Avidentifiering och pseudonymisering av svensk textdata – erfarenheter från tidigare ansatser främst inom den medicinska/kliniska domänen

Dimitrios Kokkinakis
Språkbanken, institutionen för svenska
Göteborgs universitet
dimitrios.kokkinakis@svenska.gu.se

Bakgrund: den ökande användning av informationsteknik inom sjukvården har medfört en kraftig ökning av elektronisk dokumentation som rör patientens hälsotillstånd, vård och behandling. Vårddokumentationen blir både mer detaljerad och mer individuell, samtidigt som den uppdateras och förändras regelbundet. Patientjournalen är i första hand till för att bidra till en god och säker vård av patienten, men också en viktig informationskälla för FoU. Ett stort hinder för utnyttjandet av journalinformation som forskningskälla är de etiska och rättsliga problemen inte minst den nya dataskyddsförordningen. För att kunna hantera och utnyttja dessa stora och ständigt växande informationsmängder ställs därmed högre krav på säker, skyddad och effektiv informationshantering.

Metod: det kliniska språket kännetecknas av en blandning av och en hög användning av termer; gränssnittsterminologi och jargong; förkortningar, korta, ofullständiga meningar mm. Fortfarande har de stora mängder ostrukturerad text som existerar inom vården återanvänts vid väldigt få tillfällen men uteckling inom språkteknologi har skapat nya möjligheter att stödja FoU även inom den kliniska domänen. För flera år sedan (Kokkinakis & Thurin 2007a,b; Kokkinakis 2010a,b) implementerade vi en pipeline av olika språkteknologiska verktyg som på ett synergistiskt sätt kunde avidentifiera eller pseudonymisera kliniska el. andra typer av textdata. Pipelinen stödde både igenkänning av medicinsk terminologi men även personuppgifter av olika slag (främst namnentitetstyper). Termer ersattes med generiska ID-nummer från olika terminologier (t ex 230739000 för 'ryggradsslag') medan entiteter kunde ersättas av generiska etiketter (t ex PERSON). När det gällde entiteterna kunde man även erbjuda möjlighet att ersätta dessa med neutrala tecken som respekterade ordens längd och utseende (t ex '***-****' eller 'U**-U***' som en ersättare av entiteten 'Ann-Mari').

Resultat: systemet hade inte utvärderats på ett systematiskt sätt pga brist på facit/reference resources ("gold standard") även om de ingående komponenterna hade utvärderats med bra resultatet i tidigare studier.

Sammanfattning och framtidsutsikter: ett ständigt återkommande problem inom vetenskaplig forskning, som begränsar den fria användningen av forskningsdatan, är att känslig information lätt kan spridas till obehöriga. Vi har arbetat med en metod för anonymisering/pseudonymisering som gör att patientdata, med liten risk för obehörig spridning, kunde användas för forskning. Risken är dock kvar och mer arbete återstår för att kunna nå ett ännu bättre resultat. Den skisserade pipelinen har sporadiskt använts i olika ministudier där pseudonymisering har en avgörande betydelse, bland annat i samarbete med polismyndighetens enhet *Nationella underrättelsesystem*. En del av pipelinen har även implementerats om i *Helsinki Finite-State Technology* (HFST); Kokkinakis et al. (2014).

References

Kokkinakis D. and Thurin A. (2007a*). Anonymisation of Swedish Clinical Data*. The 11th Conference on Artificial Intelligence in Medicine (AIME). Pp. 237-241. Amsterdam, Netherlands.

Kokkinakis D. and Thurin A. (2007b). *Identification of Entity References in Hospital Discharge Letters.* The 16th Nordic Conference of Computational Linguistics (NODALIDA). Pp. 329-332. Tartu, Estonia.

Kokkinakis D. (2010a). *Is data scrubbing useful for anonymizing sensitive data?* The 3rd Swedish Language Technology Conference (SLTC-2010). Linköping and also CLT-2010 in Gullmarsstrand.

Kokkinakis D. (2010b). *Är "data scrubbing" en användbar metod för att anonymisera känsliga patientdata?* Svenska Läkaresällskapets Riksstämman. 1-3 Dec. 2010. Göteborg.

Kokkinakis D., Niemi J., Hardwick S., Lindén K. and Borin L. (2014). HFST-SweNER: a New NER Resource for Swedish. The 9th edition of the Language Resources and Evaluation Conference (LREC). Reykjavik, Iceland.