hogy egy jó kiindulási alapból lehet fejleszteni, de nem határok nélkül. A természetesnyelv-feldolgozó rendszerek esetén figyelembe kell venni az adott nyelv jellegzetességeit, szókészletét és a nyelv elterjedtségét is, és a rendelkezésre álló elemzési eszköztárt, és a munka nagy részét nem csupán egy más forrásból származó ötlet lemásolása és reprodukálása jelenti. Az implementálás mögött hosszas szakértői munka van, mely segítségével létrehozható az a témaspecifikus tudás, amely nélkül önmagában egy idegen forrásból származó programkód nem tud lefutni. Úgy gondolom, hogy ez a határa a mai természetesnyelv-feldolgozás térhódításának, ugyanakkor ebben a témaspecifikus szakértői tudásban rejlik a természetesnyelv-feldolgozás továbbfejlesztése is.

Jelen tanulmány iránymutatásai a magyar nyelvű szöveganalitika egy részébe engednek betekintést, de korántsem tekinthetők teljes leírásnak. A magyar nyelvű témaspecifikus források repertoárjának bővítése hozzájárulhat a kutatási terület fejlődéséhez, és ennek megfelelő mértékű növekedése esetén könnyebbé válhat az idegennyelvű módszertanok magyar nyelven való implementálása. Azonban a nyelvek jellegzetessé-

geinek mindig lesznek olyan pontjai, amelyeket figyelembe kell venni egy ehhez hasonló feladat esetén, és ezen szempontok megtalálása a további hasonló témájú kutatások létjogosultságát adhatja.

Hivatkozások

Lapata, M., Barzilay, R. Automatic evaluation of text coherence: Models and representations. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence. (2005)

Li, Y., McLean, D., Bandar, Z., O'Shea, J., Crockett, K.: Sentence similarity using semantic nets and corpus statistics. In: IEEE Transactions on Knowledge and Data Engineering. 18. pp. 1138-1150. (2006)

Mihalcea, R., Corley, C., Strapparava, C.: Corpus-based and Knowledge-based Measures of Text Semantic Similarity In: Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1. pp. 775-780. Boston, Massachussets (2006)

Miháltz, M., Hatvani, Cs., Kuti, J., Szarvas, Gy., Csirik, J., Prószéky, G., Váradi, T.: Methods and Results of the Hungarian WordNet Project. In: Proceedings of the Fourth Global WordNet Conference GWC, pp. 310-320. (2008)

Papineni, K., Roukos, S., Ward, T., Zhu, W. J.: BLEU: a method for automatic evaluation of machine translation (PDF). ACL-2002: 40th Annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)

Salton, G., Singhal, A., Mitra, M., Buckley, C.: Automatic text structuring and summarization. Information Processing and Management 2(32). pp. 193-207.(1997)

Slimani, T.: Description and Evaluation of Semantic Similarity Measures Approaches. International Journal of Computer Applications 80.10 pp. 25–33. (2013)