

Normalisering av patientjournaler skrivna på svenska.

Examensarbetsförslag

Vi har inom KEA Kunskapsextraktionsagentprojektet, <http://people.dsv.su.se/~hercules/kea.html> tillgång till ett stort antal avidentifierade patientjournaler. Vi skall från dessa journaler göra textmining och klustring och genom dessa tekniker få fram ny okänd information som t.ex samband mellan sjukdomar, läkemedel, kost, dryckesvanor, socialt liv, m.m. Tyvärr är patientjournalerna ibland snabbt skrivna med många ej standardiserade förkortningar, journalerna innehåller också ett flertal stavfel eller alternativa stavningar. Detta gör det svårt för våra textmining- och klustringsverktyg att verka på rätt sätt; varför vi skulle behöva ”tvätta” patientjournalerna. Det betyder att vi hittar olika förkortningsvarianter och för ihop dem till samma term eller att felstavade ord korrigeras. Genom att göra detta blir patientjournalerna normaliserade med avseende på felstavningar och tvetydiga förkortningar.

Exempel på förkortningar är *rtg* som oftast betyder *röntgen*, och *pat* som kan betyda både *patient* eller *patologisk*. Här kan man t.ex. titta på ordklassinformation och omkringliggande ord.

Exempel på felstavningar är *undersöknar*, *utersökning* som båda troligtvis betyder *undersökning*

Stöd för att korrigera dessa termer är färdiga förkortningslistor som skriver ut förkortningarna i fullständiga form. Vi kan även använda ett rättstavningsprogram som t.ex. Stava eller ordklasstaggare GTA, båda från KTH, som kan hjälpa oss att komma med förslag på rättstavade termer, ordklass och frasgränsinformation. Ordlistorna som vi kan använda är till exempel FASS och slutligen titta på alla patientjournalers index och frekvenser.

Handledare: Dr. Hercules Dalianis, docent, Institutionen för data- och systemvetenskap
DSV/KTH Stockholms universitet, tel 070-568 13 59 e-post: hercules@dsv.su.se