Egy magyar nyelvű szentimentkorpusz létrehozásának tapasztalatai

Szabó Martina Katalin[1,2], Vincze Veronika[3,4]

[1] Precognox Informatikai Kft.
[2] Szegedi Tudományegyetem, Orosz Filológiai Tanszék
mszabo@precognox.com; szabomartinakatalin@gmail.com
[3] MTA-SZTE Mesterséges Intelligencia Kutatócsoport
[4] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
vinczev@inf.u-szeged.hu

Kivonat: A jelen dolgozat egy magyar nyelvű kézzel annotált szentimentkorpusz létrehozásáról számol be. A korpusz építésének célja, hogy megfelelő segédletet teremtsünk a magyar nyelvű szövegek véleménykivonatolásával kapcsolatos nyelvtechnológiai feladatok, köztük a szentimentlexikonunk és az automatikus szentimentelemző rendszerünk hatékonyságának teszteléséhez és fejlesztéséhez. A korpusz emellett lehetőséget kíván nyújtani a magyar nyelvű szövegek szentimentelemzését érintő elméleti nyelvészeti problémák feltárására is, amely nélkülözhetetlen a szentimentelemző rendszer hatékony működésének biztosításához.

1 Bevezetés

A jelen dolgozatban a magyar nyelvű szövegek automatikus szentimentelemzését célzó kutatómunkánk egyik részfeladatáról, egy szentimentekre annotált korpusz létrehozásáról számolunk be.

A *szentimentelemzés* vagy *véleménykivonatolás* (*sentiment analysis* vagy *opinion mining*) a természetesnyelv-feldolgozás részterülete, amely a szerzői attitűdöt tükröző nyelvi elemek detektálására, valamint értékének (*sentiment orientation*) és tárgyának (*target*) a megállapítására törekszik automatikus megoldások segítségével.

A szentimentelemzés a nemzetközi kutatásban és fejlesztésben egyre nagyobb figyelmet kap, amelynek oka egyrészt a feladat elméleti nyelvészeti, valamint nyelvtechnológiai kihívásaiban, másrészt az eredmények gazdasági hasznosítási lehetőségeiben keresendő (pl. a tőzsdeindex mozgásának előrejelzése; a fogyasztói csoport benyomásai, tapasztalatai bizonyos termékek és szolgáltatások vonatkozásában; politikusokkal, politikai eseményekkel kapcsolatos attitűdök felmérése; választási előrejelzések stb.). Ugyanakkor, e növekvő nemzetközi figyelem ellenére a magyar nyelvű szövegek véleménykivonatolási feladatával csupán rendkívül csekély számú dolgozat foglalkozik. Emeljük ki közülük Berend és Farkas [1] dolgozatát, amely a kettős állampolgárság témájához kapcsolódó szövegek gépi tanuláson alapuló feldolgozását célozza, valamint az *Opinhu* rendszert [2], illetve az *OpinHuBank* projektet [3], amely

az internetes hírportálokon, blogokon és közösségi oldalakon publikált szövegek szentimentszintű annotálásának megoldására törekszik automatikus és manuális megoldások segítségével.

Ami a magyar nyelvű szövegek szentimentannotálását illeti, jelenleg egyetlen magyar nyelvű korpuszról van tudomásunk, az *OpinHuBank*ról [3], amelyben a korpusz építői a munka során a szentimentek annotálását célozták. Ugyanakkor, az elkészült korpusz több lényegi sajátsága okán elemzési és tesztelési célokra csupán korlátozottan alkalmazható. Egyrészt, a szövegekben a szentimentkifejezéseket egyenként nem annotálták a korpusz építői, a szentimentértékeket (pozitív vagy negatív) ugyanis magasabb, a mondatok vagy a tagmondatok szintjén határozták meg, az azon belüli további elemzés nélkül. Másrészt, az annotátoroknak az aktuális mondat szentimentértékének pozitív vagy negatív voltáról a mondatban szereplő tulajdonnévi entitás viszonylatában kellett döntést hozniuk, azaz arra kérték őket, hogy ítéljék meg, vajon pozitív vagy negatív ítéletet fejez-e ki az elemzett mondat a bennfoglalt PERSON (személynév) típusú entitás vonatkozásában. Mindez azért is problematikus, mert a szentiment targetjének szerepét a mondatban a személynéven kívül számtalan elem (pl. egy hely, egy esemény, egy termék vagy akár a termék egy aspektusa is) betöltheti. Az a sajátság tehát, miszerint a korpuszban kizárólag személynév tölti be a target szerepét, nyilvánvalóan jelentősen korlátozza az eszköz alkalmazhatóságát. Ugyanakkor, a legnagyobb problémát nem is ez a korlátozás jelenti. Bár a korpusz készítői hangsúlyozzák, hogy automatikus, majd kézi módszerrel kiszűrték azokat az eseteket, ahol a PERSON típusú entitás nem az adott mondat targetje, hanem a mondatban megfogalmazott vélemény forrása volt, a korpusz sajnálatos módon számos ilyen esetet tartalmaz; pl.

(1) Martonyi János leszögezte: noha a jelenlegi szlovák kormánykoalíció egyik pártjának vezetői gyakran elfogadhatatlan kijelentéseket tesznek, a magyar kormány nem ilyen stílusban fog reagálni (…)
[http://www.belfoldihirek.com/belfold/martonyi-janos-szlovakiaba-latogat]

A korpuszból idézett példa beláthatóan értékítéletet fogalmaz meg, azonban azt nem a mondat tulajdonnévvel jelölt entitásának viszonylatában teszi.

A fentebb leírt sajátságokat és problémákat megfontolva úgy döntöttünk, hogy szentimentelemző rendszerünk teszteléséhez és fejlesztéséhez, valamint a szentimentelemzés problémaköréhez kapcsolódó elméleti nyelvészeti és nyelvtechnológiai kutatások támogatása céljából létrehozunk egy olyan manuálisan annotált korpuszt, amely képes a magyar nyelvű szövegek véleménykivonatolásával kapcsolatos kutatói és fejlesztői feladatok hatékony támogatására.

2 A korpuszannotálás alapelvei és eszközei

A korpusz szöveganyagát a [http://divany.hu/] honlap termékvéleményeiből állítottuk össze. A honlap készítői időközönként bizonyos termékcsoportokat tesztelnek, s közzéteszik a tesztelők véleményét. A honlap szövegeiből 111-et gyűjtöttünk össze. A

nyers korpusz jelenleg összesen mintegy 13 000 mondatot és 190 000 tokent tartalmaz.

A manuális annotálás keretében a teljes értékelő kifejezést, azon belül pedig a pozitív és negatív polaritású szentimentkifejezéseket, azok targetjeit, valamint esetleges siftereit jelöltük be a korpuszban [4,5]. Szentimentkifejezésnek olyan egy szóból álló, vagy állandósult többszavas szókapcsolatokat tekintettünk, amelyek lexikai szinten értékítéletet hordoznak valamely target vonatkozásában [6,7]. Azokat a nyelvi elemeket, amelyek valamilyen módon hatást gyakorolnak a szövegekben megfogalmazott értékelő tartalmakra, az angol nyelvű terminológia alapján *szentimentsifterek*nek nevezzük, és külön taggel látjuk el a korpuszban [8,9].

2.1 A szentimentsifterek annotálása

A szentimentsiftereken belül két alapvető csoportot különböztethetünk meg. Az egyikbe azok az elemek tartoznak, amelyek a szentimentkifejezések szintaktikai kontextusában befolyásolják azok lexikális szintű, prior szentimentértékét, a másikba azok, amelyek a prior szentimentértékeket nem változtatják meg ugyan, azonban lehetetlenné teszik az értékelést megfogalmazó szövegrész faktív olvasatát. Az alábbiakban rövid áttekintést adunk e két átfogó kategóriáról.

Az első típusba az ún. negáló és az intenzifikáló elemek tartoznak. A szentimentértékek negálói a következő közös sajátsággal bírnak: vagy az ellenkezőjére változtatják a kifejezés prior értékét (2a), vagy pedig törlik azt (2b); pl.

(2) a. Mari nem szép. (’Mari csúnya’)
b. A béka nem gusztustalan. (nem jelenti azt, hogy ’gusztusos, tetszetős’)

A szentimentértékek negálói többek között lehetnek tagadószók (pl. *ne, sem, dehogy*), a létige tagadó alakjával (*nincs, nincsen, sincs, sincsen*), tagadó névutóval (pl. *hiányában, nélkül*) és egyéb módosítószók (pl. *aligha, látszatra*) [10].

A szentimentértékek ún. intenzifikáló elemei közé soroljuk azokat a nyelvi elemeket, amelyek a közös jellemzője, hogy a prior szentimentértéket egy bizonyos mértékben, valamilyen irányban módosítják, mégpedig úgy, hogy azt vagy erősítik (3a), vagy ellenkezőleg, csökkentik (3b); pl.

(3) a. A hangminőség nagyon jó.
b. A hangminőség aránylag jó.

A szentimentértékek intenzitásának befolyásolására számtalan elem alkalmas lehet, pl. rendkívül, rendkívüli módon, borzasztóan, elképesztően, valamennyire, valamelyest, feliből-nagyjából, részben, kevésbé stb. [11,12].

Ugyanakkor jegyezzük meg, hogy egy adott szentimentkifejezés prior értékére egy negáló és egy intenzifikáló elem is hatást gyakorolhat egyszerre; pl.

(4) A hangminőség nem nagyon jó.

A szentimentsifterek másik nagy kategóriájának elemeit irreálóknak nevezzük, és közéjük tartozónak tekintünk minden olyan nyelvi eszközt, amely lehetetlenné teszik

az értékelést megfogalmazó szövegrész faktív olvasatát. Másképpen, az irreálók megakadályozzák, hogy az adott szentimentet a megfogalmazó által tényként kezelt információként fogadjuk el. Vessük össze az (5) alatti, faktív olvasatú példát a (6) alatti, nem faktív olvasatú példákkal!

(5) A hangminőség jó.
(6) a. A hangminőség valószínűleg jó.
b. Lehet, hogy a hangminőség jó.
c. Jó a hangminőség?
d. Nem tudom, hogy a hangminőség jó-e.
e. A hangminőség jó lehet.

Amint látjuk, amíg az (5) alatti példában az értékelés megfogalmazója elkötelezi magát a propozíció igazsága iránt, addig a (6) alatti példák esetében nem, ennek következtében azok értékelő tartalmát nem is kezelhetjük a szentimentelemzés során teljes értékű adatként. Minden olyan elemet tehát, amely azt jelöli, hogy az értékelés propozíciós tartalmát a beszélő nem tényként tekinti, külön taggel láttuk el a korpuszban.

2.2 Az annotáció bemutatása

A feldolgozott szövegek sajátsága okán úgy döntöttünk, hogy a tesztelt termékek címbeli elnevezéseit *topic* címkével látjuk el, míg az egyes szentimentekhez kapcsolódó targetek *target* címkét kapnak.

A topikok és a targetek annotációszintű elkülönítése indokolható, hiszen a szentimentelemzés egy fontos része abban áll, hogy meg kell tudnunk különböztetnünk egymástól az entitásokat (*entity*), valamint azok aspektusait (*aspect*) [9]. Ennek a különbségtételnek a szentimentértékek súlyozásában jelentős szerepe van; egy adott szentiment ugyanis mind egy adott entitáshoz, mind annak csupán egy adott aspektusához is kapcsolódhat. Például, egy fényképezőgép mint entitás többek között a képminőség, a szín és az ár aspektusokkal rendelkezik. Az, hogy az értékelő az entitás, illetve az egyes aspektusok vonatkozásában milyen értékítéleteket közöl, nyilvánvalóan nagy jelentőséggel bír annak szempontjából, hogy magát az entitást hogyan értékeli; pl.

(7) Bár az ára nem volt alacsony, nagyon megérte ez a fényképezőgép.

Amint azt a fentebbi példa is mutatja, egy adott entitás egy adott aspektusáról tett negatív értékítélet nem jelent feltétlenül negatív értékítéletet a teljes entitás vonatkozásában. Ily módon az entitás–aspektus-kettősség az egyes szentimentértékek súlyozásában, ezáltal az aktuálisan elemzett szöveg összesített szentimentértékének a kiszámításában lényegi szereppel bír.

A korpuszban alkalmazott annotációt, miszerint a topikot megkülönböztetjük a targettől, a jövőben az entitás−aspektus-kettősség automatikus feldolgozásában is ki szeretnénk aknázni.

A korpusz annotációs megoldását az alábbi példával szemléltetjük:

```
(8) Negyedik helyezett: <topic>Kolios goat's
cheese</topic>
„<SentNeg> <target>Állagra</target> olyan, mint a
<SentiWordNeg>gumi</SentiWordNeg> </SentNeg>, <SentNeg>
<target>ízre</target> pedig
<SentiWordNeg>fanyar</SentiWordNeg> </SentNeg>.
<SentNeg> Nekem <ShiftNeg>nem</ShiftNeg>
<SentiWordPos>jön be</SentiWordPos> </SentNeg>.”
```

A szentimentsifterek e kezelési megoldásával alapot kívánunk teremteni egy magyar nyelvű szövegekre alkalmazható szentimentérték-kalkulátor, a *SOCal-Hun* létrehozásához [5,13].

3 A korpusz adatai

Az annotálás során a nyers szövegkorpuszból 15 szöveget dolgoztunk fel, ami összesen 1834 mondatot és 26 503 tokent tartalmaz.

Az annotáció egyetértési adatait az alábbi táblázat foglalja össze:

1. táblázat. Az annotáció egyetértési adatai

az annotált tag	F-mérték
PosSentiment	0,36
NegSentiment	0,40
SentiWordPos	0,68
SentiWordNeg	0,60
Topic	0,99
Target	0,53
Negation	0,68
IntensifierPlus	0,57
IntensifierMinus	0,63
Irreal	0,17
OtherShifter	0,30

Amint az a táblázat statisztikái alapján látható, a legnagyobb egyetértési arányt a topikok annotálásában értük el. Ez nem meglepő, hiszen topic címkével – a már említetteknek megfelelően (l. fentebb) – a tesztelt termékek tulajdonnévi jelölőit láttuk el, amelyek megtalálása és terjedelmének megállapítása nem okozhatott különösebb nehézséget az annotátorok számára. Megfelelő eredményességet produkáltunk továbbá a negáló kifejezések (Negation), az intenzifikáló sifterek (elsősorban az IntensifierMinus tag esetében), valamint a szentimentkifejezések (SentiWordPos és SentiWordNeg) annotálásában.

A targetek annotálásában már kevesebb eredményességgel dolgoztunk. Az annotáció kézi ellenőrzése arra mutatott rá, hogy az eltérés alapvetően a feldolgozott szöve-

gek domén-sajátságára vezethető vissza. Mivel az annotált korpusz termékvéleményeket tartalmaz, a tesztelők által megfogalmazott értékelések rendre a tesztelt termékek különböző aspektusaira irányulnak, azokat minősítik. Ennek köszönhetően a feldolgozott szövegek rendkívüli mennyiségű targetet tartalmaznak, amelyből számos példány elsikkad a feldolgozási munka során.

Még kisebb egyetértést mértünk a teljes szentimentegységek annotálását illetően, amelynek oka – a kézi ellenőrzés tapasztalatai alapján – egyértelműen abban keresendő, hogy a korpusz feldolgozását végző két annotátor eltérően kezelte a többszörös mellérendelő szerkezeteket: amíg az egyik annotátor azok tagjait rendre külön-külön egységekként annotálta, addig a másik gyakorta egyetlen szentimentként jelölte őket. Ez alapján feltétlenül szükségesnek tartjuk az erre vonatkozó annotálási alapelvek pontosabb rögzítését.

A legkisebb hatékonyságot az ún. irreáló elemek taggelésében értük el. Ennek valószínű oka az, hogy az irreálás jelensége, ahogyan azt már korábban a (6) alatti példákkal is igyekeztünk megmutatni (l. fentebb), számos formában jelenhet meg a szövegekben, és e sokféleségnek az egységes kezelése nehézséget okozhatott az annotátorok számára.

Az alábbi táblázat összefoglalja az annotált korpuszrész statisztikai adatait:

2. táblázat. Az annotáció statisztikai adatai

annotált tag	**darabszám**
PosSentiment	603
NegSentiment	743
SentiWordPos	708
SentiWordNeg	827
Topic	169
Target	528
Negation	316
IntensifierPlus	332
IntensifierMinus	68
Irreal	66
OtherShifter	30
ÖSSZESEN:	4390

Az annotáció fentebbi statisztikai adatai alapján a következő megállapításokat tehetjük:

A negatív véleményt megfogalmazó kifejezések (NegSentiment) többségben vannak a pozitív véleményt megfogalmazó kifejezésekkel (PosSentiment) szemben. Hasonló megoszlást találunk a szentimentkifejezések között is, ami azonban nem következik szükségszerűen az előbbi megállapításunkból, hiszen negatív vélemény pozitív szentimentkifejezéssel, illetve pozitív vélemény negatív szentimentkifejezéssel is megfogalmazható, amennyiben a kifejezés lexikai szintű polaritását egy sifter segítségével megváltoztatjuk. Ennek ellenére a táblázat adatai alapján azt látjuk, hogy a lexi-

kai szinten negatív polaritással rendelkező kifejezések fordulnak elő nagyobb számban a korpusz általunk feldolgozott részében. Az annotáció tapasztalatai meglepőek az ún. Pollyanna-hipotézis tükrében, amely nyelvi univerzáléként tételezi a pozitív töltetű kifejezések magasabb használati arányát a negatív töltetű nyelvi elemekkel szemben [14]. Mindezek alapján a megfigyelt jelenséget szeretnénk nagyobb mennyiségű annotált szöveganyagon behatóbb vizsgálat tárgyává tenni a jövőben.

Ugyancsak szembeötlő eltérés mutatkozik az intenzifikáló elemek gyakorisági megoszlásában, hiszen a fokozó típusúak (IntensifierPlus) túlnyomó többségben szerepelnek a mérséklő típusú elemekkel (IntensifierMinus) szemben. Valószínűsíthető, hogy a mért adatok összhangban állnak Székely megállapításával, miszerint a magyar nyelvben (s talán nem csak a magyar nyelvben) a mérséklés eszközrendszere szegényesebb a fokozás eszközrendszerénél [12].

Végezetül emeljük ki, hogy az annotált korpuszrész 316 negáló kifejezést (Negation) tartalmaz (ebből 140 pozitív és 176 negatív polaritású véleményben szerepel), ami jelentős előfordulási aránynak tekinthető annak fényében, hogy összesen 1346 szentimentet azonosítottunk a munka során. Az eredmény arra mutat, hogy a negáció feltétlen megoldást sürget a szentimentelemzés feladatában, hiszen figyelembe nem vételük jelentős torzulást okozhat az elemzés során kapott szentimentértékeket tekintve.

4 A korpusz felhasználási lehetőségei

Az annotált korpusz nyelvtechnológiai feladatokban és elméleti nyelvészeti kutatásokban − így tesztelési és fejlesztési célokra − egyaránt alkalmazható.

A kutatómunka következő lépéseként szeretnénk az annotációt nagyobb mennyiségű szövegre kiterjeszteni, majd az annotált korpuszt beható empirikus vizsgálat tárgyává tenni. Terveink szerint a korpuszban alkalmazott annotációra támaszkodva sikerül kialakítanunk egy olyan automatikus szentimentelemző rendszert, amely képes a szentimentkifejezéseket azok targetjeivel és siftereivel összefüggésben hatékonyan kezelni a jövőben.

Köszönetnyilvánítás

A jelen kutatás a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt keretében az Európai Unió támogatásával és az Európai Szociális Alap társfinanszírozásával valósult meg.

Hivatkozások

1. Berend, G., Farkas, R.: Opinion Mining in Hungarian based on textual and graphical clues. In: Proceedings of the 8th conference on Simulation, modelling and optimization. Stevens

Point, Wisconsin, USA, World Scientific and Engineering Academy and Society (WSEAS) (2008) 408–412
2. Miháltz, M.: OpinHu: online szövegek többnyelvű véleményelemzése. In: Tanács, A., Vincze, V., eds.: VII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2010), Szegedi Tudományegyetem, Szeged (2010) 14–23
3. Miháltz, M.: OpinHuBank: szabadon hozzáférhető annotált korpusz magyar nyelvű véleményelemzéshez. In: Tanács, A., Vincze, V., eds.: IX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2013), Szegedi Tudományegyetem, Szeged (2013) 343–345
4. Ding, X., Liu, B., Yu S., Ph.: A holistic lexicon-based approach to opinion mining. In: Najork, M., Broder, A. Z., Chakrabarti, S. eds.: Proceedings of the Conference on Web Search and Web Data Mining (WSDM-2008), New York, NY, USA (2008) 231–240
5. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics 37/2, Association for Computational Linguistics, MA, USA, MIT Press Cambridge (2011) 267–307 http://dl.acm.org/citation.cfm?id=2000518
6. Vincze, V.: Félig kompozicionális főnév + ige szerkezetek a Szeged Korpuszban. In: Gecső, T., Sárdi, Cs., eds.: Új módszerek az alkalmazott nyelvészeti kutatásban, Budapest, Tinta (2010) 327–332
7. Szabó, M. K.: Egy magyar nyelvű szentimentlexikon létrehozásának tapasztalatai. In: Nyelv, kultúra, társadalom konferencia konferenciakötete (2014) (megjelenés előtt)
8. Szabó, M. K.: A magyar nyelvű szövegek szentimentelemzésének dilemmái, különös tekintettel a szentimentsifterek kezelésére. LingDok 18. Nyelvészdoktoranduszok 18. országos konferenciája, Szeged (2014)
9. Liu, B.: Sentiment Analysis and Opinion Mining. Draft (2012) http://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf
10. Pete, I.: Az állító és tagadó mondatok szinonimiája a magyarban. Magyar Nyelv 95/3. (1999) 305–312
11. Moilanen, K., Pulman, S.: Sentiment Composition. In: Proceedings of Recent Advances in Natural Language Processing (RANLP 2007) (2007) 378–382
12. Székely, G.: Egy sajátos nyelvi jelenség, a fokozás. In: Segédkönyvek a nyelvészet tanulmányozásához 66. Budapest, Tinta (2007)
13. Brooke, J., Tofiloski, M., Taboada, M.: Cross-linguistic sentiment analysis: From English to Spanish. In: Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing, Borovets (2009) 50–54
14. Boucher, J., Osgood, C.: The Pollyanna hypothesis. Journal of Verbal and Learning Behavior 8/1 (1969) 1–8