Eljárás radiológiai leletek automatikus BNO kódolására

Farkas Richárd[1] és Szarvas György[2]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
rfarkas@inf.u-szeged.hu
[2] MTA-SZTE, Mesterséges Intelligencia Tanszéki Kutatócsoport
szarvas@inf.u-szeged.hu

Kivonat: Cikkünkben egy amerikai kórházak és kutatóintézetek által, 2007 tavaszán rendezett nyílt verseny eredményeiről számolunk be. A verseny célja radiológiai leletek automatikus címkézése volt ICD-9-CM kódokkal (a Betegségek Nemzetközi Osztályozásával /BNO/ megegyező, számlázáshoz használt kódrendszer). A feladat érdekességét más, korábbi szövegfeldolgozási versenyekhez hasonlítva a szöveghez rendelendő kódok nagy száma, illetve a kódrendszer címkéi közti belső összefüggések adták (összesen 45 kód 96-féle különböző kombinációja fordult elő a korpuszban). A leletek automatikus osztályozását lehetővé tevő számítógépes eljárások fejlesztése létfontosságú, hiszen orvosi témájú szöveges dokumentumok kódolására, illetve a feladat során keletkező hibák javítására évi mintegy 25 milliárd dollárt fordítanak, pl. az Egyesült Államokban. A versenyre benyújtott rendszerek tanulsága, hogy a klinikai dokumentumok – emberi pontossághoz közelítő – eredményes feldolgozása nem lehetetlen célkitűzés a napjainkban rendelkezésre álló eszközökkel.

1 Klinikai dokumentumok feldolgozása (bevezetés)

A számítógépes ontológiák, valamint strukturált szótárak fejlesztése terén tapasztalható fejlődés ellenére a legtöbb kórházban az adatok tárolásának részben továbbra is folyó szöveg formájában történik. Ez a gyakorlat sok gigabájtnyi szöveges adatot eredményez, melynek – az adott beteg klinikai kezelésén túl – korlátozott a hasznosíthatósága, az adatok mennyisége és hozzáférhetősége folytán. Nyelvtechnológiai eszközökkel lehetséges e nagy mennyiségű, szöveges adatban a rejtett struktúra felfedezése, és a tárolt információ elérhetővé tétele. Az így kinyert strukturális információ felhasználható keresőalkalmazások, számlázás, minőségbiztosítás céljára, de gyógyszerkutatásokhoz is igen hasznos lehet (adott betegségek ill. tünetek milyen kórtörténetű páciensek, milyen lefolyással stb.). Korábbi munkákban [5] megmutattuk, hogy ez kivitelezhető akár olyan, összesített formában, mely nem sérti az egyes betegek személyiségi jogait.

A klinikai dokumentumok szövegei számos, a hétköznapi nyelvhez képest eltérő tulajdonságot mutatnak [2]. Ilyenek pl. a hiányos (pl. állítmány nélküli), rövid mondatok; a rövidítések gyakori használata; speciális írásjelezés; szokatlan metonímiák.

Ezek a jelenségek többnyire korlátozzák az általános célokra fejlesztett nyelvtechnológiai programcsomagok alkalmazhatóságát orvosi szövegeken.

1.1 Klinikai dokumentumok BNO-kódolása

Az ICD-9-CM (International Classification of Diseases, 9th Revision, Clinical Modification, magyarul Betegségek Nemzetközi Osztályozása, BNO) kódjait a kórházi ellátás során elvégzett vizsgálatok, kezelések dokumentálására használják az Egyesült Államokban. Az egészségbiztosítók kifizetései a klinikai dokumentumokhoz rendelt BNO címkék alapján történnek, melyeket a kezelés után, utólag rendelnek a kórházi dokumentációhoz. Maga a címkézési eljárás, melyet szakképzett munkaerő (pl. orvosok) végez, illetve az ezzel kapcsolatos hibák javításának költségét évente mintegy 25 milliárd dollárra becsülik [4] az USA-ban. Emiatt a címkézési eljárás automatizálása, illetve támogatása nyelvtechnológiai megoldásokkal intenzíven kutatott, piaci szempontból is fontos feladat.

Maga a címkézés hivatalos kódolási útmutatók alapján történik (pl. [3]), melyekben számos, számítógépes nyelvészeti szempontból is érdekes előírást találunk. Ilyenek pl.:

- Bizonytalan diagnózis semmi esetben sem kódolható (bizonytalanság, spekulációk, illetve tagadás azonosítása).
- Tünetek kódjait tilos a dokumentumhoz rendelni, amennyiben egy kapcsolódó betegség kódja már hozzárendelésre kerül (címkék közti összefüggések modellezése).
- Múltbeli betegségek, kezelések, illetve amelyek közvetlenül nem kapcsolódnak a kezelésekhez, nem szabad kódolni (múlt idő felismerése, alany meghatározása). Gyakran felsorolnak kórtörténeti utalásokat, illetve a család és tágabb környezet tagjainak problémáit.

Mivel a címkéket elsősorban könyvelési és dokumentációs célokra használják, a címkézés precizitása anyagi szempontból is fontos az egészségügyi intézetek számára. A dokumentumhoz hozzá nem rendelt kódok bevételkieséssel járnak a kórháznak, míg minden tévesen hozzárendelt kódért - ha a tévedésre fény derül - a számlázott összeg háromszorosával büntetik az intézményt (valamint a csalásnak egyéb jogi következményei is lehetnek).

1.2 A verseny leírása

A 2007 nyarán megrendezett nemzetközi versenyre[1] [4] a szervezők mintegy 2000 radiológiai leletet láttak el a megfelelő ICD-9-CM címkékkel. Az elkészült korpusz (melynek etalon címkézése három, a kódolást egymástól függetlenül elvégző szervezet többségi annotációja lett) felét adták ki a résztvevőknek tanító adathalmazként, míg a fennmaradó dokumentumhalmazon a szervezők végezték el a beérkező eredmények kiértékelését. A kiértékeléshez, illetve a beküldött rendszerek rangsorolására

[1] részletes leírása található a www.computationalmedicine.org/challenge oldalakon

címke szintű $F_{\beta=1}$ értéket használtak (a továbbiakban minden eredményt mi is eszerint közlünk).

2 Eredményeink

Az alábbiakban ismertetjük a versenyre benyújtott rendszerünk főbb jellemzőit, illetve annak fejlesztése és a későbbi kapcsolódó munkáink során nyert tapasztalatokat.

2.1 Nyelvi modell

Tapasztalataink szerint a feladat megoldásában szerepet játszó legfontosabb nyelvi jelenségek a következők voltak: tagadás (egy betegség vagy tünet nemléte nem kódolandó), feltételes/spekulatív nyelvhasználat (a kódolási útmutatók szerint bizonytalan diagnózist semmilyen esetben sem szabad kódolni), illetve múlt idő kezelése (olyan tünetek és betegségek jelennek meg a dokumentum címkézésében, melyeknek hatása van a dokumentum keletkezésekor végzett vizsgálatokra, a fennálló betegségre).

A versenyre benyújtott modellünkben a tagadás és a spekuláció kezelésére készítettünk szótáron és kötött illesztési szabályokon alapuló megoldást. Az időbeliséget a szakértők inkonzisztens címkekódolása miatt nem kezeltük (volt olyan betegség, melynek jelen és múltbeli lefolyásához külön kód tartozott, de a két címke használatában semmi szabályszerűséget sem találtunk a manuálisan annotált korpszban). A spekulatív vagy tagadó szövegelemek hatókörét az írásjelek segítségével állapítottuk meg. Ez a megoldás a tagmondathatárok azonosításának egy durva egyszerűsítése, ami a feldolgozandó szövegek speciális volta miatt itt hatékonynak bizonyult.

A feladat megoldása során döntő tényező volt a lényeges nyelvi jelenségek pontos kezelése (a tagadás szűrése 10.66%, a feltételes mód kezelése 9.6%, a kettő együtt 18.56% pontosságnövekményt eredményezett, míg a végleges rendszer hibáinak egy számottevő részét épp a nem kezelt múltbeli hivatkozások adják).

2.2 Sokosztályos dokumentumosztályozási feladat

A fenti nyelvi előfeldolgozási lépések után minden címke hozzárendelését önálló osztályozási feladatként oldottuk meg (45 db bináris osztályozási feladat). Az így óhatatlanul előálló érvénytelen kombinációkat nem kezeltük. Az egyes feladatokra felírt modellünk a kódolási útmutatók előírásait követő szabályalapú rendszer és a dokumentumok vektortérmodellben ábrázolt példáin tanított statisztikai osztályozók (melyek a szabályalapú rendszer hibás predikcióit igyekeztek modellezni) kombinációjából állt. Ezeket külön-külön részletesen bemutatjuk a következő fejezetekben.

2.3 Szabályalapú BNO-kódoló alkalmazás

A BNO-kódoláshoz használható többé-kevésbé strukturált útmutatók számos helyen elérhetők az Interneten[2]. Egy – ezek valamelyikének felhasználásával készített – automatikus BNO-kódoló, mely az útmutatóban talált kifejezések illesztésével rendel kódokat a leletekhez, minimális emberi beavatkozással előállítható.

1. Táblázat: Hivatalos útmutató konverziója BNO-kódoló szabályrendszerré

KÓDOLÁSI ÚTMUTATÓ	**GENERÁLT DÖNTÉSI SZABÁLYOK**
label *518.0*	**if document contains**
Pulmonary collapse	pulmonary collapse **OR**
Atelectasis	atelectasis **OR**
Collapse of lung	collapse of lung **OR**
Middle lobe syndrome	middle lobe syndrome
Excludes:	**AND document NOT contains**
atelectasis:	
congenital (partial) (770.5)	congenital atelectasis **AND**
primary (770.4)	primary atelectasis **AND**
tuberculous, current disease (011.8)	tuberculous atelectasis
	add label *518.0*

Egy ilyen egyszerű rendszer, melynek kifejlesztése címkézett példák meglétét sem igényli (nincs modelltanítási fázis), meglepően jó pontosságot ér el a leletek osztályozásában (a megfelelő nyelvi előfeldolgozás után). Az NCSH alapján készült szakértői modellünk a verseny tanító adatbázisán 84.07%, a teszt adatokon 83.21% pontosságot mutatott. A módszer legnyilvánvalóbb korlátai, hogy az útmutatókban található kifejezés-lista fedése nem teljes, valamint, hogy azokban nincs utalás a címkeközi (betegség-tünet) összefüggésekre, amiket így a modell nem kezel.

2.4 Címkeközi összefüggések feltárása

A címkeközi összefüggések elsősorban betegségek és tünetek kódjai közt állnak fent. Például a *tüdőgyulladás* (pneumonia, 486-os kód) szövegbeli megjelenése magával hozza a *köhögés* (kód 786.2) és a *láz* (kód 780.6) tünetek előfordulását is a szövegben, azonban ilyen esetekben csak a betegséget szabad címkézni. Az ilyen jellegű összefüggéseket nem tartalmazza a kódolási útmutató, így annak alkalmazása igen gyakran túlkódolja a dokumentumokat tünet-címkékkel.

Ennek orvoslására azt a gépi tanulási feladatot fogalmaztuk meg aminek outputja olyan szabályokat tartalmaz, amelyek bizonyos betegség kódolása esetén bizonyos tüntetet eltávolít a predikált címkehalmazból (hamis pozitív címkézések leválasztása).

---

[2] Unified Medical Language System (UMLS) (http://www.nlm.nih.gov/research/umls/)
National Center for Health Statistics - Classification of Diseases, Functioning and Disability (http://www.cdc.gov/nchs/icd9.htm)
ICD9Data.com: Free 2007 ICD-9-CM Medical Coding Database (http://www.icd9data.com/)

Az így nyert címkeközi relációkkal (5 db. szabály) bővített modell körülbelül 1.5%-os javulást hozott, a tanító adatbázison 85.57%, a teszt halmazon 84.85%-os pontosságot elérve.

Megvizsgáltuk manuálisan is, hogy milyen címkeközi relációkat lehet érdemes felvenni. A manuális és a statisztikai módon nyert címkeközi szabályhalmaz megegyezett, ez a lépés tehát – ezen adatbázis alapján állíthatjuk – gépi tanulási módszerekkel kivitelezhető.

2.5 A nyelvi modellen alapuló adatreprezentáció

A statisztikai modellek alkalmazása nyelvtechnológiai problémákra a vizsgált szövegek gépi feldolgozásra alkalmas ábrázolását teszi szükségessé. Dokumentumosztályozási feladatok megoldására használt leggyakoribb adatreprezentációs modell az úgynevezett vektortérmodell. A vektortérmodell egy sokdimenziós vektortérben ábrázolja a dokumentumokat, a vektortér dimenziói pedig a dokumentumgyűjteményben előforduló egyedi szavak. Hátrány, hogy ebben a reprezentációban a szavak szövegen belüli pozíciójára és sorrendjére vonatkozó információ elvész. Ennek ellenére számos feladatra nagyon jól működő, vektortérmodelles megoldást fejlesztettek ki.

A BNO kódolási feladathoz szükség volt a fentebb ismertetett nyelvi előfeldolgozási lépések elvégzése a dokumentumokon. Azaz a nyelvi modellek által megjelölt tagadott vagy spekulatív szövegrészeket eltávolítottuk a szövegekből, és csak a maradék, tényszerű információkat tartalmazó szövegrészeket használtuk a dokumentumok ábrázolásához.

2.6 Vektortérmodellen alapuló statisztikai BNO-kódoló alkalmazás

A fenti szabályalapú osztályozóhoz hasonló feladatot ellátó statisztikai modell építhető címkézett példák segítségével a (nyelvi előfeldolgozás utáni) dokumentumok vektortérmodellbeli ábrázolását használva, gépi tanulási módszerek segítségével. Ennek a reprezentációnak, és a versenyen közzétett tanító adatbázisnak a használatával a tanító adatokon 88.20%, a tesztadatbázison 86.69% pontosságot mutató modellt kaptunk. Ennek a megközelítésnek a hátránya, hogy azokra a címkékre, amelyekre csak nagyon kevés példa állt rendelkezésre (22 olyan címke volt a 45-ből, aminek a tanító-adatbázisbeli gyakorisága 6 vagy az alatt volt) nem építhető megbízható statisztikai alapú modell, míg előnye, hogy az útmutatóban nem szereplő szinonimákat, rövidítéseket is fel tudja fedezni.

2.7 Statisztikai modell és szakértői szabályrendszer kombinálása

A további kísérletekben azokra a kérdésekre kerestük a választ, hogy a statisztikai alapú (tanító-adatbázison épített) modell és a külső szakértői tudásbázis (tanító adatbázistól független információ) milyen módszerekkel kombinálható és ezzel a hibrid

modellel milyen eredmények érhetőek el. A két modell előnyeinek egyesítésére számos lehetőség kínálkozik:

- A szabályalapú rendszer illesztési szabályainak bővítése, illetve finomítása a címkézett tanító adatok alapján: Ez a folyamat történhet a kiinduló szabályrendszer hibáira (hamis pozitív és negatív címkék) felírt gépi tanulási problémák megoldásával. Ezt a megközelítést döntési fa illetve Maximum Entrópia osztályozók használatával is teszteltük. C4.5 döntési fa osztályozó segítségével 90.22% és 88.92% pontosságot értünk el a tanító ill. tesztadatbázisokon, míg Maximum Entrópia modellel 90.26% és 88.93% pontosságot sikerült elérni.
- A gépi tanulási modell kibővítése a szabályalapú rendszer tudásbázisával: Ennek a megoldásnak a legegyszerűbb formája, a dokumentum vektortérmodellbeli ábrázolását további jellemzőkkel bővítjük, melyek a szabályalapú rendszer címkézését (kimenetét) írják le. Ekkor a gépi tanulási modellnek lehetősége van kiaknázni mind a kódolási útmutatóban rejlő tudást, mind a címkézett adatok vizsgálatával nyerhető mintákat. Az ilyen kombináció egyik nyilvánvaló gyenge pontja, hogy a statisztikai modell csak a szabályalapú rendszer megfelelő számú példával alátámasztott jelöléseit képes integrálni a tanult modellbe. Ez a módszer 90.62% és 87.92% pontosságot mutatott a tanító és tesztadatbázisokon.
- A szabályalapú modell, valamint a gépi tanulási modell együttes használata (kaszkád modell). Ekkor a modellek kombinációja helyett együttesen alkalmazzuk a két osztályozót, azaz a két modell által predikált címkehalmazok uniója lesz a rendszer végső kimenete. Ez a megközelítés 90.53% és 89.33% pontosságot ad.

A szabályalapú rendszer illesztési szabályainak bővítése a címkézett tanító adatok alapján (beleértve a címkeközi összefüggések kezelését is) történhet a leletek tanulmányozásával, manuális módon is. E módszer – noha nagyon jó modellt eredményez – munkaigényessége, illetve skálázhatósága miatt (több ezer BNO kódra kivitelezhetetlenné válik a munkaigénye miatt) gyakorlati szempontból nem ideális megoldás. Egy, ezt a megközelítést követő modell 90.02% valamint 89.41% pontosságot mutatott a tanító illetve tesztadatokon. Ezek az eredmények tekinthetőek a fenti három hibrid modell által elérhető elméleti felső korlátnak.

2. Táblázat: A különböző modellek, és pontosságuk

	tanító adatbázis	teszt adatbázis
45-osztályos statisztikai modell	88.20%	86.69%
Szabályalapú BNO-kódoló	84.07%	83.21%
Szabályalapú r. címkeközi összefüggésekkel	85.57%	84.85%
Hibrid1 (kiterjesztett szabályalapú)	90.26%	88.93%
Hibrid2 (kiterjesztett statisztikai)	90.62%	87.92%
Hibrid3 (szabályalapú + statisztikai kaszkád)	90.53%	89.33%
Kézileg fejlesztett szabályalapú modell	90.02%	89.41%

3 Értékelés

Az alábbiakban értékeljük az általunk adott BNO-kódolási modell hatékonyságát az emberi címkézéshez, illetve a versenyre benyújtott modellekhez viszonyítva.

3.1 Egyetértés az annotálást végző szervezetek és a modelljeink között

A fentebb ismertetett eredmények megközelítik azt a pontosságot, amit képzett szakemberek képesek elérni a címkézési eljárás végrehajtásában. Az etalon címkézés 3 független (BNO-kódolással a versenytől függetlenül is foglalkozó) egészségügyi szervezet jelöléseinek többségi szavazásával állt elő, azaz minden olyan címke szerepelt az etalon címkézésben, amit legalább két szervezet javasolt.

3. Táblázat: Az annotátorok egyetértési rátái egymással, az etalon címkézéssel valamint két modellünkkel.

	A1	A2	A3	GS	Szabály	Hibrid
Annotátor1	—	73.97/75.79	65.61/67.28	83.67/84.62	75.11/75.56	78.39/79.42
Annotátor2	73.97/75.79	—	70.89/72.68	88.48/89.63	78.52/78.43	83.60/83.14
Annotátor3	65.61/67.28	70.89/72.68	—	82.01/82.64	75.48/74.29	80.00/78.80
Gold Standard	83.67/84.62	88.48/89.63	82.01/82.64	—	85.57/84.85	90.53/89.33
Szabályalapú r.	75.11/75.56	78.52/78.43	75.48/74.29	85.57/84.85	—	—
Hybrid	78.39/79.42	83.60/83.14	80.00/78.80	90.53/89.33	—	—

Fontos megjegyezni, hogy a címkézést végző szervezeteknek nem volt hozzáférésük az etalon címkékhez, míg az adatbázison tanított statisztikai modellek közvetlenül az etalon címkézést modellezhették. Ennek megfelelően, ha a címkézést végző szervezeteknek lehetőségük lett volna a többségi címkézés jellegzetességeit tanulmányozni, várhatóan nagyobb egyetértési rátát lennének képesek. Másrészről viszont a 3 annotálást végző szervezetnek hatása volt az etalon címkékre, mivel azok az ő többségi szavazásukkal álltak elő. Ez a tény magyarázza, hogy mindhárom szervezet nagyobb egyetértési rátát mutat a többségi címkézéssel, mint a szervezetek címkézései egymással összehasonlítva. Fair összehasonlítást egy olyan, negyedik szervezet által a dokumentumokhoz rendelt címkézés és az etalon jelölés között lehetne tenni, akik előzetesen tanulmányozhatták az etalon címkéket, de arra nem volt kihatásuk. Ez utóbbi statisztika mutatná jól a feladatban elérhető elvi felső korlátot, azaz a képzett szakemberek teljesítményét a BNO-kódolási feladaton.

A különböző szervezetek jelölései között megfigyelhető, feltűnően alacsony egyetértési ráták arra engednek következtetni, hogy az egyes szervezeteknek saját BNO-kódolási stílusuk, szokásaik vannak. A táblázatban szerepeltetjük azon, szabályalapú modellünk és a szervezetek egyetértési rátáit is, mely a BNO-kódolási útmutatók (és nem a címkézett adatok) modellezésével készült. Ez a rendszer az útmutató kifejezéseit illeszti, és kezeli a címkeközi függéseket (az útmutatók általános előírásai szerint), azaz fair összehasonlítást ad a jelölést végző szervezetekkel. Az a tény, hogy a 3

szervezet kissé magasabb egyetértést mutat ezzel a modellel, mint egymással, azt sejteti, hogy a szervezetek sajátos kódolási szokásai a hivatalos BNO-kódolási útmutató eltérő értelmezéséből fakad; illetve, hogy a többségi címkézés jobban közelíti a hivatalos előírásokat, mint a szervezetek egyedi címkézései.

3.2 Összehasonlítás a versenyre benyújtott rendszerekkel

Összesen 50 résztvevő indult a versenyen, a beküldött modellek 89.08% és 15.41% közötti pontosságot mutattak, 76.7% átlagos érték mellett[3]. Összesen 21 rendszer ért el 80% feletti teljesítményt, mely – korlátozott mértékben – összevethető az emberi annotáció pontosságával. A legjobb modellek az emberi címkézés pontosságát közelítik, azaz a klinikai szövegek jó pontosságú gépi feldolgozása reális célkitűzés.

A versenyre beküldött, második helyezést elérő rendszer [1] 88.55% pontosságot ért el, azaz az általunk automatikusan készített legjobb modell (89.33%) megközelíti, vagy kissé meghaladja a más megközelítésekkel elért eredményeket. Ez alapján elmondható, hogy a cikkünkben ismertetett, szabályalapú és gépi tanulási modellek kombinációján alapuló megoldás versenyképes eredményt ad klinikai dokumentumok osztályozásában. Mivel a kézi szabályrendszer kifejlesztésének főbb, munkaigényes lépéseit nagyrészt sikerült automatikusan is reprodukálnunk, – az eredmények jelentős romlása nélkül –, a rendszerünk a versenyben használtnál jelentősen nagyobb számú BNO-kódra is megvalósítható, skálázható lenne.

3 Konklúzió

A kódolási útmutató előírásait alapul vevő szabályalapú szakértői rendszerek meglepően hatékonynak bizonyultak a radiológiai leletek BNO kódolásánál. Ezeken a külső szakértői tudást reprezentáló rendszereken további javítást tudtunk elérni statisztikai gépi tanulási modellek és a szabályhalmazok megfelelő összekapcsolásával. A versenyre beküldött rendszerünk 89.08% pontosságot ért el a leletek egészségügyi kódrendszer kategóriáiba való besorolásában.

A verseny fejlesztési és értékelési időszakát követően végzett, a modellünk skálázhatóságát célzó kutatásaink azt mutatták, hogy a kézileg kiélezett szabályrendszerek helyett (melyek kifejlesztése több ezer kódra időigényes, kivitelezhetetlen lenne) hasonló eredményt érhetünk el, ha a kódolási útmutatókból többé-kevésbé automatikus konverzióval egy kezdeti szabályrendszert állítunk elő, majd ezt a címkézett adatok felhasználásával gépi tanulási modellekkel teljesen automatikusan fejlesztjük tovább.

Hasonló szövegfeldolgozási problémák magyar nyelven való vizsgálatához jelenleg kutatási partnereket keresünk.

[3] Részletes kimutatás található a http://www.computationalmedicine.org/challenge/res.php oldalon.

Bibliográfia

1. Goldstein, I., Arzumtsyan, A., Uzuner, Ö.: Three Approaches to Automatic Assignment of ICD-9-CM Codes to Radiology Reports. Proceedings of the Fall Symposium of the American Medical Informatics Association (AMIA 2007), Chicago, IL, November 10-14, (2007)
2. Lang, D.: Natural Language Processing in the Health Care Industry, Consultant Report, Cincinnati Children's Hospital Medical Center, Cincinatti, Ohio, USA (2006)
3. Moisio M. A.: A Guide to Health Insurance Billing. Thomson Delmar Learning, USA (2006)
4. Pestian J. P., Brew C., Matykiewicz P., Hovermale D. J., Johnson N., Cohen K. B., Duch W.: A shared task involving multi-label classification of clinical free text, In Biological, translational, and clinical language processing, Prague, Czech Republic, 97–104 (2007)
5. György Szarvas, Richárd Farkas, Róbert Busa-Fekete: State-of-the-art anonymisation of medical records using an iterative machine learning framework. Journal of the American Medical Informatics Association, Volume 14, Issue 5, pp 574-580 (2007)