4. Cunningham H., Maynard D., Bontcheva K., Tablan V.: GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. in proc. of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL '02), Philadelphia, 2002.
5. Dill S., Eiron N. et al.: A Case for Automated Large-Scale Semantic Annotation. Journal of Web Semantics, 2003.
6. Fabian M. Suchanek, Mauro Sozio, Gerhard Weikum: SOFIE: A Self-Organizing Framework for Information Extraction 18th International World Wide Web conference (WWW 2009), Madrid
7. Laclavík Michal, Seleng Martin, Ciglan Marek, Hluchy Ladislav: ONTEA: Platform for pattern based automated semantic annotation Computing and Informatics, Vol. 28, 2009, 555–579, V 2009-Sep-16
8. Trón Viktor, László Németh, Péter Halácsy, András Kornai, György Gyepesi, and Dániel Varga: Hunmorph: open source word analysis in proc. of ACL., 2005
9. Silviu Cucerzan: Large-Scale Named Entity Disambiguation Based on Wikipedia Data in proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 708–716, Prague, June 2007.
10. Zeno Gantner,Lars Schmidt-Thieme: Automatic Content-based Categorization of Wikipedia Articles in proc. of the 2009 Workshop on the People's Web Meets NLP, ACL-IJCNLP 2009, pages 32–37, Suntec, Singapore, 7 August 2009.

Kontextualizált névelem-felismerés és relációkinyerés kórházi zárójelentésekben

Solt Illés[1,2], Szidarovszky P. Ferenc[1], Tikk Domonkos[1,2]

[1] Budapesti Műszaki és Gazdaságtud. Egyetem, Távközlési és Médiainf. Tanszék
H-1117 Budapest, Magyar Tud. krt. 2, e-mail: {solt,szidarovszky,tikk}@tmit.bme.hu
[2] Humboldt-Universität zu Berlin, Knowledge Management in Bioinformatics
D-10099 Berlin, Unter den Linden 6, e-mail: {solt,tikk}@informatik.hu-berlin.de

Kivonat Cikkünkben a kórházi zárójelentések szövegbányászati feldolgozásával foglalkozó i2b2 szervezet 2010-es, információkinyeréssel kapcsolatos feladatára (Fourth i2b2/VA Shared-Task) készített megoldásunkat ismertetjük. Az első, névelem-felismerési feladatban három entitástípus szövegbeli előfordulásait, pontosabban egy szűk bennfoglaló nyelvtani egységet kellett megjelölni. A második, állításosztályozási feladatban ezen entitások említésének jellegét (kijelentő, tagadó, spekulatív stb.) kellett osztályozni. Végül a harmadik, relációkinyerési feladatban az egy mondatban szereplő entitások között fennálló kapcsolat meglétét és pozitív esetben a típusát kellett megállapítani. Megoldásainkban kontextusra épülő, a rendelkezésünkre bocsátott tanítóadaton betanított – részben szabályalapú, részben felügyelt gépi tanuláson alapuló – módszereket alkalmaztunk. Munkánkban elemezzük az egyes eljárások hatékonyságát és megvizsgálunk néhány lehetséges továbbfejlesztési irányt.

1. Bevezetés

Az orvosbiológia a szövegbányászat egyik vezető alkalmazási területe, mivel számos feladattípus esetén sikerült hatékony eljárásokat fejleszteni, amelyek képesek például kutatóbiológusokat, klinikai és kutatóorvosokat, betegbiztosítási szakértőket a mindennapi munkájukban támogatni (l. pl. a [3] áttekintő tanulmányt). A már széles körben elterjedt információ-visszakeresési megoldások mellett manapság már egyre jellemzőbb az alkalmazási területtől jobban függő, és gyakran lényegesen bonyolultabb információkinyerő módszerek gyakorlati alkalmazása, vagy legalábbis ezek kísérleti bevezetése [7].

Három alapvető információkinyerési feladattípus a névelem-felismerés (named entity recognition, NER) [11], állításosztályozás (assertion classification) [8,21] és a relációkinyerés (relation extraction, RE) [5,8,9]. Névelem-felismerésnél az adott feladat szempontjából releváns névelem-, vagy más szóval entitástípusok előfordulásait kell egy szövegben azonosítani; orvosbiológiai alkalmazásokban ezek többnyire fehérjék, gének, mutációk, betegségek, gyógyszerek, szimptómák, tünetek stb. nevei. Állításosztályozásnál a feladat az entitás- és/vagy reláció-előfordulások szemantikai értékének meghatározása, pl. az állítás, tagadás és

spekuláció megkülönböztetése. Relációkinyerésnél az első lépésben azonosított vagy esetleg már eleve adott, entitás-előfordulások közötti kapcsolatok meglétét és típusát kell meghatározni; jellemző feladatok közé tartoznak a fehérje–fehérje kölcsönhatások (protein-protein interaction, PPI), betegség–kezelés, ill. gén–betegség összefüggések kinyerése.

A kórházi zárójelentések számos értékes információt tartalmaznak, amelyek segítségére lehetnek az orvoskutatóknak gyógykezeléseknek a páciensekre gyakorolt hatásának tanulmányozásában, a betegségek és gyógyszerezésük hatásvizsgálatában, az elvégzett vizsgálatok és betegségek felderítési arányának feltérképezésében stb. Ahhoz azonban, hogy a szöveges zárójelentésekből — melyek gyakran nem felelnek meg a nyelvtan és a helyesírás szabályainak, valamint számos szaknyelvi rövidítést is tartalmaznak — további kutatási célra felhasználható, statisztikailag szignifikáns mennyiségű céladathoz jussunk, ki kell nyerni a szövegből a releváns információkat és azok kapcsolatát. Az i2b2 (Informatics for Integrating Biology & the Bedside)[3] szakcsoport 2010-ben immár negyedik alkalommal[4] rendezett nemzetközi megmérettetést; minden évben más-más klinikai szövegekre vonatkozó aktuális szövegbányászati problémát helyezve a verseny fókuszába. A versenyen hagyományosan jól szerepelnek a magyar csapatok: 2006-ban 2. helyezést értek el Szarvas György és kollégái [18], 2008-ban pedig a mi csapatunk végzett az első helyen [17]. Jelen munkánkban kutatócsoportunknak a 2010-es kiírás feladataira adott megoldásait és az utólagos elemzések tanulságait ismertetjük. Fontosnak tartjuk megjegyezni, hogy bár a versenyen angol nyelvű szövegeken kellett dolgozni, a megoldásaink nagy része átvihető magyar nyelvű szövegfeldolgozásra is a megfelelő nyelvi eszközök magyar verziójának behelyettesítésével, így az általunk javasolt megoldások a hazai szakemberek részéről is érdeklődésre tarthatnak számot.

2. Feladatok

2.1. Névelem-felismerés

A 2010. évi verseny három egymásra épülő feladatból állt [1]. Az első feladat alapvetően névelem-felismerés volt, ahol három entitástípus felismerése volt a cél:

1. általános egészségügyi probléma (*medical problem*): ide tartoznak betegségek, tünetek, szimptómák, sérülések, abnormalitások stb.;
2. gyógykezelés (*treatment*): ide tartoznak a gyógyszerek, biológiai anyagok, gyógyszeradagolók, gyógyászati segédeszközök stb.;
3. vizsgálat (*test*): vizsgálati kezelések, testnedveken végzett laborvizsgálatok, életjelfunkciók mérési eredményei.

A feladatban a szokásos, kizárólag az entitásra koncentráló névelem-annotációs feladatokon túlmutatóan az entitás előfordulásának értelmezését támogatandó

[3] `https://www.i2b2.org/`

[4] `https://www.i2b2.org/NLP/Relations/`

az entitást fejként tartalmazó főnévi vagy melléknévi csoportokat kellett annotálni. Emellett az ún. prepozíciós szabály szerint bővíteni kellett az annotációkat a főnevet követő első prepozíciós szerkezettel, ha az nem tartalmazott amúgy is annotálandó entitást. Így a „pain in chest" szerkezetet egyben kellett felismerni, a „removal of mass"-t viszont két entitásként (l. még [1]). Az annotálás módja az orvosbiológiai korpuszokon gyakran alkalmazott tokenszintű annotáció. Szemben a karakterszintű annotációval, ez a jelölés emberi annotációnál kevesebb időráfordítással állítható elő, ellenben kevésbé pontos jelölést tesz lehetővé, mint amilyet a klinikai szövegek szintaktikai gyakorlata indokolna (pl. egybeírások, torlódó szavak, szokatlan rövidítések).

2.2. Állításosztályozás

A második feladatban az egészségügyi probléma entitástípus előfordulásait kellett a következő 6 szemantikai osztály valamelyikébe sorolni: megfigyelhető (állítás), nem figyelhető meg (tagadás), lehetséges, feltételezett, feltételhez kötött, vagy mással (nem a páciensszel) kapcsolatos.

2.3. Relációkinyerés

A harmadik feladatban azt kellett meghatározni, hogy milyen kapcsolat van – ha van egyáltalán – az egy mondatban szereplő entitások között. Itt 8 reláció szerint kellett entitáspárokat vizsgálni. A relációk egyet kivéve (*problem indicates problem*, PIP) szimmetrikusak voltak, az irányított relációnál természetesen a megfelelő irányt is meg kellett határozni. A reláció egyértelműen meghatározta, hogy milyen entitástípusok között állhat fent.

2.4. Kiértékelés, lebonyolítás

Minden feladatra három megoldást lehetett csapatonként beküldeni. Az első feladat esetében elsődlegesen a pontos illeszkedés mikro F-mérték, pontosság, felidézés hármast alapján rangsorolták a versenyzőket. Ugyanerre a három mértékre átfedő illeszkedés szerint is kiértékelték a megoldásokat, vagyis ekkor már helyesnek számított az a predikció, ahol tokenszinten van átfedés a helyes és a predikált entitások között. A két osztályozási feladatnál az összes osztályra vetített mikro F-mérték, pontosság, felidézés szerint értékelték a beküldéseket.

A verseny lebonyolítása az alábbi módon igazodott a feladatok sorrendjéhez. Egy-egy feladat leadási határideje között 24 óra állt a versenyzők rendelkezésére. A névelem-felismerési feladat teljesítése után közzétették a résztvevők számára a tesztadatokhoz tartozó helyes értékeket (ground truth), amit a második és harmadik feladat megoldásánál így fel lehetett használni; hasonlóan nyilvánosságra hozták az állításosztályozás megoldókulcsát is a feladat teljesítése után. Ez az egymásra épülő kiértékelési módszer lehetővé tette az egyes részfeladatokra beadott megoldások egyenkénti értékelését, hiszen így nem adódtak össze a hibák. Gyakorlati alkalmazásokban ugyanakkor az állításosztályozásnál és a relációkinyerésnél természetesen magasabb hibaértékkel szükséges kalkulálni, amikor a

három részfeladatot egymásra épülve hajtják végre. A hibanövekedés nagyságrendjét például a [7] cikk alapján becsülhetjük meg, ahol a névelem-felismerés és a PPI-kinyerés kombinált folyamatának hatékonyságát vizsgálták.

3. Módszerek

3.1. Lexikon és szintaktikai minták alapján működő névelem-felismerő

Az első feladat egy többcímkés szekvenciaannotálási feladatként fogható fel, ahol a címkék a három entitástípushoz tartoznak. A feladatkiírás logikáját követve elsőként a fejként szereplő entitásokat határoztuk meg (illesztés), majd kiterjesztettük az entitás szövegkörnyezetét az előírt feltételek szerint (kiterjesztés).

Az illesztési feladatban a fejentitások felismeréséhez a tanító halmazból indultunk ki. Egy entitást akkor tekintettünk jelöltnek az adott osztályhoz, ha az átfedő pontosság – az előfordulások, amelyek átfednek az adott osztály annotációival osztva az összes előfordulással – meghaladott egy bizonyos küszöbértéket. F-mértékre optimalizálva 0,6 bizonyult az optimális küszöbértéknek a tanító halmazon. A lexikont hasonló kiértékelési módszerrel bővítettük az i2b2 Obesity Challenge [20] megoldásánál általunk összeállított fogalomlexikon [17], valamint internetes forrásokból származó adatok alapján. A kiterjesztés alapjául tehát az így összeállított névelemlistának a tesztszövegre illeszkedő tokenjei szolgáltak.

A szövegkörnyezet kiterjesztéséhez több angol nyelvi elemző kimenetét elemeztük abból a szempontból, hogy ezek hogyan illeszkednek a meglehetősen körmönfont annotációs irányelvekhez. Összehasonlításunkban a Stanford szintaktikai elemzőt[5], a Charniak–Lease elemzőnek[6] a McClosky-féle biológiai szövegeken tanított modellel [12] való verzióját, az Enju elemzőt[7], valamint a GENIA chunkert[8] vizsgáltuk. Azt tapasztaltuk, hogy a jól formált mondatokon a Stanford szintaktikai kimenete egyezett meg leginkább a versenyen elvárttal, míg a nyelvtanilag helytelen (pl. felsorolás jellegű) mondatoknál a GENIA chunker bizonyult a legjobbnak.

Bizonyos esetekben azonban az elemző kimenete módosítást igényelt, hogy megfeleljen az annotációs irányelveknek. Az egyik tipikus példa erre, hogy az elemzők a főnévi csoportokat jellemzően nem vágják kötőszavak mentén, azaz a „The patient experienced X and Y" mondatban az X entitást tartalmazó főnévi csoportot az elemzők „X and Y"-ként azonosítják, míg a versenyben a legszűkebb bennfoglaló nyelvtani egységet X-szel kell annotálni. Ilyenkor tehát kettévágjuk a kötőszavak mentén a főnévi és melléknévi csoportokat. Szintén eltért az elemzők kimenete az annotációs irányelvektől a bizonyosságot jelentő határozószók esetén („likely" stb.). Bár az elemzők – nyelvtanilag helyesen – ezeket a főnévi

[5] `http://nlp.stanford.edu/software/lex-parser.shtml`
[6] `ftp://ftp.cs.brown.edu/pub/nlparser/reranking-parserAug06.tar.gz`
[7] `http://www-tsujii.is.s.u-tokyo.ac.jp/enju/`
[8] `http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/postagger/geniatagger-3.0.1.tar.gz`

csoportba vonták be, az állításosztályozásra való tekintettel (l. ott) ez ellenkezett az irányelvekkel.

Az annotációs irányelveket szintaktikai gráfokra vonatkozó illeszkedési szabályokra írtuk át, és ezekkel határoztuk meg az entitások szövegkörnyezetét. A szabályok az illeszkedő fejentitást kibővítették a tartalmazó főnévi, ill. melléknévi csoportra, valamint további bővítést végeztek a prepozíciós szabály szerint. A kiterjesztésnél a Stanford elemző és a GENIA chunker kimenetének unióját vettük, ugyanis ezek akkor is átlapolódhatnak, ha a fejentitásokra diszjunktak. Például az *egészségügyi probléma* entitástípus tartalmazza mind a „reflux" és „disease" szavakat, de a „reflux disease" kifejezést nem. Ekkor a „The patient has reflux disease." mondatban mindkét szót külön entitásként felismerjük, amelyeknek a kiterjesztése egybeesik. Ilyen esetekben — ha az entitások azonos osztályba tartoztak — az egyesített kiterjesztést vettük.

3.2. Állításosztályozás szabályalapú módszerrel

A második feladatot szabályalapú megközelítéssel oldottuk meg, amely az entitás szövegkörnyezete alapján határozta meg az állításosztályt. Az entitást tartalmazó mondatot az előfordulást megelőző, tartalmazó, követő, illetve körülvevő részre bontottuk. A tanító adat alapján azonosítottuk az egyes állításosztályokra utaló kulcsszavakat és hogy melyik szövegrészletre vonatkozik a hatásköre, l. például az 1. táblázatot.

1. táblázat. Példák az állításosztályokra utaló kulcsszavakra

Osztály	**Kulcsszó**		
	megelőző	**követő**	**körülvevő**
Megfigyelhető (állítás)	due to		
Nem figyelhető meg (tagadás)	without		non
Lehetséges	possible		
Feltételezett	if you		any
Feltételhez kötött		when	exertion
Mással kapcsolatos		wife	

Korábbi kontextusalapú állításosztályozási eljárásunkat [17] használtuk fel frázis-, illetve mondatszintű negáció, allergia és családdal kapcsolatos szövegrészletek, illetve hatáskörök bejelölésére, és az adott hatáskörbe eső entitásokat rendre a *nem figyelhető meg*, *feltételhez kötött* és *mással kapcsolatos* osztályba soroltuk. A különböző módszerek által adott eredmények közötti esetleges ellentmondás esetén a nagyobb *a priori* valószínűségű állításosztályt választottuk. Azokat az entitásokat, amelyeket a fenti módszerek egyike sem sorolt be valamelyik osztályba sem, a *megfigyelhető* osztályhoz rendeltük.

3.3. Relációkinyerés felügyelt tanulókkal

A verseny relációkinyeréssel kapcsolatos feladata többcímkés osztályzásnak tekinthető, ahol minden adott feltételnek megfelelő — azaz a megfelelő entitásosztályokba tartozó — entitáspárt valamely relációtípushoz kell hozzárendelni. A feladatot relációtípusonként bináris osztályozók alkalmazásával oldottuk meg, azaz összesen 9 modellt építettünk: egy-egy osztályozót a szimmetrikus relációtípusokra, és irányonként egy osztályozót a PIP-típusra. Végül feloldottuk az esetleges ellentmondásokat, vagyis amikor több osztályozó is a pozitív osztályba sorolt egy entitáspárt.

A relációtípus meghatározza, hogy milyen típusú entitások között állhatnak fenn: 5 relációtípust definiáltak egészségügyi problémák és kezelések között, kettőt egészségügyi problémák és vizsgálatok között, míg az aszimmetrikus PIP reláció – amelyet szétbontottunk PIP ($\curvearrowright$) és PIP ($\curvearrowleft$) irányfüggő altípusokra – két egészségügyi probléma közt állhat fent. Vegyük észre, hogy minden relációtípus legalább egy egészségügyi probléma entitást tartalmaz (bővebb statisztikákkal a 2. táblázat szolgál).

Két különböző megközelítést alkalmaztunk a feladat megoldására. Alapmódszerként egy együttes előfordulás (kollokáció) alapú módszert alkalmaztunk, amelyik az entitások között gyakorta előforduló szavakat és szó n-gramokat azonosítja relációtípusonként, és amely az így felismert minták alapján osztályozza a tesztpéldányokat.

A második módszer gépi tanulási problémaként közelíti meg a feladatot, és szupport vektor gépeket (SVM) alkalmaz különböző magfüggvényekkel (kernel) az osztályozó modellek megtanulásához, természetesen mind a 9 relációtípusra külön modellt építve. A gépi tanulókat 10-szeres keresztvalidálással értékeltük ki, amihez 10 nagyjából azonos méretű részre osztottuk a tanítóhalmazt. Összehasonlításként az n-gram alapú módszert is kiértékeltük ily módon.

Szó n-gram alapú relációkinyerés. A (szó) n-gram alapú módszer a tanító halmazon megfigyelt szósorozatok előfordulási statisztikái alapján működik. Minden osztályra külön modellt készítettünk.

Elsőként tokenizáltuk a mondatokat, és a tokeneket az alábbi 4 tokenosztályba soroltuk:

1. entitás (a tanítóhalmazon definiálva);
2. számok (csak számjegyeket tartalmazó tokenek);
3. egyéb szavak;
4. írásjelek.

Az entitások előfordulását az entitástípus címkéjével (entity blinding), míg a számtokeneket egységesen egyazon címkével helyettesítettük, majd készítettünk egy n-gram szótárat a tanítóhalmazon. Különböző beállításokkal futtattuk a kísérleteket az n-gramok minimális és maximális hosszát, valamint az összesített minimális előfordulás mennyiségét ($\min_{\text{freq}}$)változtatva.

Egy n-gramnak valamely relációtípus szerinti osztályozási pontosságát a (típus szerinti) pozitív és az összes előfordulás arányának alapján határozzuk meg.

2. táblázat. Relációtípusokra vonatkozó statisztikák és keresztvalidációval mért eredmények a tanító adatokon. A pozitív példák a tanító adatokban szereplő relációk, negatív példák a mondatokban található fennmaradó típushelyes entitáspárok. Mikro F-mérték eredmények, a legjobb eredmény típusonként félkövérrel szedve.

Reláció	**Statisztika**				**Módszer**					
	Poz	**Neg**	**P/N**	**P%**	**kBSPS**	**SpT**	**SL**	**PT**	**APG**	**n-gram**
TERP	1 711	1 858	0,92	48%	0,78	0,74	**0,83**	0,74	**0,83**	0,68
TECP	295	3 274	0,09	8%	0,49	0,44	**0,51**	0,41	0,45	0,35
TRAP	1 413	2 874	0,49	33%	0,64	0,64	0,71	0,64	**0,74**	0,52
TRCP	294	3 993	0,07	7%	**0,45**	0,34	0,42	0,35	0,43	0,29
TRIP	107	4 180	0,03	2%	**0,40**	0,23	0,38	0,24	0,28	0,37
TRNAP	106	4 181	0,03	2%	**0,44**	0,27	0,37	0,23	0,37	0,27
TRWP	56	4 231	0,01	1%	**0,19**	0,13	0,02	0,13	0,11	0,11
PIP ($\curvearrowright$)	900	7 688	0,12	10%	0,49	0,00	**0,55**	0,00	–	0,29
PIP ($\curvearrowleft$)	320	8 268	0,04	4%	0,17	0,00	**0,27**	0,00	–	0,12
Összes	**5 202**	**40 547**	**0,13**	**11%**	**0,60**	**0,47**	**0,65**	**0,47**	**0,52**	**0,54**

Egy entitást is tartalmazó, pozitív példamondatban előforduló n-gramot csak akkor tekintünk pozitív példának, ha az entitás része az annotált relációnak.

Osztályozásnál egy mondatot akkor tekintünk pozitívnak, ha a mondatbeli legmagasabb n-gram pontossági érték elér egy előre definiált küszöbértéket; itt természetesen csak az n-gram szótár elemeit tekintjük. A pontosság kiszámításához a fenti 10-szeres keresztvalidációt alkalmaztuk és minden részhalmaz 10%-án állítottuk be a pontossági küszöbértéket.

Az optimális paraméterértékek meghatározásánál az $n = 1, \ldots, 4$, ill. $\min_{\text{freq}} = 4$ értékek eredményezték a legjobb átlagos keresztvalidált F-mértéket. Mondatszintű osztályozásnál futtattunk kísérleteket a maximumon kívül más aggregáló függvénnyel is, de mindegyik hatékonysága elmaradt a maximumétól.

Kernelfüggvény alapú osztályozás SVM-mel. A szupport vektor gépek adott tanítóhalmaz esetén azt a lineáris hipersíkot határozzák meg, amely a legjobban szeparálja a pozitív és negatív tanító adatokat [6]. Ha a két halmaz lineárisan nem szeparálható, akkor kernelfüggvények segítségével a feladat egy nemlineáris, jellemzően magasabb dimenziójú térbe transzformálható, ahol már fennállhat a szeparálhatóság [16]. A kernelfüggvény egy adott párhoz egy hasonlósági értéket rendel, amely a pár közti belső szorzatként egyszerűen számolható, és lehetővé teszi sokdimenziós problématerek használatát, amelyek például a mondatok strukturális jegyeit jobban leíró, bonyolultabb nyelvtani reprezentációk esetén szükségesek lehetnek.

A kernelfüggvényekkel kapcsolatos kísérleteinkben felhasználtuk azt a kernelösszehasonlító keretrendszerünket, amelyet eredetileg fehérjeinterakciók (PPI)

kinyeréséhez fejlesztettünk [19]. Az abban rendelkezésre álló 13 kernelfüggvényből 5-tel folytattunk kísérleteket:

- a *shallow linguistic* (SL) [4] kizárólag felszíni nyelvtani jegyekkel operál (szófajok, tokenek, lemmák stb.);
- a *partial tree* (PT) [13] és a *spectrum tree* (SpT) [10] kernelfüggvények a mondat szintaktikai fáján dolgoznak;
- a *k-band shortest path spectrum* (kBSPS) [14,19], és az *all-paths graph* (APG) [2] kernelek pedig a mondat függőségigráf-reprezentációja alapján definiálnak hasonlósági függvényt.

A kernelek alkalmazása előtt át kellett alakítani a rendelkezésre bocsátott zárójelentés-dokumentumokat és ezek mondataihoz generált nyelvtani elemzéseket a PPI-relációkinyerésnél *de facto* standardként használt XML formátumra [15], hogy a kerneleket alkalmazni lehessen. A kernelek különböző mondatreprezentációt használnak, ezért az összes reprezentációs formátummal gazdagítani kellett a dokumentumokat. Az SL kernelhez a GENIA taggert alkalmaztuk a lemmák meghatározásához. A PT és SpT kernelekhez szükséges szintaktikai fákhoz a Charniak-Lease elemzőt alkalmaztuk a McClosky-féle biológiai modellel, míg a függőségi gráfokat a Stanford konverterrel állítottuk elő a szintaktikai fákból.

Különböző paraméterbeállításokkal futtattunk keresztvalidációs kísérleteket a kernelekkel, hogy megtaláljuk a legjobb beállításokat, amelyekkel azután az egész korpuszon betanítottuk a modellt, és ezt alkalmaztuk a tesztadatokon. A 2. táblázatban összehasonlítjuk a kernelekkel elért eredményeket és az alapmódszerként használt n-gram alapú módszert a keresztvalidációs adatokon. Minden kernelre csak a legjobb beállítással elért eredményt közöljük.

3. táblázat. A három feladatban elért eredmények a tesztadatokon

Feladat	Módszer	TP	FN	FP	R	P	F
Névelem-felismerés	Illesztés + kiterjesztés	24 892	20 117	16 962	0,55	0,59	0,57
Állításosztályozás	Kulcsszó- és kontextusalapú szabályok	15 805	2 745	2 745	0,85	0,85	0,85
Relációkinyerés	SVM SL kernellel [4]	6 301	2 769	3 122	0,69	0,67	0,68
	kBSPS kernellel [14]	447	8 623	3 772	0,05	0,11	0,07
	Szó n-gram alapon	6 040	3 030	23 697	0,67	0,20	0,31
	SL és kBSPS kernelek kombinációja	4 639	4 431	8 636	0,51	0,35	0,42

4. Eredmények

A 3. táblázat összefoglalja a három feladatban elért eredményeinket, amelyeket már a helyes referenciaadatokkal ellátott tesztadatok segítségével számoltunk ki (a szervezők által rendelkezésre bocsátott kiértékelő ugyanis pontatlan volt). A 4. táblázatban részletesen ismertetjük a névelem-felismerési feladat eredményeit, míg az állításosztályozás részeredményei az 5. táblázatban találhatók. A

névelem-felismerésben 0,57 F-mértéket értünk el pontos, és 0,80-at eredményt átfedő illeszkedés esetén. Állításosztályozásnál 0,92, míg relációkinyerésnél 0,68 volt ez az érték.

4. táblázat. Névelem-felismerési eredmények entitástípusonként a tesztadatokon

Illeszkedés	Entitás	TP	FN	FP	R	P	F
Pontos	Egészségügyi probléma	11,4k	7,2k	5,4k	0,61	0,68	0,64
	Gyógykezelés	7,3k	6,3k	5,6k	0,54	0,57	0,55
	Orvosi vizsgálat	6,3k	6,6k	6,1k	0,49	0,51	0,50
	Összes	**24,9k**	**20,1k**	**17,0k**	**0,55**	**0,59**	**0,57**
Átfedő	Egészségügyi probléma	14,2k	4,3k	2,1k	0,77	0,87	0,82
	Gyógykezelés	10,3k	3,3k	2,4k	0,76	0,81	0,79
	Orvosi vizsgálat	10,0k	2,9k	1,8k	0,77	0,84	0,81
	Összes	**36,3k**	**8,7k**	**6,3k**	**0,77**	**0,85**	**0,80**

5. táblázat. Állításosztályozási eredmények a tesztadatokon

Osztály	TP	FP	FN	R	P	F
Megfigyelhető (állítás)	11 754	1 300	1 271	0,90	0,90	0,90
Nem figyelhető meg (tagadás)	2 934	843	675	0,81	0,78	0,79
Lehetséges	596	392	287	0,67	0,60	0,64
Feltételezett	361	55	356	0,50	0,87	0,64
Feltételhez kötött	26	131	145	0,15	0,17	0,16
Mással kapcsolatos	134	24	11	0,92	0,85	0,88
Összes	**15 805**	**2 745**	**2 745**	**0,85**	**0,85**	**0,85**

5. Diszkusszió

A 4. táblázatban a névelem-felismerésnél látható a pontos és az átfedő illeszkedéssel számolt eredmények közötti jelentős különbség (0,23-os eltérés F-mértékben) arra utal, hogy az entitások hatáskörének helyes kiterjesztése gyakran nehéz feladatnak bizonyult. A pontosság rendre magasabb a felidézésnél a teszthalmazon, ami a F-mértékre vonatkozó lexikonoptimalizálásnak az eredménye.

Az állításosztályozási feladatnál az osztályozási hatékonyság a rendelkezésre álló tanító adatok számával korrelál, l. 5. táblázat. Kivételt csak a *mással kapcsolatos* osztály jelent, amely könnyen felismerhető volt egyes rovatfejlécbeli kulcsszavak gyakori előfordulása miatt (pl. „family history”).

A relációkinyerésnél használt módszerek elemzésénél (2. táblázat) — egybehangzóan a [19] munka megállapításaival — azt találtuk, hogy a szintaktikai elemzési fa alapú kernelfüggvények (PT, SpT) kevésbé képesek a relációtípusra jellemző jegyek kiemelésére, mint a felszíni jegyeken vagy függőségi gráfokon alapuló módszerek (SL, kBSPS és APG), ezért a további kísérletekben nem használtuk fel őket. Az APG kernelt implementációjának lassú sebessége miatt voltunk kénytelenek kizárni (10–50-szeres különbség, l. még 24 órás limit a versenyben). Következésképpen a versenyben az SL a kBSPS kernelek, továbbá a kombinációjuk és az alapmódszer által adott eredményeket vettük számításba. A kBSPS nagyon kevés pozitív osztálycímkét predikált, és így alacsony felidézést, valamint F-mértéket ért el. Az eredmények utólagos kiértékelésénél kiderült, hogy ez a kBSPS kernel paraméterbeállításra való érzékenységének köszönhető: a 9 relációtípusból csak 2-nél adott értékelhető eredményt a keresztvalidációval legjobbnak bizonyult beállítással. Ezzel ellentétben az SL, amely az alapbeállítással is kedvező teszteredményeket produkált, lényegesen robosztusabb, ami valószínűleg azzal is összefügg, hogy ez a megoldás a legkevésbé érzékeny az agrammatikus mondatokból adódó elemzési hibákra, hiszen kizárólag felszíni jegyekkel operál.

6. Összefoglalás

Munkánkban ismertettük a 2010 i2b2/VA Shared-Task nemzetközi verseny három lényegesen különböző részfeladatára elkészített megoldásainkat. Névelemfelismerésnél először lexikonalapú illesztést, majd utána szintaktikai elemzés alapú hatókör-kiterjesztést alkalmaztunk; állításosztályozásnál kulcsszavakon alapuló szabályalapú rendszert fejlesztettünk ki; relációkinyerésnél egy kollokációalapú alapmódszert és több kernelfüggvényt használó SVM-et alkalmaztunk megoldásainkban. Ugyan a versenyben elért helyezéseket a szervezők csak 2010 novemberében, a kapcsolódó konferencián hozzák nyilvánosságra, mindhárom feladattípusnál sikerült a szakterület jelenlegi sztenderdjét legalább elérő eredményeket hozni.

Az utólagos elemzések több olyan irányt mutatnak, amellyel az egyes feladatokra kapott eredmények pontossága esetleg javítható, amelyeket tehát a további munkáinkban vizsgálni kívánunk:

- Névelem-felismerésnél a lexikonalapú illesztés, majd kiterjesztés helyett célszerű lenne az elemzési hibákra érzéketlenebb conditional random fields (CRF) tanuló módszer alkalmazása.
- Relációkinyerésnél számos módon javítható a hatékonyság. A kernelfüggvények esetén a hasonlósági értékek feladatspecifikus módosításával elérhető lenne az agrammatikus mondatok hatékonyabb kezelése. A kernelek optimalizálásánál a keresztvalidációs halmazon való optimalizálás helyett alternatív módon a pozitív/negatív osztálycímkék tanító halmazon mért arányának teszthalmazon való közelítésével kiszűrhetők a nyilvánvalóan téves eredmények (l. kBSPS) — feltéve, hogy ez az arány hasonló a teszthalmazon is. Mivel az adott feladatban nagyon eltért az egyes osztályok egyedszáma, ezért

a több tanító adattal bíró osztályok felülsúlyozásával kedvezőbb mikro F-mérték lett volna elérhető. Végül pedig a kernelek kombinálásával, akár relációtípusonként, akár mondattípusonként is lehet a teljesítményt lényegesen javítani.

Köszönetnyilvánítás

Tikk Domonkost az Alexander von Humboldt Alapítvány, Solt Illést a DAAD támogatta.

Hivatkozások

1. 2010 i2b2/VA Challenge Documentation. `https://www.i2b2.org/NLP/Relations/Documentation.php`.
2. A. Airola, S. Pyysalo, J. Björne, T. Pahikkala, F. Ginter, and T. Salakoski. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC Bioinformatics*, 9(Suppl 11):S2, 2008.
3. A.M. Cohen and W.R. Hersh. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1):57–71, 2005.
4. C. Giuliano, A. Lavelli, and L. Romano. Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proc. of the 11st Conf. of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pages 401–408, Trento, Italy, 2006. The Association for Computer Linguistics.
5. T. Ideker and R. Sharan. Protein networks in disease. *Genome Research*, 18(4):644–52, 2008.
6. T. Joachims. *Making large-scale support vector machine learning practical, Advances in kernel methods: support vector learning*. MIT Press, Cambridge, MA, 1999.
7. R. Kabiljo, A. Clegg, and A. Shepherd. A realistic assessment of methods for extracting gene/protein interactions from free text. *BMC Bioinformatics*, 10(1):233, 2009.
8. J. D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii. Overview of BioNLP'09 shared task on event extraction. In *BioNLP'09: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 1–9. Association for Computational Linguistics, 2009.
9. M. Krallinger, R. A. Erhardt, and A. Valencia. Text-mining approaches in molecular biology and biomedicine. *Drug Discovery Today*, 10(6):439–45, 2005.
10. Tetsuji Kuboyama, Kouichi Hirata, Hisashi Kashima, Kiyoko F. Aoki-Kinoshita, and Hiroshi Yasuda. A spectrum tree kernel. *Information and Media Technologies*, 2(1):292–299, 2007.
11. U. Leser and J. Hakenberg. What makes a gene name? Named entity recognition in the biomedical literature. *Briefings in Bioinformatics*, 6(4):357–69, 2005.
12. D. McClosky. *Any Domain Parsing: Automatic Domain Adaptation for Natural Language Parsing*. PhD thesis, Department of Computer Science, Brown University, 2009.
13. A. Moschitti. Efficient convolution kernels for dependency and constituent syntactic trees. In *Proc. of The 17th European Conf. on Machine Learning*, pages 318–329, Berlin, Germany, 2006.

14. P. Palaga. Extracting relations from biomedical texts using syntactic information. Master's Thesis, Technische Universität Berlin, May 2009.
15. S. Pyysalo, A. Airola, J. Heimonen, J. Björne, F. Ginter, and T. Salakoski. Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, 9 Suppl 3:S6, 2008.
16. B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors. *Advances in kernel methods: support vector learning*. The MIT Press, 1999.
17. I. Solt, D. Tikk, V. Gál, and Zs. T. Kardkovács. Context-aware rule based classifier for semantic classification of diseases in discharge summaries. *J. Am. Med. Inform. Assoc.*, 16(4):580–4, July/August 2009.
18. Gy. Szarvas, R. Farkas, and R. Busa-Fekete. State-of-the-art anonymization of medical records using an iterative machine learning framework. *J Am Med Inform Assoc.*, 14:574–80, Sep-Oct 2007.
19. D. Tikk, P. Thomas, P. Palaga, J. Hakenberg, and U. Leser. A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature. *PLoS Comput Biol*, 6(7):e1000837, July 2010.
20. Ö. Uzuner. Recognizing obesity and comorbidities in sparse data. *Journal of the American Medical Informatics Association*, 16(4):561–70, 2009.
21. Ö. Uzuner, X. Zhang, and T. Sibanda. Two approaches to assertion classification. In *AMIA Annual Symposium Proceedings*, volume 2008, page 752. American Medical Informatics Association, 2008.