Panaszlevelek automatikus kategorizálása szerkezeti egységek és jellemző kifejezések figyelembevételével

Bártházi Eszter[1,2⋆], Héder Mihály[3,4] ⋆⋆

[1] MTA SZTAKI Géppel Támogatott Megértés Kutatócsoport, barthazi@sztaki.hu
[2] Szegedi Tudományegyetem Nyelvtudományi Doktori Iskola Elméleti Nyelvészet Program
[3] MTA SZTAKI Internet Technológiák és Alkalmazások Központ, mihaly.heder@sztaki.hu
[4] Budapesti Műszaki és Gazdaságtudományi Egyetem Filozófia és Tudománytörténet Tanszék

1. Bevezető

2008-ban indult kutatásunk célja, hogy egy rendszert készítsünk, amely egyszerre könnyíti meg valamely hivatal és a hozzá forduló ügyfelek dolgát. A gép közreműködésének lényege, hogy az ügyfél számára egy felületet nyújt, ahol panaszát, hozzászólását (továbbiakban levelét) megfogalmazhatja. A levél írása során az elképzelt rendszer dialógusok formájában kapcsolatot teremt a levélíróval, kérdések segítségével pontosabb információkat kér, megpróbálja eldönteni, hogy a levél milyen hivatali kategóriába tartozik.

Ezen elképzelt rendszer megvalósításához mindenek előtt kiterjedt alapozó kutatások szükségesek. A jelen cikk ezen kutatásokat, kísérleteket mutatja be, melyeket az Igazságügyi Minisztériumtól kapott, nagyon változatos, közel 900 levélből álló korpuszon (a továbbiakban korpusz) hajtottunk végre. A korpuszért külön köszönetet szeretnénk mondani dr. Vörös Editnek, az Igazságügyi és Rendészeti Minisztérium Társadalmi Kapcsolatok Osztálya vezetőjének, aki nemcsak rendelkezünkre bocsátotta a szövegkorpuszt, hanem gondoskodott a felhasználás jogi és előfeldolgozási körülményeiről.

A korpusszal végzett munka első lépése az előfeldolgozás volt. Ehhez egy alkalmas keretrendszert készítettünk, amely integrálja a magyar nyelvre elérhető különféle elemző eszközök jelentős részét. A 2. fejezet ezt a keretrendszert mutatja be.

A feladat sajátossága, hogy a korpusz szókincse kivételesen terebélyes, inhomogén, hivatalosnak egyáltalán nem nevezhető. A levelek fogalmazása is gyakran hiányos, nehezen értelmezhető, és sok helyesírási hibát, elírást tartalmaz. Mivel az automatikus kategorizáló rendszerek igazán jó teljesítményt csak egy

⋆ PhD hallgató, témavezető: Németh T. Enikő
⋆⋆ PhD hallgató, témavezető: Vámos Tibor

jól behatárolható terület szaknyelvi kontextusában szoktak elérni, kiemelten sok energiát kell fordítanunk arra, hogy a hétköznapi és a szaknyelv között kapcsolatot teremtsünk, illetve hogy a levelek által meghatározott túlságosan tág, emiatt nehezen kezelhező kontextust leszűkítsük. Egy kísérlet a kontextus szűkítésére a szerkezeti egységek detektálása és felhasználása a kategorizálási kísérletek során. A szerkezeti egységek jelentősége, hogy segítséget nyújtanak abban, hogy bizonyos típusú információkat hol érdemes keresni. Például a kategorizálás szempontjából lényeges részek többnyire a problémát bemutató szerkezeti egységben találhatók, míg a levélíróról rendelkezésre álló adatok jellemzően a Bemutatkozás szerkezeti egységben keresendők. Ezt az elképzelést részletesebben a 3. fejezet bontja ki.

A következő lépés a rendszer megvalósítása felé a jó eredményekkel működő osztályozási és csoportosítási algoritmusok kipróbálása a korpuszon. Ez egyrészt információt szolgáltat számunkra a meghaladni kívánt pontosságról, másrészt a végleges rendszerben is fel szeretnénk használni a kategorizálás, illetve a kategorizálási javaslatra vonatkozó dialógus megvalósításánál. A részleteket a 4. fejezet tartalmazza.

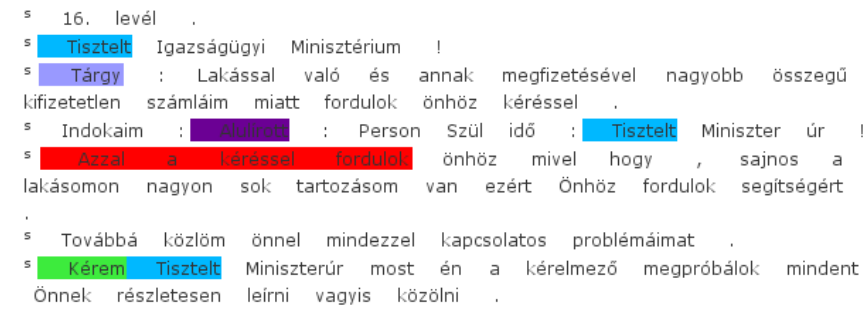
A szerkezeti egységek ismeretében már lehetőségünk van egy speciálisabb, az ügyintéző feltételezett gondolatmenetét modellező kategorizálási eljárás készítésére. Feltevésünk szerint az ügyintéző a levelek feldolgozásánál forgatókönyveket követ. A KATEGORIZÁLÁS forgatókönyv egy általunk formalizált algoritmus, ami megadja a panaszkategóriába való soroláshoz vezető lépéseket. A gép az algoritmust bejárván, miután azonosította az ügyfelet, azokat a szerkezeti egységeket vizsgálja meg, amelyek a kategorizálás szempontjából releváns kifejezéseket taralmazzák. A KATEGORIZÁLÁS forgatókönyv részleteit az 5.1. fejezet fejti ki. A jövőben a besorolástól függően újabb, immár kategóriaspecifikus forgatókönyvekkel is szeretnénk kísérletezni.

Számos, a kategorizálás szempontjából irreleváns, ugyanakkor egyéb – szociológiai, valamint pszichológiai – szempontból fontos információ is található a levelekben. A levelek besorolását, és még inkább a dialógusok alakítását befolyásolja a levélíró pszichés-szociológiai profilja, amelyet a használt kifejezések és fordulatok, a szerkesztési jegyek alapján folyamatosan építünk. A profil meghatározásához egy, A Magyar nyelv értelmező szótárára [10] épülő, a szavak stilisztikai jegyeit tartalmazó listát használtunk. A részleteket az 5.2. fejezetben mutatjuk be.

Ezen kutatás közvetlen előzményének tekinthető Héder diplomaterve [6], amely szemantikus annotációk géppel támogatott elhelyezését tárgyalja webes dokumentumokban. Abból a munkából eszközöket és sok tapasztalatot sikerült átmenteni, de hiányzott belőle a szöveg nyelvi, szerkezeti elemzése és a profil készítése.

2. A használt keretrendszer

Kutatásunkhoz egy egyszerűen használható, általános előfeldolgozó, illetve nyelvi elemző rendszert készítettünk, melynek segítségével sok különféle, kész eszközt



1. ábra. Egy DMD fájl vizuális megjelenése

homogén módon tudunk kezelni. A fejlesztés fő követelményeiként a könnyű használhatóságot, az új eszközök minél egyszerűbb integrációját és a robusztusságot jelöltük meg. Mivel a fő célunk nem eszközfejlesztés, törekedtünk minden elérhető megoldás beépítésére.

Az így elkészült rendszer bemenete egy egyszerű szövegfájl vagy strukturált XML dokumentum lehet. A kimenet egy úgynevezett Docuphet Mixed Document (DMD) típusú XML fájl, amelyet több névtérből gyúrtunk egybe, úgy, hogy az lehetőleg minden elképzelhető annotációtípust hordozni tudjon. A DMD saját hordozó névterén kívül definiáltunk egy névteret a projektben létrehozott eszközeink számára is. A többi névtér a felhasznált külső eszközök annotációit reprezentálja. Használtuk a Hitec projekt [4] kapcsán kifejlesztett fulldoc formátum egyes elemeit és a Huntools jól ismert komponenseit, a Huntokent, a Hunmorphot és a Hunpost.

A névterek éles megkülönböztetése révén megpróbáljuk a jövőbeli feldolgozó eszközök számára minél egyszerűbbé tenni az általuk ismert névterek elemeinek kezelését, miközben az ismeretlen névtereket figyelmen kívül hagyhatják. Ezzel egyidejűleg lehetővé tesszük a rendszerünk zökkenőmentes kiterjesztését is.

A DMD fáljoknak van egy egyszerű, informatívnak és tetszetősnek szánt XHTML megjelenítése is (2 ábra). A DMD fájl XHTML formátumba való konvertálását XSLT 2 transzformációval végezzük.

Korábbi saját fejlesztés [6] az eredetileg névelemek, később a tipikus szerkezeti egységek (lásd 3. fejezet) felismerésére használt JNER rendszer. A java nyelven íródott eszköz szabályok és katalógusok segítségével végzi feladatát.

Az egyes névterekkel jelölt annotációkat különféle szkriptek lefuttatása állítja elő. Vannak a Huntools egyes elemeit, illetve a JNER-t egy-egy fájlon lefuttató szkriptek, mások minden feladatot kötegelten, esetleg egész könyvtárakra hajtanak végre. Készítettünk eszközöket a szó, szótő és egyéb típusú statisztikák gyűjtésére is. Megemlítendő, hogy az integrált, minden elemzést egyben elvégző megoldáshoz webes felületet is készítettünk, ahol a beírt szöveg AJAX technológia segítségével a háttérben feldolgozásra kerül.

3. Szerkezeti egységek annotálása

A vizsgált panaszlevelek esetében megfigyelhető, hogy az állampolgárok jelentős része a hétköznapi szókincsére támaszkodva pontatlanul, hiányosan, sok esetben nehezen érthetően fogalmazza meg a panaszát, és számos, az ügyintézés szempontból irreleváns információt is közöl. Ezzel megjósolhatatlanul tág kontextusba helyezi a levélben megfogalmazottakat. Továbbá a levelek szerkezeti felépítése is igen változatos, ezért pusztán a közigazgatási területekre jellemző terminológiára támaszkodva egy bottom-up megközelítéssel nagyon nehéz jó eredményt elérni. Ennek a problémának a megoldásaként, a leveleket alaposan megvizsgálva tizenkét szerkezeti egységet találtunk, amelyek egyben kontextusként, értelmezési keretként is szolgálnak a bennük előforduló kifejezések interpretálásához. A tizenkét szerkezeti egység a következő:

1. **Megszólítás**: a levélíró valamilyen módon kifejezi, hogy kinek szánja levelét, pl.: *Tisztelt [személynév/titulus/intézménynév/stb.]*
2. **Bemutatkozás**: a levélíró azonosításához szükséges adatokat tartalmazza, pl.: *Alulírott, [személynév], született [évszám], anyja neve [személynév]* stb.
3. **Cél**: a levélíró még a panasza ismertetése előtt kifejezi, hogy milyen területen vár segítséget, pl.: *Tárgy: nyugdíjügy.*
4. **Előzmény**: a jelenlegi problémát megelőző, de ahhoz kapcsolódó események ismertetése, pl.: *Kértem a miniszter urat, hogy...*
5. **Probléma**: a levélíró a problémáját részletezi, pl.: *Az alábbi problémámra várnám a segítséget.*
6. **Javaslat**: a Probléma szerkezeti egység alternatívája, amikor a levélíró nem egy megoldásra váró problémával fordul a minisztériumhoz, csupán egy javaslatot tesz valamivel kapcsolatban, pl.: *A következő javaslattal fordulok Önökhöz...*
7. **Vádaskodás**: a levélíró indulatait, kétségeit fejezi ki, erősen emocionális módon, pl.: *Hol itt a törvény?*
8. **Elismerés**: a levélíró elismerését fejezi ki a levél címzettjének eddigi tevékenységével szemben, pl.: *Engedje meg, hogy gratuláljak.*
9. **Egyéb körülmények**: a levélíró a problémájához szorosan nem vagy egyáltalán nem kapcsolódó egyéb problémáját, életkörülményeit, egészségügyi állapotát stb. ecseteli, pl.: *Az igaz hogy jobb kezem az ujjam hegyétől a vállamig és az egész törzsem a derekamtól a fejem hegyéig zsibog a jobboldalamon* — egy öregségi nyugdíjemelési kérelemről szóló levélben.
10. **Elvárás**: a levélíró azt fogalmazza meg, hogy milyen viselkedést, intézkedést vár el az ügyintéző részéről, pl.: *A fentiek alapján kérem...*
11. **Köszönet**: a levélíró megköszöni az eddigi intézkedést, türelmet, illetve előre is megköszöni a további intézkedéseket, pl.: *Előre is köszönöm, hogy válaszlevelével megtisztel.*
12. **Lezárás**: a levélíró egy adott formulával befejezi a levelét, pl.: *Minden jót.*

Az egyes szerkezeti részek sorrendje levelenként eltérő lehet, és természetesen nem minden szerkezeti egység található meg minden levélben. Az azonban,

hogy mely szerkezeti egységek fordulnak elő egy adott levélben, valamint az is, hogy milyen sorrendben, további információval szolgálhat a levélíróval kapcsolatban. A szerkezeti egységeknek köszönhetően az információkinyerés egyszerre bottom-up (jellemző kifejezések figyelembevétele a kategorizálás során) és top-down folyamatok eredménye (egy bizonyos kontextusban/értelmezési keretben történik), amely azért is fontos, mivel a humán megértés során a kontextus ismerete éppúgy irányítja az interpretációt, mint az egyes kifejezések jelentése (a kompozicionalitás és a kontextualitás elvének együttműködése, lásd [8]).

A szerkezeti egységek felismeréséhez a levelek 10%-ának manuális elemzésével elkészítettük az egyes egységeket tipikusan jelölő definitív kifejezések listáját. A lista alapján a JNER segítségével annotáltuk a leveleket, az annotációk megjelenítéséhez színkódokat használtunk (lásd a 2. ábrát).

A megoldás tesztelése azt az eredményt hozta, hogy a lista még kiegészítésre szorul, ugyanis sok levélben csak kevés szerkezeti egységet találtunk így. Ennek oka feltételezhetően kettős: egyrészt az általunk vizsgált 89 levél valószínűleg nem reprezentálja a teljes korpuszt megfelelően; másrészt a levelekre jellemző szóhasználat sokkal változatosabb annál, mint amit ezzel az egyszerű módszerrel jelenleg kezelni tudunk. Ugyanakkor jó eredménynek tartjuk, hogy a felismert szerkezeti egységek többnyire helytállóak. A helyesen felismert szerkezeti egységek százalékos arányát megfelelő tesztadatok hiányában egyelőre nem tudjuk megállapítani.

4. Osztályozási és csoportosítási kísérletek

4.1. A korpusz

A kutatás alanyául szolgáló korpusz az Igazságügyi Minisztériumhoz beérkezett 888 levél digitális, anonimizált verziójából állt. A levelekről általánosan elmondható, hogy igen szerteágazó témakörökben és nagyon változó stílusban, illetve helyesírással íródtak. Továbbá sok levél nyilvánvalóan felfokozott érzelmi állapotban (düh, elkeseredettség) íródott, értékes alapanyagot szolgáltatva ezáltal a levélírók különféle profiljainak meghatározásában.

A közel kilencszáz levélből Kabai Dóra munkája[2] nyomán 210-hez rendelkezésünkre állt kategória információ is. Ezen levelek 10 kategóriába voltak besorolva. Némely levél több kategóriába is tartozott egyszerre, így a kategóriabesorolások összesített levélszáma 330 volt.

4.2. Szűrés

A korpuszon először különféle szűrési eljárásokat próbáltunk ki. Feltételezésünk szerint a szűrésnek nagy szerepe lehet a csoportosítás és osztályozás hatékonyságának növelésében, de még nagyobb a gépi megértést nem befolyásoló, vagy zavaró zaj csökkentésében.

Az egyes szűrési eljárásokkal eredeti szóalakokat tartalmazó, illetve csak szótöveket tartalmazó tanulóadat-verziót is előállítottunk. A szótöveket minden esetben a HumMorph segítségével állapítottuk meg.

A legegyszerűbb szűrésünk azon szóalakok kihagyása volt, amelyek több, mint a levelek 50 %-ában szerepelnek. Épp 50 ilyen szóalakot találtunk. Ide soroltuk továbbá az egyéb karaktereket is. Ezt a szűrési típust a továbbiakban H betűvel jelöljük.

Az egyszerű, ökölszabályszerű H szűrés mellett kézzel is készítettünk egy szűrőlistát. A lista elkészítése során figyelembe vettük a szavak eloszlását is a levelekben, ebben a Weka rendszer volt segítségünkre [5]. Ez a lista a H listával szemben nem szóalakokat tartalmaz, hanem 235 szótövet (pl.: ha, mert, stb.) , illetve a HunMorph különféle morfológiai elemzési kimenetei közül 111-et (pl.: DET, ART, PUNCT, stb.). Ezt a szűrőlistát elsősorban a csoportosítás, illetve osztályozás hatékonyságának növelésére szántuk. Az információkinyerés és megértés szempontjából alkalmazásuk nem feltétlenül célszerű, mert kiszűri többek között a tagadószavakat, számneveket, illetve a létigéket is, ezáltal információveszteséget termelve. Ezt a szűrést a továbbiakban K-nak nevezzük.

A H és K globálisan alkalmazott szűrési eljárások mellett nagy reményekkel kísérletezünk egy, az egyes leveleken külön-külön kiértékelendő szűrési metódussal is. Ennek során megkíséreljük azonosítani a levelek szerkezeti egységeit, és a kategorizálás szempontjából irreleváns mondatokat — jelenleg: Megszólítás, Lezárás, Vádaskodás — teljes egészében kivesszük. A strukturális elemek azonosításáról a 3. fejezet szól. Ezen szűrést a továbbiakban S-nek nevezzük.

A szűretlen levelek összesen 425 ezer tokenből állnak — így kb. 450 token/levél adódik. Ha kiszűrjük a többszörös előfordulásokat, 53 ezer különböző tokent kapunk. Szótövekre ugyanezen számok 318 ezer (az egyéb karaktereknek, mondatvégi jeleknek nincs szótöve, ezért a különbség) és 13,5 ezer. Az egyes szűrések alkalmazásával az összméret kevesebb, mint a felére csökkenthető, és az egyedi szóalakok illetve szótövek száma is csökken. Külön kiemelendő, hogy az S szűrés kb. 30 ezerrel csökkenti az összesített méretet, de még az egyedi számokat is csökkenti néhány százzal.

Az adathalmazokból a Weka által feldolgozható Vektor (arff) fájlokat készítettünk. Itt két további szűrést alkalmaztunk: elhagytuk a kevesebb, mint háromszor szereplő elemeket, illetve összevontuk a kicsi és nagybetűs verziókat. Más korlátozást a vektor dimenzióinak méretére (az arff attribútumok számára) nem tettünk.

4.3. Kategorizálás

A célunk az volt, hogyan megvizsgáljuk, javítható-e a kategorizálás pontossága és hatékonysága a különféle szűrések segítségével. A kísérleteket a 210 kategorizált levéllel végeztük, úgy, hogy a levelek kétharmadát tanításra, a fennmaradó egyharmadot tesztelésre használtuk fel. Két elterjedtnek mondható algoritmust is kipróbáltunk.

A Naive Bayes [9] és az SVM[1] eljárásokat a Weka keretrendszer által alapértelmezettként felkínált paraméterekkel futtattuk. SVM implementáció gyanánt a libsvm rendszert vettük igénybe a Wekán keresztül.

A kísérleteket elvégeztük a szűretlen leveleken, illetve a szűrők H, H+K, S, S+H, S+H+K kombinációival átrostált leveleken is. Minden tesztet elvégeztünk

a szótő, illetve szóalak vektorokon is. A 1. táblázat tájékoztat az egyes lefutások időigényéről, illetve a helyesen kategorizált levelek százalékáról, zárójelben a pontos levélszámmal.

Elmondható, hogy bár a pontosságot nem befolyásolta lényegesen a szűrések alkalmazása, a legjobb eredményeket döntően az összes szűrő együttes alkalmazásával kaptuk. Eközben a futási idők jelentősen csökkentek. Az is kiemelendő, hogy ezen a korpuszon a szóalak és szótő vizsgálata között az összes szűrés együttes használata mellett nincs különbség. Érdekesség ugyanakkor, hogy a Bayes módszernél sokkal jobb pontosságú SVM kiegyensúlyozott eredményt hoz a szóalakok esetén a szűréstől függetlenül, de érzékeny a szűrés hiányára a szótövek esetében. A Bayes módszer ezzel szemben a szóalakok esetében jobban működik ha erős szűrést alkalmazunk, a szótövek esetében viszont épp fordítva: gyengülő teljesítményt mutat a szűrések hatására. A szóalak vektorok dimenzióinak száma kb. kétszer nagyobb, mint a szótövek dimenzióinak száma, ami feltehetően szerepet játszik a tapasztalt eltérésben.

Összegzésként elmondható, hogy — bár a kategorizált levelek kis száma nem engedi meg nagyon erős általánosítások tételét — a szűrés semmi esetre sem rontott az osztályozás pontosságán, ugyanakkor a futási időket és a feldolgozandó adatmennyiséget jelentősen csökkentette.

1. táblázat. Az osztályozási kísérletek eredményei

Típus	Szűrés típus	Naive Bayes	Futási idő(s)	SVM	Futási idő(s)
Szóalak		14.2857 % (16)	8.65	29.4643 % (33)	2.06
Szóalak	H	14.2857 % (16)	8.49	30.3571 % (34)	2.07
Szóalak	H+K	16.0714 % (18)	6.05	30.3571 % (34)	1.13
Szóalak	S	15.1786 % (17)	7.66	29.4643 % (33)	2.4
Szóalak	S+H	15.1786 % (17)	7.49	30.3571 % (34)	2.13
Szóalak	S+H+K	16.9643 % (19)	5.32	30.3571 % (34)	1.26
Szótő		16.9643 % (19)	4.47	25.8929 % (29)	1.59
Szótő	H	16.9643 % (19)	4.1	28.5714 % (32)	2.09
Szótő	H+K	16.9643 % (19)	3.49	30.3571 % (34)	1.11
Szótő	S	16.9643 % (19)	4.02	25.8929 % (29)	1.57
Szótő	S+H	16.0714 % (18)	3.85	28.5714 % (32)	1.42
Szótő	S+H+K	15.1786 % (17)	3.08	30.3571 % (34)	1.17

4.4. Csoportosítás

Néhány kísérletet elvégeztünk az X-mérték [7] csoportosítási algoritmussal is. Ezen algoritmus sajátossága, hogy a csoportok számát is képes adaptívan meghatározni, ugyanakkor a vágáshoz egyszerű K-mérték eljárást használ. A számunkra érdekes kérdés az volt, hogy a sok helyen előforduló, de lényegi információt nem tartalmazó szavak/mondatok szűrése segíti-e a csoportok elkülönülését. Ezért az algoritmust minden esetben tíz iterációban futtattuk, a kapott csoportok számát 2 és 30 közé limitálva.

Ahogy a 2. táblázatban látható, a csoportok számát a szűrések nem módosítják, ellenben a csoporthozzárendelésekre jelentős hatásuk van. Ezen jelenség okát az attribútumeloszlások és csoport hozzárendelések emberi vizsgálata tárhatná fel.

2. táblázat. A csoportosítási kísérletek eredményei

Típus	Szűrés típus	Csoportok száma	eloszlások
Token		4	104(12%), 329(37%), 209(24%), 245 (28%)
Szótő		4	67(8%), 214(24%), 271(31%), 335(38%)
Szótő	H	4	79(9%), 297(33%), 207(23%), 304(34%)
Szótő	H+K	4	106(12%), 299(34%), 188(21%), 294(33%)
Szótő	S+H+K	4	97(11%), 285(32%), 213(24%), 292(33%)

5. Forgatókönyv és profil

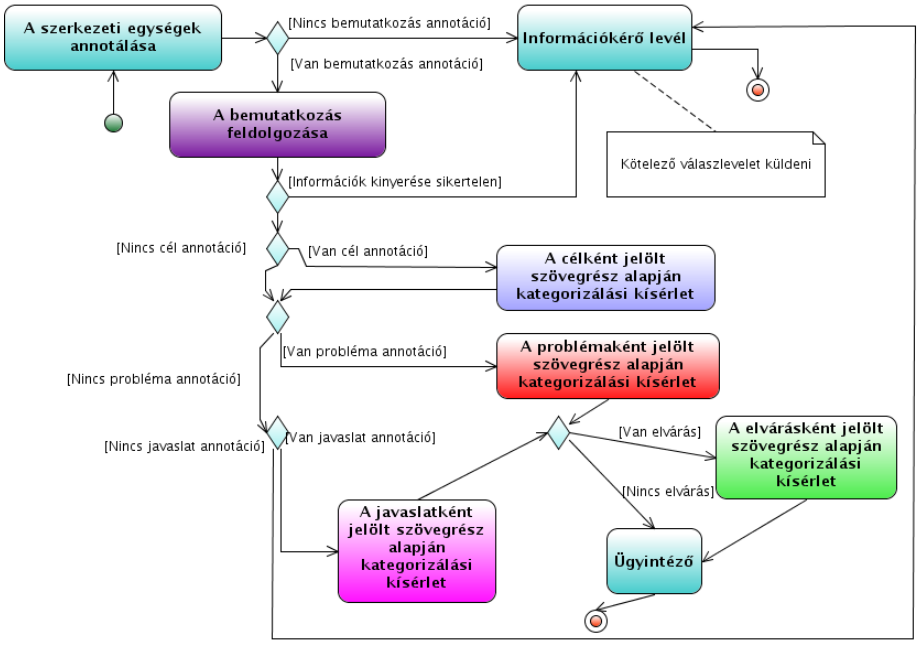
Az egyes panaszlevelekből két célból szeretnénk információt kinyerni. Az egyik cél az, hogy megállapítsuk, hogy melyik panaszkategóriába tartozik az adott panaszlevél, a másik, hogy az azt beküldő ügyfélről egy profilt alakíthassunk ki, az ügyfél aktuális érzelmi állapotáról, szociális körülményeiről tehessünk megállapításokat, amelynek majd a későbbiekben, a dialógusokban lesz fontos szerepe.

5.1. A kategorizálás forgatókönyv

A kategorizáláshoz elképzelésünk szerint a következő szerkezeti egységeket kell figyelembe venni: Bemutatkozás, Cél, Probléma, Javaslat, Elvárás. Ezek a levélnek azon részei, amelyek a panaszkategória megállapításához szükséges definitív kifejezéseket tartalmazzák, tehát azokat a nyelvi elemeket, amelyek alapján eldönthető, hogy az adott levél írója milyen kategóriájú panasszal fordul az ügyfélszolgálathoz. A Bemutatkozás szerkezeti egység figyelembevétele pedig azért alapvető, hogy az állampolgár egyértelmű azonosítása lehetővé váljon.

A kategorizálás általános forgatókönyvét a 5.1. ábrán látható aktivitás diagram mutatja be. Az algoritmus először a Bemutatkozás szerkezeti egységet keresi, hogy ezt feldolgozva kinyerhesse azokat az információkat, amelyek segítségével egyértelműen azonosítható az ügyfél. Ezen szerkezeti egység azonosítása a jellemző definitív kifejezések alapján történik.

A definitív kifejezések kétfélék lehetnek: kategóriasemlegesek vagy kategóriaspecifikusak. A kategóriasemleges kifejezések kizárólag az adott szerkezeti egység beazonosításában játszanak szerepet. A kategóriaspecifikus kifejezések szintén segíthetnek az adott szerkezeti egység beazonosításában, de nem ez az elsődleges feladatuk, hanem az, hogy az egységen belül a kategorizáláshoz szükséges, tartalmi szempontból releváns információkat hordozzák. Másrészt a kategóriaspecifikus kifejezések megtalálását a kategóriasemleges kifejezések segíthetik.



2. ábra. A KATEGORIZÁLÁS forgatókönyv aktivitás diagramja

Ha a rendszer nem talál bemutatkozás annotációt, azaz a Bemutatkozás szerkezeti egységet nem sikerül azonosítani, illetve ha az azonosítás sikerült, de az információk kinyerése sikertelen, akkor egy levél kerül kiküldésre az ügyfélhez, amely egy arra vonatkozó kérést tartalmaz az ügyfél felé, hogy pótolja a hiányzó adatokat. Az információkérő levél küldésével egyben az az elvárás is teljesül, miszerint a hivatalnak kötelező válaszlevelet küldeni minden egyes panaszlevélre egy meghatározott időn belül.

Amennyiben a gépnek sikerült azonosítania az ügyfelet, a következő lépés a Cél szerkezeti egység keresése. Ha a rendszer talál cél annotációt, azaz a Cél szerkezeti egységre jellemző kategóriasemleges- vagy kategóriaspecifikus definitív kifejezések alapján képes beazonosítani azt, akkor az ezen a szerkezeti egységen belüli kategóriaspecifikus definitív kifejezések segítségével (ha vannak ilyenek) azonosíthatóvá válik a panaszkategória. Minden lehetséges esetben, azaz ha a Cél szerkezeti egység hiányzik, vagy ha a szerkezeti egységet sikerült ugyan azonosítani, de kategorizálás szempontjából releváns információt nem sikerült belőle kinyerni (azaz a rendszer nem talált kategóriaspecifikus definitív kifejezéseket), vagy harmadik lehetőségként, ha a panaszkategóriát sikerült azonosítani, a keresés a Probléma szerkezeti egységgel folytatódik.

Amennyiben a Probléma szerkezeti egységnek az azonosítása megtörtént, akkor, feltéve, hogy a definitív kifejezések között talál kategóriaspecifikusakat, a gép újra elvégzi a kategorizálási lépést, most már ebben az egységben talált definitív

kifejezések figyelembevételével. Ez a panaszkategória lehet azonos az előzővel, de lehet ettől eltérő is.

Miután sikerült a levélhez panaszkategóriát rendelni, a keresés az Elvárás szerkezeti egységgel folytatódik tovább. Ezt a szerkezeti egységet szintén a kategóriasemleges, valamint a kategóriaspecifikus kifejezések figyelembevételével azonosítja a rendszer, és csakúgy mint az előző esetekben, a kategóriaspecifikus kifejezések alapján újra egy panaszkategóriát rendel a levélhez.

Az eddigiek alapján tehát az algoritmus ezen pontján a következő esetek állhatnak fenn: a Cél, a Probléma, valamint az Elvárás szerkezeti egységek alapján a gép egy, kettő vagy három különböző panaszkategóriát rendelt a levélhez. Az első esetben az algoritmus következő lépése, hogy a levelet a megállapított panaszkategóriában jártas ügyintőzőhöz továbbítja, míg a második és harmadik esetben, hogy a két-, illetve három, az adott panaszkategóriában jártas ügyintézőkhöz kerül a levél továbbításra. Azok a levelek, amelyek több ügyintézőhöz is eljutnak, tartalmazzák azt az információt, hogy kik a címzettek, hiszen ez az ügyintézők számára releváns lehet.

Abban az esetben, ha az algoritmus a Probléma szerkezeti egységre utaló kategóriasemleges és kategóriaspecifikus kifejezéseket nem talál a levélben, úgy megvizsgálja, hogy annak alternatívájaként Javaslat szerkezeti egységet talál-e. Amennyiben igen, úgy a kategóriaspecifikus kifejezések alapján kikalkulált panaszkategória megállapítása után a folyamat az Elvárás szerkezeti egység keresésével folytatódik. Amennyiben nem, úgy információkérő levél kerül kiküldésre az ügyfélnek, amelyben kérik, hogy tisztázza, hogy pontosan milyen ügyben fordult a minisztériumhoz.

Abban az esetben, ha a rendszer nem talál elvárás annotációt a levélben, vagy nem sikerül abból releváns információt kinyernie, a levél a valamelyik korábbi szerkezeti egység alapján megállapított panaszkategóriában jártas ügyintézőhöz kerül továbbításra.

Az ábra és a fentiek alapján is látható, hogy amennyiben egy levélből kinyerhető információ arra vonatkozóan, hogy az állampolgár milyen közigazgatási kategóriának megfelelő panasszal fordult az ügyfélszolgálathoz, úgy azt a gép az adott témában szakértő ügyintézőhöz juttatja el, aki válaszol arra a megszabott határidőn belül, ellenkező esetben pedig a rendszer automatikusan is generálhat egy információkérő levelet. Hogy milyen információ hiányzik a levélből, az megállapítható annak alapján, hogy az algoritmus milyen lépéseket járt be, mielőtt az információkérő levél ponthoz ért volna.

A kategorizálás forgatókönyv tesztelése eddig a kategóriasemleges definitív kifejezések figyelembevételével történt, azaz azt teszteltük, hogy az algoritmus, illetve a kategóriasemleges kifejezések listája alapján a gép milyen mértékben képes beazonosítani a kategorizáláshoz szükséges szerkezeti egységeket (Bemutatkozás, Cél, Probléma, Javaslat, Elvárás), és jut el az Ügyintéző pontig (lásd 5.1. ábra). Az eredmény, hogy a 888 panaszlevélből 156 levél esetében a levél az ügyintézőig jut, a többi esetben azonban valamelyik szerkezeti egység annotációjának hiányában az algoritmus információkérő levelet küld ki. Feltételezésünk szerint a számos információkérő levél küldésének egyik fő oka a szerkezeti egysé-

gekről szóló fejezetben is említett definitív kifejezések listájának a hiányossága. Elvásáraink szerint a lista bővítésével, pontosításával az eredemények jelentősen javulni fognak.

5.2. A profil

A profil kialakításához az összes szerkezeti egységet figyelembe kell venni, azaz a Megszólítást, az Elismerést, az Előzményt, az Egyéb körülményeket, a Vádaskodást, a Köszönetet és a Lezárást, valamint a problémakategória megállapításához figyelembe vett szerkezeti egységeket. Ezeknek a jelenléte, illetve a hiánya önmagában is árulkodó lehet, valamint egymáshoz viszonyított sorrendjük, arányaik is hordozhatnak fontos információkat. Ugyanakkor az ezekben az egységekben előforduló kifejezések stilisztikai jellemzői is értékesek lehetnek. Természetesen nem állítjuk, hogy egy pontos szociológiai, illetve pszichológiai profilt lehet ezek alapján az információk alapján felállítani az illető ügyfélről, azonban elvárásaink szerint bizonyos következtetések levonhatóak.

Az ügyfél aktuális (a levél írásának pillanatában fennálló) érzelmi állapotára következtethetünk a szótárunkban durva vagy bizalmas stílusúként jelölt kifejezések használatából. A levelek ilyen célú vizsgálata után azt mondhatjuk, hogy ha egy levél legalább egy durva vagy bizalmas stílusjegyű kifejezést tartalmaz (bármely szerkezeti egységben), akkor az ügyfél aktuális érzelmi állapota zaklatott. Amennyiben a levélben előforduló kifejezések legalább 0,5%-a, de legfeljebb 1%-a durva és/vagy bizalmas kifejezéseket tartalmaz, akkor az ügyfél erősen zaklatott érzelmi állapotban írta a levelet, ha pedig ez az érték 1% fölötti, akkor a levélíró aktuális érzelmi állapota szélsőségesen zaklatottnak tekinthető.

Az ügyfél szociológiai és pszichológiai profiljának felállítása a dialógusok kialakítása során lesz majd nagyon fontos, hiszen ha a használt kifejezésekből, a levél szerkezetéből, illetve tartalmi jellemzőkből tudunk következtetni az ügyfél életkörülményeire, iskolázottságára, vagy épp az aktuális érzelmi állapotára, az megszabhatja a kérdések és a válaszok formáját, illetve tartalmát egyaránt.

6. Összefoglalás

Jelen cikkben egy összetett cél elérése érdekében folytatott kutatás első eredményeit tárgyaltuk. Ezek közül az első egy egységes, robusztus előfeldolgozó keretrendszer és az ehhez kapcsolódó formátum elkészítése volt. Az eszköz és a formátum lehetővé teszi, hogy különféle, már korábban rendelkezésre álló nyelvfeldolgozó eszközöket, illetve a saját fejlesztéseinket egységesen kezeljük és a későbbiekben újabb komponensekkel egészítsük ki. Meglátásunk szerint az igazi értéke a rendszernek a szolgáltatásként igénybe vehető interfészekben, formátumokban rejlik. Hosszú távon tervezzük az implementáció UIMA [3] alapokra való helyezését.

A nyelvhasználattal, fogalmazással kapcsolatos problémák leküzdésében részeredményeket értünk el a szerkezeti egységek beazonosítása és a szűrésben való felhasználásuk által. A szerkezeti egységek jelölése — amennyiben létrejön —

kellően pontos. Azonban javítanunk kell még a felismerés hatékonyságán, amelyet a definitív kifejezések listájának bővítésével remélünk elérni.

A szerkezeti egységek felismerésének másik hozadéka az, hogy lehetővé teszik a levelek bizonyos részeinek elhagyását, és ezáltal megkönnyítik a kategorizálást. Ez a megközelítés túlmutat az egyszerű, egész korpuszra jellemző stopszólisták használatán, mivel ez a szűrés minden levélre külön-külön elvégezhető. Feltehetőleg a kategorizált levelek kis száma miatt a szűrés csak minimális javulást hozott a pontosság terén. Másrészt a különféle szűrési eljárásaink jelentős futási idő megtakarítást eredményeztek.

Fontos iránynak tartjuk a profilok építését a használt kifejezések alapján, valamint ezeknek a dialógusokban történő alkalmazását. Ezen a korpuszon alapvetően a profil szociológiai és pszichológiai dimenziójának felépítését tervezzük. A profilt nem csak egyszerű adatgyűjtési igények kielégítése miatt építjük. Fontos szerepet szánunk neki a levélíróval folytatott dialógus paraméterezésében: a kérdések és válaszok nyelvezetének meghatározásában, és az ügyfél várható reakcióinak megbecslésében. Ezen reakcióktól függővé tehetjük azt is, hogy felteszünk-e egyáltalán egy adott kérdést. A profilok segítségével szélsőségesen zaklatott felhasználók gyakran zavaros leveleit is felismerhetjük és megfelelően kezelhetjük.

Az itt bemutatott eredmények szándékaink szerint csupán az alapját képezik egy hoszabb kutatómunkának, amely során a rendszerünk kísérleti alkalmazását szeretnénk elérni. A munka része lesz más, eltérő tulajdonságokkal rendelkező korpuszok kipróbálása, valamint egy géppel támogatott megértést szemantikus keretek és hálók segítségével megvalósító eszköz elkészítése is.

Hivatkozások

1. Steven Busuttil. Support vector machines, 2003.
2. Kabai Dóra. Automatikus tartalmi kódolás és osztályozás kidolgozása az igazságügyi minisztérium ügyfélszolgálatára beérkező állampolgári levelekre, 2006.
3. David Ferrucci and Adam Lally. Uima: an architectural approach to unstructured information processing in the corporate research environment. *Nat. Lang. Eng.*, 10(3-4):327–348, 2004.
4. Hitec. *categorizer.tmit.bme.hu/trac/wiki*.
5. G. Holmes, A. Donkin, and I.H. Witten. Weka: A machine learning workbench. In *Proc Second Australia and New Zealand Conference on Intelligent Information Systems*, Brisbane, Australia, 1994.
6. Héder Mihály. Szemantikusan annotált dokumentumok létrehozása szövegfeldolgozó eszközök támogatásával, 2009.
7. Dau Pelleg and Andrew Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *In Proceedings of the 17th International Conf. on Machine Learning*, pages 727–734. Morgan Kaufmann, 2000.
8. Hans Rott. *Words in Context: Fregean Elucidations*, volume 23, pages 621–641. 2000.
9. Duane Szafron, Russell Greiner, Paul Lu, David Wishart, Cam Macdonell, John Anvik, Brett Poulin, and Zhiyong Lu. Explaining naive bayes classifications. Technical report, 2003.
10. Bárczi Géza és Országh László. *A Magyar Nyelv Értelmező Szótára (CD)*. Arcanum Adatbázis Kft, 1994.

A. További példák annotált levelekre

s 1. levél .
s Hozzá nem értésük már nagyon dühítő .
s Mit képzelnek maguk tulajdonképpen Mi a maguk munkája .
s Törvény szerint legkésőbb 30 napon belül válaszolni kell .
s Megmagyaráznák .
s Majd segítek .
s Amit maguk már művelnek egyenlő a Szervezett . bűnözéssel . .
s Maguk nagyon sok kárt okoznak .
s Ez már egyértelmű Bűnpártolás .
s Kértem két fontos címet 3 napon belül itt legyen vagy kénytelen leszek magukat is feljelenteni szándékos károkozásért .
s Nem lopom a pénzemet .
s Az az egyetemeken szokás .
s . Bűnöző . disznók miatt én nem fogom veszni hagyni a pénzemet ami jogosan jár .
s Mit képzeltek ti .
s Köztörvényes gazemberek Az erdőkből kitakarodni .
s Maguk csak . . hülyék . .

s 95. levél .
s Igazságügyi Minisztérium Budapest V. ker. Kossuth tér 2/4 Tisztelt Name Miniszter Úr Tisztelettel Kérem a Miniszter Úr Segítségét hogy a boszorkány ügyembe segítsen panasszal élek a Name ellen mert felbérelt kettő boszorkányokat ellenem lakása Place utca 22 szám nagyon rossz így élni ettől az orr viszketéstől nagyobb fájdalmat és szenvedést nem lehet okozni mert ez a legidegesítőbb az egész világon azért csinálják a boszorkányok élvezik azt hogy szenvedést okoznak ez a boszorkány ügyem nehéz ügy mert nem tudom bizonyítani és nem lehet mert nem ismerem azt a kettő boszorkányokat akik üldöznek csak a Name ismeri őket .
s Tisztelettel kérem azt a hivatali személyt aki az ügyemmel foglalkozik hogy próbálja valahogyan . szóra . bírni .
s A Name hogy mondja meg a kettő boszorkányoknak nevüket és . lakcímüket . még akkor is ha letagadja és hazudik és nem ismeri el azt hogy boszorkányokkal üldöztet engem .
s Tisztelt Minisztérium Tudatom önnel Miniszter úr és önökkel azt is hogy leveleket írtam a Kék Fény Szerkesztőségének a Magyar Rádió Szerkesztőségének a TV RTL Klub Fókusz Szerkesztőségének és annyi Tiszteletet nem érdemlek meg hogy a panasz leveleimre válaszoljanak nyilvánosság elé is akartam vinni az ügyemet de nem sikerült mert a három Szerkesztőségek nem veszik komolyan az ügyemet tudom hogy tele vannak panaszos ügyekkel de azért válaszolhattak volna csak egyedül az Ombudsman válaszolt a levelemre és . nyilvántartásba . vették