Under ytan på svenska patientjournaler
– ett språkteknologiskt perspektiv

Dimitrios Kokkinakis, forskarassistent i svenska

<*individ*>**Somalisk kvinna**</*individ*> , <*ålder*>**48 år gammal**
</*ålder*> , <*individ*>**8 barn**</*individ*> . **Talar dålig svenska**
så det är svårt att få exakt anamnes . Har kontakt med
<*individ*>**husläk**</*individ*> **i** <*plats*>**Hjällbo**</*plats*> . <*tid*>**För**
8 mån sedan</*tid*> **insatt på beh med** <*substans*>**Seloken**
</*substans*> **samt** <*substans*>**Sorbangil**</*substans*> .
<*tid*>**För 3 mån sedan**</*tid*> **dessutom insatt på beh med**
<*substans*>**Amaryl**</*substans*> **mot** <*sjukdom*>**diabetes typ II**
</*sjukdom*> .

(Utdrag ur en annoterad patientjournal)

Ett problem inom medicinsk och vetenskaplig forskning är att känslig information lätt kan spridas till obehöriga. Den här artikeln ska handla om en metod för anonymisering som gör att patientdata, med minimal risk för obehörig spridning, kan användas för forskningsändamål. Pågående arbeten med att utveckla nationella system för anonymisering i olika länder kan skapa ökade möjligheter att använda individdata från olika källor utan risk för patientintegriteten.

Inom ramen för ett skandinaviskt samarbetsprojekt, Nomen Nescio (latin för 'namnet känner jag icke'), förkortat NN, har ett system för namnigenkänning utvecklats. Systemet har testats för att kunna användas till att anonymisera patientdata och man har utvecklat en kunskapsplattform som innehåller verktyg för namnigenkänning. Verktygen består av språkliga resurser, såsom namnlistor och algoritmiska metoder. Algoritmiska metoder används i datorprogram, som kan känna igen och märka upp namn i

svenska texter. Systemet har under det senaste året utökats med ytterligare funktioner för att kunna känna igen och märka upp medicinska begrepp i texter, t.ex. sjukdomar, läkemedelsnamn samt personreferenser baserade på nyckelord som t.ex. nationalitets- och yrkesbeteckningar.

Vad är anonymisering?

Anonymisering definieras som en process där man ändrar, ersätter eller på något sätt ”döljer” sekretesskänslig information i elektroniska texter, som t.ex. patientjournaler i vår studie. Vilken typ av information det gäller kan variera från fall till fall och från tillämpning till tillämpning, men i princip rör det sig om personuppgifter och demografiska uppgifter (t.ex. namn, kön, ålder, nationalitet, personnummer), platsuppgifter (t.ex. städer, byar, fysiska olycksplatser) eller organisationsnamn (t.ex. företag och myndigheter). Annan information som också kan vara av intresse är t.ex. medicinska begrepp (såsom ovanliga sjukdomstillstånd) samt numeriska uttryck (såsom ”120 mmHg”, ”1200 BTU/min”) och tidsuttryck (såsom ”igår kväll”, ”för 20 år sedan”).

Begreppet *anonymisering* kan vidare indelas i två underavdelningar, nämligen *avidentifiering* och *pseudonymisering*. *Avidentifiering* innebär att man avlägsnar alla uppgifter som kan möjliggöra identifiering på ett sådant sätt att en återkoppling till det ursprungliga identifierbara materialet aldrig kan göras. *Pseudonymisering*, även kallat *kodifiering*, innebär att en nominell identifierare överförs till en annan identifierare som inte på ett enkelt sätt kan kopplas till den ursprungliga. Detta innebär att en markering, en ”nyckel”, kvarstår i den elektroniska texten. Då kan man om behov uppstår återskapa de ursprungliga identitetsuppgifterna med hjälp av ”nyckeln”.

Behovet av anomymisering

Behovet av att förbättra möjligheterna till elektronisk datakommunikation ökar successivt i takt med samhällets utveckling. Till det som bromsar datoriseringen av patientrelaterad information hör en berättigad oro för att information ska komma på avvägar. Information som behandlas inom hälso- och sjukvården rör ofta patienternas personliga integritet och är därmed mycket känslig. Det finns en uppsjö av möjliga praktiska tillämpningar samt forskning av medicinisk och vetenskaplig natur där det är cen-

tralt att skydda den personliga integriteten. Dessa behov kan sammanfattas i följande icke-uttömmande lista:

- klinisk och epidemiologisk forskning som bygger på riktade uttag ur patientjournaler
- möjlighet till sökning i centrala register för forskning och planering av vårdbehov som t.ex. Cancer- eller Abortregistret
- biobanker (dvs. samlingar av humanbiologiskt material som upprättats för kliniska ändamål eller för forskning) som en följd av kartläggningen av människans arvsmassa, och andra framsteg inom den moderna genforskningen, då det finns ett stort intresse bland forskare att använda dessa biobanker för forskningsändamål
- annan medicinsk och vetenskaplig forskning, t.ex. framställning av statistik
- läkemedelsprövningar
- undervisning av t.ex. (nya) läkare eller läkarsekreterare
- hjälp för patienter och anhöriga som vill använda webbtjänster eller annan medicinsk service

Namnigenkänning, stödteknologi för anonymisering

Namnigenkänning är en specialiserad datalingvistisk stödteknologi som går ut på att automatiskt känna igen och förse namn och namnliknande uttryck (kallas också *entiteter*) i löpande text med etiketter, som t.ex. person eller plats. Det igenkänningssystem som vi har utvecklat kan känna igen namn och entitet av åtta olika huvudtyper, nämligen *personer*, *lokaliteter/platser*, *organisationer*, *händelser*, *produkter*, *andliga produkter* (dvs. namn på böcker, filmer och teaterpjäser), *numeriska uttryck* och *tidsuttryck*. Systemet består av ett antal delar som t.ex. en uppsättning grammatikregler för varje namnkategori, listor med flerordsnamn (t.ex. ”American Association of Cancer Research”), listor med enkla namn (t.ex. ”Sverige”) osv. Under 2004 och 2005 har systemet vidareutvecklats för att kunna känna igen och markera även medicinska termer och begrepp, såsom *läkemedelsnamn*, *kemiska substanser*, *sjukdomar*, *symptom*, *mikroorganismer* samt *anatomiska termer* (dvs. kroppsdelar). De flesta av de medicinska begreppen (glosor, ordlistor osv.) har hämtats automatiskt från olika medicinorienterade sidor på Internet, som t.ex. *www.fass.se* och *www.netdoktor.se*.

Medicinska texter för begreppskomplettering

De officiella och inofficiella termlistor vi har hämtat från nätet kommer aldrig att bli helt kompletta eller uppdaterade med det senaste, eftersom utvecklingen inom det medicinska området går så snabbt. De medicinska texter som insamlats från Internet i syfte att komplettera de medicinska termlistorna har även märkts upp efter ordklass för att möjliggöra extrahering av termer baserade på mönstermatchning. Ett mönster som t. ex. ”egennamn+substantiv” har använts för att skapa listor av sjukdomar bestående av flerordsuttryck, (t.ex. ”Biermers[*egennamn*] anemi[*subst.*]”, ”Zollinger-Ellisons[*egennamn*] syndrom[*subst.*]” och ”Ewings[*egennamn*] sarkom[*subst.*]”). Hösten 2005 uppgick den medicinska korpusen till drygt 4 miljoner löpande ord.

Symptombeskrivningar har också extraherats från denna textmassa. Enstaka ord (oftast sammansättningar) eller korta fraser utgör oftast sådana beskrivningar i medicinska texter i allmänhet och patientjournaler i synnerhet. För att kunna konstruera en symptomigenkännare, speciellt i brist på lämpliga lexikala resurser, dvs. listor av symptom som är i paritet med läkemedelslistor (t.ex. från www.fass.se), har vi undersökt hur symptom konstrueras i medicinska texter. Detta genom applicering på texterna av enkla regler baserade på pålitliga sammansättningsefterled (”-rubbning”, ”-smärta”, ”-känsla”, ”-värk”, ”-känslighet”, ”-störning”, ”-problem”, ”-svaghet”, ”-nedsättning” m.m.) och pålitliga symptommönster (”ont i …”, ”smärtor runt…”, ”obehag över …”, ”kramp i …”, ”problem med…” m.m.). Resultatet som erhölls undersöktes, och nya symptomord/fraser i nära kontext med de gamla kunde identifieras och därmed kunde nya uttryck samlas i datorprogrammen och tillämpas på nya texter.

Animathet

Förutom de tidigare beskrivna entiteterna och begreppen finns det också andra informationstyper av intresse i anonymiseringen. Det kan röra sig om ord eller korta fraser som kan signalera animathet, dvs. individer i texten (animat = ’levande’). Detta görs oftast med hjälp av igenkänning av nationalitetsord (t.ex. ”tysken”), yrkesrelaterade ord (t.ex. ”läkaren”) och familjerelationsord (t.ex. ”svärson”).

Utvärdering

Vid utvärderingen av systemet användes ca 25 000 ord i slumpvis extraherade meningar från patientjournaler. Följande autentiska exempel visar mer i detalj olika förenklade uppmärkningar som systemet har försett texten med. ”NUMEX” står för numeriska uttryck, ”TIMEX” för tidsuttryck, ”MEDEX” för medicinska begrepp och ”ENAMEX” för övriga namn och entiteter.

> ”En <ENAMEX>syster</ENAMEX> dött i <MEDEX> hjärtinfarkt</MEDEX> . Lever ensam . <ENAMEX>4 barn </ENAMEX> . Rökt <TIMEX>i många år</TIMEX> men bestämt sig för att sluta . C:a 10 cig/dag . <NUMEX>Sedan 60 års ålder</NUMEX> tilltagande intag av alkohol . <TIMEX>Nu </TIMEX> ej druckit <TIMEX>sedan 5 månader tillbaka </TIMEX> . Medicinerar med <MEDEX>Antabus</MEDEX> . @@ <MEDEX>Ulcerös enterokolit</MEDEX>”.

Syftet med utvärderingen var att beskriva täckningen och precisionen i den automatiska namn- och entitetsigenkänningen. Täckning och precision är de vanligaste sätten att mäta prestation inom analys av naturliga språk. Vi har valt att använda dessa mått individuellt för varje entitet.

Entitet	**Förekomster**	**Precision**	**Täckning**
Personer	314	91,6%	98%
Lokaliteter/Platser	242	95,5%	66%
Organisationer	47	69,9%	70,2%
Produkter	1	100%	100%
Händelser	0	---	---
Andliga produkter	2	75%	75%
Tidsuttryck	716	95,8%	84,6%
Numeriska uttryck	195	100%	---
Animathet	193	96,7%	92,7%
Anatomiska uttryck	175	96,4%	91,6%
Kemiska substanser	227	98%	93,4%
Sjukdomar	260	99,5%	88%
Symptom	227	100%	90,3%
Organismer	2	50%	50%

Precision innebär *antal korrekta enheter* dividerat med *antal identifierade enheter*, och *täckning* beskriver *antal korrekta enheter* dividerat med *totala antalet enheter*. Tabellen på föregående sida visar utvärderingsresultatet för varje typ av entitet (för numeriska uttryck har vi endast beräknat precision).

Negativt påverkande faktorer i igenkänning

Namnigenkänningen påverkas negativt vid analys av patientjournaltextens speciella karaktär och egenskaper. Ett stort antal tidsuttryck i patientjournaler är t.ex. av formatet: ”-*SiffraSiffra*”, t.ex. ”-71” eller ”-85”, ett mönster som inte finns med i systemets igenkänningsmekanism eftersom ett sådant mönster skulle övergenerera uppmärkning av ett antal uttryck som inte förekommer ofta i andra typer av texter. Analysen av resultatet visade att ett stort antal fel berodde på följande faktorer:

- Stavfel: ”senate åren”, ”Siri Lanka”, ”3 vuxna varn”, ”kronisk leukumi”
- Särskrivning av sammansättningar: ”6-barns mor”
- Ogrammatiska konstruktioner: ”1gång/vecka”
- Korta meningar eller otillräcklig kontext gör att system inte kunde märka upp vissa entiteter, särskilt tidsuttryck: ”… pensionerad -88”, ”Stroke -92”
- Förkortningar: ”… vid 59 å å” (’vid 59 års ålder’), ”Har fred och lörd med hustru delat på…” (’Har fredag och lördag med hustru delat på…’), ”ssk” (’sjuksköterska’), ”distr sköt” (’distriktssköterska’)
- Ofullständiga konstruktioner som i relation till innehållet i namnlistorna medfört vissa felaktiga uppmärkningar. I följande exempel har ”Talat”, ”Anger”, ”Ammar” och ”EEG” märkts som personnamn: ”Talat assyriska…”, ”Anger att hustrun…”, ”Ammar.”, ”…svaret från EEG”.

Sammanfattning

Patientjournaler innehåller oerhört värdefull information för både forskare, läkemedelsföretag och myndigheter, men eftersom denna information för närvarande är väl ”gömd” i sjukhusens databaser är det svårt för forskare och sjukvårdspersonal att utnyttja den.

Denna artikel har gett en översiktlig beskrivning av ett system för namn- och terminologiigenkänning, vilket har testats på texter tagna från svenska patientjournaler. Vi har endast testat systemet på en bråkdel av den stora textmassa som patientjournalerna utgör, och vår förhoppning är att vi så småningom ska kunna bidra med att göra en del av denna kunskap tillgänglig för forskning. En del av systemet har utvecklats inom Nomen Nescio-projektet, som under den senaste tiden har utökats med nya igenkänningsmöjligheter. Namnigenkänning är en icke-trivial verksamhet som innefattar olika delmoment, allt från att kunna bestämma att ett eller flera ord verkligen är ett namn, begrepp eller en namngrupp, till att kunna avgöra vilken sorts kategori det/de tillhör.