Többértelműségek magyar mondatok számítógépes elemzésében - a „meg” szó szófajának vizsgálata gyakoriságokkal

témalabor-dolgozat, Makrai Márton haramadéves matematikus (BME-TTK)
témavezető: Babarczy Anna

Többértelműségek magyar mondatok számítógépes elemzésében - a „meg” szó szófajának vizsgálata gyakoriságokkal

A félév során a magyar nyelv számítógépes feldolgozásval foglalkoztam. Azt vizsgáltam, hogy a „meg” szó egy adott mondatban kötőszó-e vagy igekötő. Olyan szabályos kifejezéseket (*regular expression*,) kerestem, melyekre illeszkedő mondatokban a „meg” vagy túlnyomó részt kötőszó, vagy túlnyomó részt igekötő. (Ha egy mondatban több „meg” is szerepel, akkor mindig csak arról (azokról) van szó, amelyek miatt a mondat illeszkedik a szóban forgó kifejezésre.) A magyar mondatok 82,66%-ára találtam olyan szabályt, ami legalább 0,95 valószínűséggel helyes jóslatot ad a „meg” szófajára.

A feladat jelentősége – A nyelvtechnológia

A huszadik század ötvenes éveiben, a számítógépek megjelenésének következtében felmerült az a igény, hogy egy gép intelligens legyen. Ennek első konkrét megfogalmazását Alan Turing adta: a Turing-tesztnek az a gép felel meg, amely úgy tud válaszolni egy ember mondataira, hogy az ne tudja megmondani, hogy géppel beszél. (Eredetileg írásbeli kommunikációról volt szó, de gondolhatunk hangzó szövegre is.) A mesterséges intelligencia kérdésének felvetése óta kiderült, hogy a feladat nagyon nehéz, mind a nyelvelemzés folyamata – melynek során géppel elemzünk egy szöveget a kisebb nyelvtani egységek (hangok) felől a nagyobbak (szöveg) felé haladva – mind a szintézis (az adott jelentésű (akár hangzó) mondat megkonstruálása) sok kihívást tartogat. A problémát ma az alkalmazás sürgeti: gépi fordítást, tolmácsgépeket szeretnénk létrehozni, illetve törekszünk az intelligens webkeresés felé, melynek során a kereső úgy ad találatokat a keresőkifejezéshez, hogy valamilyen mértékben értelmezi azt.

A nyelvtechnológia jelenlegi módszerei két fő irányzathoz sorolhatók. A szabályalapú megközelítésben a szabályok tipikusan „mindent vagy semmit” alapon mű ködnek, tehát vagy érvényesül egy szabály a feldolgozás során, vagy nem érvényesül. Ezzel szembeállítható az újabban elterjedt statisztikai módszer, ahol példatárakból (korpuszokból) automatikusan nyert információ segítségével történik az elemzés. A mintákhoz tipikusan valószínűségeket rendel az algoritmus és újabb szövegek feldolgozásakor vagy generálásakor arra törekszik, hogy az együttes valószínűség maximalizálása mellett valósítsa meg a mintaillesztést.

A magyar nyelv számítógépes elemzése különösen gyerekcipőben jár, ami megerősíti azt, hogy fontos témával van dolgunk.

Az eszköztár

A korpuszt, amelyet használtam, a BME Szociológia és Kommunikáció Tanszékén állították össze (ld. A témalaboromhoz szorosabban kapcsolódó honlapok) Ez az interneten elérhető és szabadon használható. Ebből a korpuszból 136 711 mondatot morfológiailag leelemeztem a `hunpos` nevű program (a `hunmorph` újabb változata) segítségével. Az így kapott szövegben a `grep`-pel kerestem a különböző szabályos kifejezéseknek megfelelő mondatokat.

A `hunpos` kimenete

```
Lesz/VERB ott/ADV valami/NOUN munkalehetőség/NOUN |,/PUNCT
neked/NOUN<PERS<2>><CAS<DAT>> meg/CONJ lesz/VERB időd/NOUN<POSS<2>>
gondolkozni/VERB<INF> |./PUNCT
```

Ahogy ez a péda is mutatja, a `hunpos` minden szó után „/”-rel elválasztva adja meg az elemzést, és az írásjelek (központozások) elé „|” jelet tesz.

Az egyes szavak elemzése a szófaj megjelölésével kezdődik. Az alábbi táblázat tartalmazza azoknak a szófajrövidítéseknek a jelentését, melyeket használtam a kutatás során.

ADJ	melléknév
ART	névelő
CONJ	kötőszó
DET	determináns (pl. azon,
NOUN	főnév
NUM	számnév
POSTP	névutó
PREV	igekötő
VERB	ige

A szófaj rövidítése után részletesebb elemzés következik. A

```
neked/NOUN<PERS<2>><CAS<DAT>>
```

elemzésben pédául az első zárójelben (`<PERS<2>>`) fel van tüntetve, hogy ennek az igének van egy személye, méghozzá második személyű. Azt, hogy van egy személye, úgy értjük, hogy nem harmadik személyű. A második zárójelben (`<CAS<DAT>>`) azt látjuk, hogy esete is van, azaz nem alanyesetű, hanem részes.

Ezekből a részletesebb elemzésekből csak az infinitív (főnévi igenév) jelölését (`gondolkozni/VERB<INF>`) használtam a félév során, és a végeredményben az sem jelenik meg.

A szabályos kifejezések

A `grep` segítségével kerestem a korpuszban az adott szabályos kifejezésekkel illeszkedő mondatokat. A `grep` egy változatát használtam, a kiterjesztett szabályos kifejezéseket (*extended regular expression*) elfogadó `egrep`-et.

Az egrep speciális karakterei a következők:

`(` `)` A kerek zárójel tagolásra való.

`[ ]` Szögletes zárójellel halmazt lehet kifejezni, a szögletes zárójeles kifejezés illeszkedik minden olyan karaterre, ami a zárójelek között fel van sorolva, így pédául a `v[ae]` kifejezés a „`va`” és a „`ve`” katakterláncra illeszkedik.

`^` A kalap szögletes zárójelen belül a komplementer jele, pl. a `[^/]` kifejezés minden karakterre illeszkedik a „`/`” kivételével.

`|` A „`|`” a „ vagy” jele, így pl. a `(( m)|M)` kifejezés a „ `m`” és a „`M`” karakterlánc bármelyikére illeszkedik.

`\` A „`\`” egy speciális karakter előtt azt jelöli, hogy az adott speciális karakter elveszíti a speciális jelentését és önmagára illeszkedik , pl. a `\|` kifejezés a `|` karakterre.

A „`\`” egy másik jelentésére példa a `\w` kifejezés, mely minden szókarakterre (alfanumerikus karakterre) illeszkedik.

`*` A csillag azt jelzi, hogy az adoltt karakterből vagy karakterláncból bármennyi szerepelhet az adott helyen, beleértve a nulla előfordulást is, így a `\w*` kifejezés többek közt illeszkedik bármely magyar szóra és az üres karakterláncra is.

A többi általam használt karakter (beleértve a szóközt) önmagára illeszkedik, pl. a `/NOUN` kifejezés a „`/NOUN`” karakterláncra.

A kutatás eredményei. Szabályok

Formálisan szólva, szabálynak nevezem az olyan függvényeket, amelyek szabályos kifejezéseken vannak értelmezve, és értékük CONJ vagy PREV. Arról van szó, hogy egy szabály megadja, hogy az adott kifejezésre illeszkedő mondatban a „meg” milyen szófajú.

A főnévi csoport (*noun phrase*, NP)

Ahhoz, hogy megértsük a szabályokat, emylekre jutottam, szót kell ejtenünk a szintaktikáról. A szintaktikai elemzés során azt keressük, hogy ez adott mondaton belül mely szavak tartoznak össze, és alkotnak együtt kifejezéseket. Például abban a modatban, hogy „Írd meg a receptet! ”, az „írd meg” egy igei csoport (*verb phrase*, VP), mert a feje ige, az „a receptet” pedig egy főnévi csoport, mert a feje főnév.

[Írd meg]$_{\mathrm{VP}}$ [a receptet]$_{\mathrm{NP}}$

A főnévi csoport tartalmazhat névelőt (ART), determinánst (DET), számnevet (NUM) és melléknevet (ADJ) a főnév előtt, és névutót (POSTP) a főnév után, pl. a három barna kutyával:

```
a/ART három/NUM barna/ADJ kutyával/NOUN<CAS<INS>> együtt/POSTP
```

Most megfogalmazom a hét szabályszerűséget, később pedig formalizálom őket, hogy szabályok legyenek a szó előbb definiált értelmében. Az első hét szabályszerűség olyan esetekről szól, amikor a „meg” mindig igekötő. Azt, hogy a szabályok túlnyomó része az igekötő esetére vonatkozik, némileg magyarázza, hogy a „meg” az esetek több mint négyötödében igekötő. (Ezt a kijelentést a később ismertetett eredményekre alapozom. A `hunpos` – sok esetben hibás - elemzését alapul véve az igekötői előfordulások még erősebb többsége lenne feltételezhető.)

Az alábbi hat környezetben a „meg” mindig igekötő:

1. egy ige és egy főnévi csoport között (Amerika nem a világ ura, nem *tehet meg mindent* kényére-kedvére.)
2. vessző, vagy más írásjel előtt (Vagy rossz címet adott *meg,* vagy az oldal címe megváltozott.)

3. ha két ige követi (akár infinitív; pl. Évezredek óta ezt mindig *meg lehetett állapítani*.)
4. „is” előtt (Krisztina főhercegnő kegyesen *meg is* ajándékozza egy aranyórával.)
5. ha egy ige és egy „-va/ve” végű határozószó (határozói igenév) követi (A szavazási lehetőségek száma *meg van határozva*.)
6. két ige között (A kiadók így nem *próbálták meg betiltatni* a Xerox-eljáráson alapuló fénymásolást.)

7. Két főnévi csoport között a „meg” mindig kötőszó. (A lányok egymásba karolva sétáltak a *Cifrapalota meg a szemközti zsinagóga* előtt.)

Az első szabály. Egy ige és egy főnévi csoport között a „meg” mindig igekötő.

Készítsük el azt a kifejezést, amely az ennek a szabálynak megfelelő mondatokra illeszkedik!

Az hunpos kimenetéről tudjuk, hogy az igék elemzése (a „/” után) a `VERB` lánccal kezdődik, és ez a lánc csak az igék elemzésének kezdetén fordul elő. Ez után következik a részletesebb elemzés a szóközig. Erre a részre illeszkedni fog a `[^ ]*` kifejezés. Ezt követi a szóköz, majd a meg. Itt nem feledkezünk meg a „/ ”- ről sem, hogy ne kapjunk meg minden „meg” kezdetű szót. Nem foglalkozunk azzal, hogy a `hunpos` milyen szófajúnak elemzi a szóban forgó „meg”-et, a „meg” elemzésére `[^ ]*` illeszkedik (ez lehet bármi, ami nem szóköz). A főnévi csoportban a kifejezés fejét képző főnév előtt névelőt (ART), determinánst (DET), számnevet (NUM) és/vagy melléknevet (ADJ) állhat. (Ezeknek a bővítményeknek a sorrendje szabályos a magyar nyelvben, de ez a mi szempontunkból nem érdekes.) Így egy-egy ilyen bővítmény elemzésére (beleértve az elemzést követő szóközt) illeszkedik a `(\w*/((ART)|(DET)|(NUM)|(ADJ))[^ ]* )` kifejezés. (A `\w*`, ahogy már említettük, bármely szóra illeszkedik.) Ilyen bővítményből több is lehet, és az is megengedett, hogy egy se legyen, ezt jelöli a „`)` ” utáni „`*`”. Végül szerepel maga a főnév. Így kapjuk a

```
VERB[^ ]* meg/[^ ]* (\w*/((ART)|(DET)|(NUM)|(ADJ))[^ ]* )*/w*/NOUN # PREV
```

szabályt.

Példa: `Amerika/NOUN nem/ADV a/ART világ/NOUN ura/NOUN<POSS> |,/PUNCT nem/ADV tehet/VERB meg/PREV mindent/NOUN<CAS<ACC>> kényére-kedvére/NOUN<CAS<SBL>> |./PUNCT`

A második szabály. Vessző, vagy más írásjel előtt a „meg” mindig igekötő.

Mivel a „meg” a mintázat elején szerepel, biztosítanunk kell, hogy a kifejezés ne illeszkedjen „meg” helyett egy „`meg`” végű szóra, de illeszkedjen a „`Meg`”-re is (mondat elején). A példamondat

```
Vagy/CONJ rossz/ADJ címet/NOUN<CAS<ACC>> adott/VERB<PAST> meg/PREV |,/PUNCT vagy/CONJ az/ART oldal/NOUN címe/NOUN<POSS> megváltozott/VERB |./PUNCT
```

elemzéséből látható, hogy az írásjelek elé a `hunpos` „`|`” jelet tesz. Ezt a jelet másra nem használja a `hunpos`, így jól tesszük, ha erre a karakterre keresünk. Mivel az `egrep` aktív karakteréről van szó, „`\`”-t alkalmazunk. Így kapjuk meg a szabályt:

```
(( m)|M)eg/[^ ]* \| # PREV
```

A harmadik szabály. Ha a „meg”-et két ige követi, akkor legtöbbször igekötő.

A kifejezésünk úgy fog kezdődni, hogy `(( m)|M)eg/[^/]*`. Itt a `[^/]*` illeszkedni fog a „meg” elemzésére, az azt követő szóközre, és a következő szóra (az igére) is (a „/”-re már nem). (Megjegyzem, hogy a `grep` mohó (*greedy*) algoritmust használ, például először minde karaktert a `[^/]*`-ra illeszt, amit csak lehet, így éppen az itt leírt „gondolatmenet” szerint halad.) Ezt követi a „/”, majd az ige elemzése és a következő ige, végül annak az elemzése:

```
(( m)|M)eg/[^/]*/VERB[^/]*/VERB # PREV
```

Példa: `Évezredek/NOUN<PLUR> óta/POSTP ezt/NOUN<CAS<ACC>> mindig/ADV meg/PREV lehetett/VERB<PAST> állapítani/VERB<INF>`

A negyedik szabály. Ha a „meg”-et „is” követi, akkor igekötő.

Ezt könnyű formalizálni az eddigiek alapján:

```
(( m)|M)eg[^ ]* is/ # PREV
```

Példa: `Krisztina/NOUN főhercegnő/NOUN kegyesen/ADV meg/PREV is/ADV ajándékozza/VERB<DEF> egy/ART aranyórával/NOUN<CAS<INS>> |./PUNCT`

Az ötödik szabály. Ha a „meg”-et egy ige és egy „-va/ve” végű határozószó (határozói igenév) követi, akkor igekötő.

```
(( m)|M)eg/[^/]*/VERB[^/]*v[ae]/ADV # PREV
```

Itt a második `[^/]*` illeszkedik az ige bővebb elemzésére, az ige utáni szóközre, és a határozószónak a „`va`” vagy „`ve`” előtti részére.

Példa: `A/ART szavazási/ADJ lehetőségek/NOUN<PLUR> száma/NOUN<POSS> meg/PREV van/VERB határozva/ADV |,/PUNCT`

A hatodik szabály: Két ige között a „meg” túlnyomó részt igekötő.

```
VERB[^ ]* meg/[^ ]* \w*/VERB # PREV
```

Példa: `A/ART kiadók/NOUN<PLUR> így/ADV nem/ADV próbálták/VERB<PAST><PLUR><DEF> meg/PREV betiltatni/VERB<INF> a/ART Xerox-eljáráson/NOUN<CAS<SUE>> alapuló/ADJ fénymásolást/NOUN<CAS<ACC>> |./PUNCT`

A kötőszóra vonatkozó szabály: Két főnévi csoport között a meg legtöbbször kötőszó.

Ebben a kifejezésben nem feledkezhetünk meg arról, hogy a „meg” előtti főnévi csoport névutóval is végződhet:

```
NOUN[^ ]* (\w*/POSTP[^ ]* )*meg/[^ ]* (\w*/((ART)|(DET)|(NUM)|(ADJ))[^
]* )*[^/]*/NOUN # CONJ
```

Példa: `A/ART lányok/NOUN<PLUR> egymásba/NOUN<CAS<ILL>> karolva/ADV sétáltak/VERB<PAST><PLUR> a/ART Cifrapalota/NOUN meg/CONJ a/ART szemközti/ADJ zsinagóga/NOUN előtt/POSTP |./PUNCT`

Az eredmények számszerű értékelése

Jóságmérték

Nyilván jó egy szabály, ha sok esetet lefed. A másik mértéke egy szabály jóságának az, hogy amikor érvényes, milyen arányben ad helyes következtetést. Ezt a kettőt egyszerre kell szem előtt tartani, így a kettő szorzata alkalmas jóságmértéknek (*measure of merit*, a magyarítást Kornai Andrástól hallottam). Ha a második adat (hogy az adott szabály az esetek mekkora részében ad helyes következtetést) pontosan rendelkezésre állna, akkor ez a szorzat a helyesen elemzett „meg”-ek száma lenne.)

	többség	gyakori	rel gyak	arány	conj	prev	CONJ	PREV	jóság	
VERB meg NP	PREV	2846	33,66%	1,00	0	64	30	2816	2846	
meg,	PREV	2695	31,88%	1,00	0	72	24	2671	2695	
meg VERB VERB	PREV	914	10,81%	0,99	1	92	21	893	904	
NP meg NP	CONJ	228	2,70%	0,98	64	1	202	26	224	
meg is	PREV	183	2,16%	1,00	0	50	4	179	183	
meg VERB …v[ae]/ADJ	PREV	64	0,76%	1,00	0	44	6	58	64	
VERB meg VERB	PREV	63	0,75%	0,95	3	60	2	61	60	
minden speciális		6993	82,72%							
többi	PREV	1602	18,95%	0,74	13	37	559	1043	1185	
teljes	PREV	8454	100,00%	0,86	7	43	822	7632	7270	

Ez a táblázat tartalmazza a fenti szabályokkal kapcsolatos eredményeket.

A sorok egy-egy szabálynak felelnek meg. A „NP meg NP” helyétől eltekintve ugyanabban a sorrendben állnak, mint ahogy fent szerepeltek. A „minden speciális” kezdetű sor az előző hét összesítéses. A „többi” kezdetű sor azokról a mondatokról szól, amelyek semelyik mintára sem illeszkednek. A „teljes” kezdetű sor az összes „meg”-et tartalmazó mondatról szól.

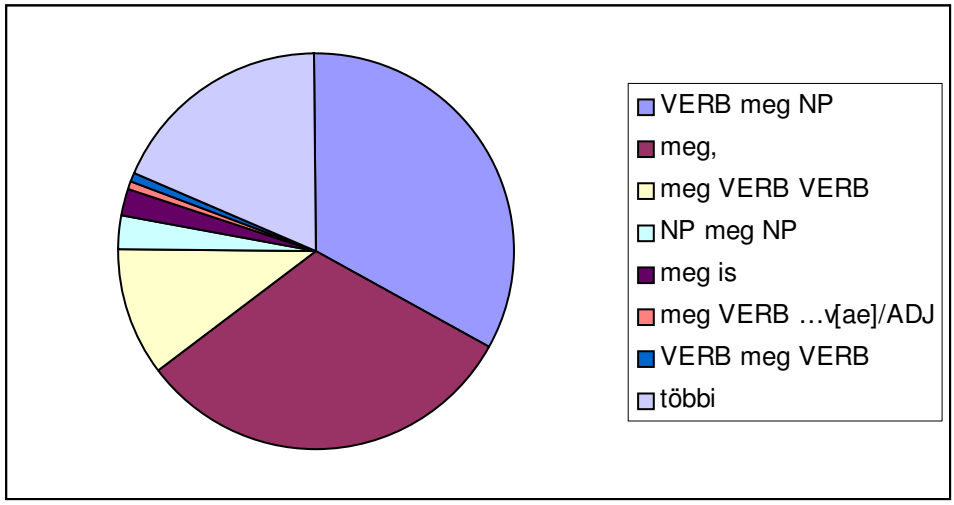
Az oszlopok. Az első oszlopban a mintázat sémája van feltüntetve. A második oszlopba azt írtam, hogy melyik a *többségi* szófaj. A szabályok esetén erre a szófajra következtetünk. A következő oszlopban az adott szabálynak megfelelő mondatok *gyakorisága*, majd *relatív gyakorisága* szerepel.

Az megfelelő mondatokból minden esetben vettem egy mintát (legalább 44, legfeljebb 93 mondatot), amelyen „kézzel” leellenőriztem a „ meg” szó szófaját. A megfelelő gyakoriságokat írtam a „*conj*” illetve a „*perv*” feliratú oszlopba, az „*arány*” feliratú oszlopba pedig a többségi szófajnak az ennek ebben a mintában való relatív gyakoriságát.

Azt is feltüntettem, hogy a `hunpos` hogyan elemzi a szóban forgó mondatokban a „meg”-et, ezt jelzi a „*CONJ*” és a „*PREV*”.

Fent bevezetett jóságmérték konkrétan a gyakoriság és a találati „arány" szorzata. Eszerint vannak a sorok rendezve: a jóságmérték szerint legjobb szabály van legfelül, és lefelé haladva csökken ez a szám.

Azt, hogy a szabályainkkal az esetek mekkora részét sikerült lefednünk, szemlélteti ez a diagram, mely az egyes szabályoknak megfelelő mondatok illetve a fennmaradó mondatok („többi”) arányát mutatja. Az esetek 18,95%-a feldolgozatlanul maradt, így további kutatás tárgya lehet.



A témalaboromhoz szorosabban kapcsolódó honlapok

Babarczy Anna témavezető
http://cogsci.bme.hu/~babarczy/
A `hunpos` korábbi változata: a `hunmorph`
http://mokk.bme.hu/resources/hunmorph
magyar vebkorpusz a BME Szociológia és Kommunikáció Tanszékén
http://mokk.bme.hu/resources/webcorpus/index_html
Ez a dokumentum letölthető a honlapomról
http://www.math.bme.hu/~makraim/temalabor