Friskare klimat med rätt information

Vi bygger MedEval, en medicinsk testkollektion för svensk forskning inom informationssökning

Karin Friberg, doktorand i språkvetenskaplig databehandling

Att hitta dokument med användbar information blir svårare i takt med att mängden information omkring oss växer. Och växer gör den, explosionsartat, i den medicinska världen och annars. Det finns sökverktyg att ta till, men än så länge är det svårt för användare och dator att kommunicera på ett sätt som gör det möjligt för datorn att förstå vad användaren är ute efter. Tänk dig själv, när du söker på Google. Visst kan du få fram intressanta dokument. Men de allra flesta i listan du får är ointressant skräp. Ett problem är att forskning inom informationssökning mest sker på engelska. För att få sökmotorer att fungera bättre på svenska, behövs forskning på svenska. Och för att forska på svenska behövs svenskt forskningsmaterial. För forskning inom informationssökning använder man sig av något som kallas testkollektioner, vilka är bearbetade samlingar av texter. Den här artikeln beskriver hur vi på Språkdata bygger upp just en sådan testkollektion med svenska medicinska texter.

Hur var det då med klimat? Vi önskar väl alla ett bättre klimat inom sjukvårdsapparaten? Om läkare och annan vårdpersonal snabbt och lätt hittar den information han eller hon är ute efter måste vården bli bättre och mer effektiv. Patienter kan också söka information själva. Och informerade patienter är trygga patienter. Ett bättre klimat. Eller hur?

Syfte

Informationssökning (eng. *information retrieval*, IR) handlar om att lagra dokument så att de kan hittas och återvinnas när de är intressanta för en användare som behöver information. Med hjälp av en testkollektion där det är känt vilka dokument som är relevanta för vilka frågeställningar kan

forskaren undersöka olika sökstrategier och se vilken typ av sökfrågor som ger bäst resultat.

En testkollektion består av tre delar. En mängd dokument (eventuellt inom ett speciellt område, som i MedEval), en mängd informationsbehov (ämnen att söka information om) och en mängd relevanta dokument för vart och ett av informationsbehoven. Dessa dokument är delmängder av den större mängden dokument.

För närvarande finns inte någon svensk testkollektion med medicinska dokument och informationsbehov. MedEval-projektet syftar till att skapa just en sådan, så att domänspecifik IR-forskning kan bedrivas på svenska språket. Det är viktigt eftersom engelska skiljer sig från svenska och andra germanska språk med avseende på bildning av sammansättningar och ordens böjningsformer. Därför kan resultat som nåtts inom engelsk informationssökning inte omedelbart överföras till svensk.

MedEval byggs upp till en testkollektion genom att omvandla och utöka den medicinska korpusen, MedLex (en stor samling medicinska texter) som utvecklas vid Språkdata.

Informationssökning

Vid informationssökning finns dokumenten som det söks bland representerade i index. I index finns alla ord från texterna i samlingen med pekare till vart och ett av dokumenten som innehåller detta ord. Användaren ställer sökfrågor till systemet. Orden i sökfrågorna matchas mot orden i index. Slutligen rankar en rankningsfunktion dokumenten i förhållande till deras beräknade relevans gentemot sökfrågan. Ju bättre matchningen är mellan sökfråga och dokument desto högre rankas ett dokument. På så sätt kommer förhoppningsvis relevanta dokument att sökas ut och rankas högt, medan irrelevanta dokument inte kommer att sökas ut, eller åtminstone rankas lägre än mer relevanta dokument och alltså komma längre ner på resultatlistan.

Målet med informationsökning är att återvinna så många dokument som möjligt som är relevanta för informationsbehov, så kallad hög *recall*, och dessutom att ha med så låg andel irrelevanta dokument som möjligt, hög precision. Om precisionen är låg kan de intressanta dokument som återfinns vara svåra att hitta. Användaren kanske inte har tid eller

tålamod att titta igenom mer än de allra högst rankade dokumenten och hittar då inte intressanta dokument som blandats med ointressanta.

Dokumentsamlingen

Insamling av dokument till den medicinska korpusen MedLex har pågått en tid. Korpusen består för tillfället av cirka 48 000 dokument med text från tidskrifter, internet, hälsovårdsmyndigheter och liknande. Förhoppningen är att utöka MedLex-samlingen med någon form av patientdokumentation, i bästa fall patientjournaler, i andra hand medicinska utlåtanden från läkare.

Det första steget i utvecklingen av testkollektionen MedEval går till så att dokumenten från MedLex tokeniseras, det vill säga orden skiljs ut, bland annat genom blanksteg mellan ord och skiljetecken. Dokumenten konverteras till XML-format och märks med taggar (etiketter). Varje dokument har en id-tagg som ger dokumentet en unik identitet. Id-märkningen består av fyra bokstäver vilka representerar källan och ett löpnummer. Datumtagg finns med när datumet då dokumentet publicerades är känt. Det finns också en tagg med adress till källan och slutligen själva brödtexten.

```
<art id="FLKB-0004">
<title> Cell , vävnad , kroppens organisation </title>
<date = "2006-04-XX" />
<source = "http://www.folkbildning.net/konmat/m1_3.htm" />

Någon gång för drygt tre miljarder år sedan föddes den första cellen .
Den uppstod under speciella betingelser i urhavet .
Troligen bildades först s.k. smarta molekyler i form av RNA
( ribonukleinsyra ) .
...
Till flytande vävnad räknas blod och lymfa .
I både blod och lymfa är intercellulärsubstansen flytande och vävnadens
celler cirkulerar i denna vätska .
</art>
```

Exempel på hur dokument i testkollektionen MedEval märks upp med identitet, titel, datum och källa.

När dokumenten är bearbetade på detta sätt förs de över till index. I index finns alla ord från alla dokument listade, och för varje ord finns pekare som talar om i vilka dokument orden återfinns, och i vilken position. Olika index skapas för olika sökmetoder. Index kan exempelvis innehålla orden i deras ursprungliga form eller orden omvandlade till grundform. Sammansättningar kan stå som de är eller tillsammans med ordleden var för sig. Exempelvis kan insulinspruta indexeras som insulinspruta, insulin och spruta.

Informationsbehov

När dokumenten är insamlade skapas informationsbehov (eng. *information need* eller *topic*). Dessa behov används som utgångspunkt för att ställa sökfrågor till systemet. Eftersom det i MedEval är fråga om en specifik kunskapsdomän, den medicinska, har vi anlitat personer med några års läkarstudier bakom sig. Dessa personer har studerat dokumenten i samlingen och skapat behov som ska vara realistiska frågeställningar för två kategorier tilltänkta användare: läkare eller presumtiva patienter.

Drygt 100 informationsbehov skapades i ett första skede, men när de mest lämpliga har valts ut väntar vi oss att det ska landa runt 50 stycken. Behovsskaparna har kontrollerat att det finns ett lämpligt antal relevanta dokument i samlingen för varje behov. Antalet relevanta dokument kan variera men de bör inte vara färre än 5, hellre 50 stycken eller fler för ett behov. Det är bra om det bara är ett fåtal behov som har ett så lågt antal relevanta dokument som 5 eller 10 eftersom recall och precision i dessa fall får en mycket grov variation.

Alla behov består av ett topic-nummer, en titel, <TITLE>, som oftast är en fras, en beskrivning, <DESC>, som oftast är en fråga eller en uppmaning och en narrativ, <NARR>, som är en utförligare beskrivning av ämnet.

<TOP>
<TOPNO>23</TOPNO>
<TITLE> Risker vid användning av neuroleptika </TITLE>
<DESC> Vilka risker är förknippade med användandet av neuroleptika?
</DESC>

<NARR> Relevanta dokument skall innehålla generell information gällande neuroleptika, dess indikationer, biverkningar och behandlingsalternativ. Information om de olika sjukdomstillstånd där neuroleptika används för behandling är relevant. </NARR>
</TOP>

Informationsbehov med taggar från MedEval-kollektionen. Ett informationsbehov består av ett nummer, en titel, en beskrivning och en utförligare narrativ.

Pooling

För att ta reda på vilka dokument som är relevanta för vilka behov skulle det ultimata vara att för varje behov gå igenom alla dokument för att bedöma relevans. Men detta är alldeles för tidskrävande och skulle innebära orimlig arbetsbörda. Om vi räknar med att det tar 10 minuter att bedöma ett dokument och det finns 48 000 dokument och 100 informationsbehov, då skulle det ta fyra personer som arbetade 40 timmar i veckan närmare 100 år att gå igenom samlingen. Detta är naturligtvis inte rimligt. Istället skapar man en så kallad pool av dokument för relevansbedömning.

En pool skapas genom en grov sållning av samlingen med enkla sökfrågor. Output blir en mindre mängd dokument med större sannolikhet att vara relevanta för respektive behov än resten av samlingen. Den strategi som tillämpats för MedEval är att köra fyra olika sökmetoder och för var och en av dem välja ut de 100 högst rankade dokumenten. Detta innebär att man för varje informationsbehov får mellan 100 och 400 dokument att bedöma. 100 dokument om varje sökmetod sammanfaller helt och hållet vad gäller de 100 högst rankade dokumenten. 400 om de fyra listorna inte har några dokument gemensamma på de 100 första platserna. Detta är en avsevärd lättnad mot att för varje behov bedöma kollektionens samtliga 48 000 dokument.

Att skapa en dokumentpool på detta vis gör arbetsbördan rimlig. Samtidigt innebär detta att det inte är säkert att man fångar alla relevanta dokument. Det kan finnas dokument som är relevanta bland dem som inte har kommit med i poolen och alltså inte blivit bedömda.

Relevansbedömning

När dokumenten finns i poolen ska de bedömas. Bedömningarna lagras i en lista där varje informationsbehov har en post för varje dokument som bedömts för relevans till detta behov. Varje post innehåller id för ett informationsbehov, id för ett dokument, eventuellt datum då dokumentet skrevs, en siffra för grad av relevans och en bokstav som representerar vilken målgrupp dokumentet riktar sig till.

Relevansbedömningen görs på en fyrgradig skala, 0–3. Fyra nivåer tillåter differentiering mellan olika sökstrategier, inte bara vad gäller relevanta och irrelevanta dokument, utan även när det gäller graden av återvinning av mycket relevanta dokument jämfört med återvinning av mindre relevanta dokument.

Vid relevansbedömningen har även en bedömning av tilltänkt målgrupp gjorts. Bedömarna har märkt dokumenten med P för målgruppen patienter och M för målgruppen medicinskt utbildad personal. Detta för att sökstrategier ska kunna utvärderas inte bara med tanke på om de återvunna dokumenten är relevanta för det aktuella behovet, utan även huruvida de riktar sig till rätt målgrupp. Förhoppningsvis ska frågor som typiskt ställs av patienter eller typiskt av medicinskt utbildad personal generera dokument avsedda för respektive målgrupp.

Värde	Relevans	Beskrivning
0	Irrelevant	Dokumentet innehåller inte någon information om informationsbehovet.
1	Marginellt relevant	Dokumentet innehåller inte mer information om informationsbehovet än vad som finns i beskrivningen av informationsbehovet.
2	Något relevant	Dokumentet innehåller mer information om informationsbehovet än beskrivningen, men presentationen är inte heltäckande. Om det är ett behov med flera aspekter är endast några av aspekterna berörda.
3	Mycket relevant	Dokumentet diskuterar informationsbehovet i sin helhet. Om det är ett behov med flera aspekter, så berörs alla eller de flesta aspekter i dokumentet.

Den fyrgradiga relevansskalan som används i MedEval-kollektionen.

Materialet

MedEval-projekt handlar om att bygga en svensk testkollektion baserad på insamlade medicinska dokument. Samlingen innefattar det som skrivs på svenska inom den medicinska domänen (utom böcker) som vi kan få tillgång till utan att det strider mot copyright-villkor. Det rör sig om material från medicinska tidskrifter, hälsoinformation från myndigheter, information från FASS och material som finns på internet, till exempel från Sjukvårdsrådgivningen. En viktig del i insamlingsarbetet, som ännu inte är färdig, är att fråga dokumentens upphovsmän om tillstånd att använda deras texter i testkollektionen.

Inom informationssökning vill man forska på material som är så likt det som används i verkligheten som möjligt. En situation där det söks efter dokument är när läkare söker bland patientjournaler, för att hitta tidigare fall som påminner om nuvarande fall och på så vis få hjälp med diagnos och råd om behandling. Därför vill vi i framtiden inkludera avidentifierade patientjournaler i kollektionen. Det får inte finnas risk att någon patient går att identifiera. Därför är det viktigt att arbetet med att anonymisera journaler går framåt.

Framöver

Att kunna bedriva forskning om informationssökning specifikt på medicinska dokument med medicinskt språk är betydelsefullt, inte bara för IR-forskningen, utan även för medicinsk forskning och behandling och för fortbildning av läkare och vårdpersonal.

En svensk medicinsk testkollektion som MedEval kommer dessutom att fylla ett stort behov inom domänspecifik IR-forskning. Den blir förhoppningsvis en resurs för många, inte bara forskare vid Göteborgs universitet. Vi hoppas bli föregångare för liknande kollektioner på flera språk. Om vi kan visa skillnad i resultat för sökstrategier utförda på svenska jämfört med engelska blir det ett incitament för forskare med andra modersmål att skapa testkollektioner på ännu fler språk.

Vi planerar att låta MedEval-kollektionen vara tillgänglig för den som önskar använda den, vilket troligtvis kommer att ge svensk forskning inom informationssökning en skjuts framåt. Vi får ett bättre forsknings-klimat inom svensk domänspecifik informationssökning!