INSTITUTIONEN FÖR BIOMEDICINSK TEKNIK | LUNDS TEKNISKA HÖGSKOLA | PRESENTERAD 2021-06-04

EXAMENSARBETE Natural Language Processing for Patient Data in Clinical Decision Support Systems
Naturlig språkbehandling av patientjournaler till kliniskt beslutsstöd
STUDENTER Amanda Nilsson, Lina Samnegård
HANDLEDARE Pierre Nugues (LTH), Oskar Strand (Siemens Healthineers)
EXAMINATOR Martin Stridh (LTH)

Naturlig språkbehandling och prostatacancer

POPULÄRVETENSKAPLIG SAMMANFATTNING **Amanda Nilsson, Lina Samnegård**

Naturlig språkbehandling är ett område inom maskininlärning för att förstå och generera text och tal. I detta arbete använder vi naturlig språkbehandling på patientjournaltexter relaterade till prostatacancer, för att klassificera dem samt extrahera relevant information.

Prostatacancer är den vanligaste typen av cancer bland män och varje år dör 2400 i sjukdomen. Vårdflödet för prostatacancer är komplext och innefattar många olika steg. Risken att insjukna i prostatacancer ökar drastiskt med åldern och vid 80 års ålder har 1/5 män sjukdomen. Med en befolkning som blir äldre, kommer allt fler få sjukdomen och vårdbehovet kommer att öka. Med en redan pressad sjukvård, är digitalisering nödvändigt för att förenkla för sjukvårdspersonal. Ett verktyg som kan användas är kliniska beslutsstöd. De kan implementeras med hjälp av naturlig språkbehandling, för att på ett strukturerat sätt visa patientens sjukdomsinformation.

Syftet med vårt projekt var att göra ett proof of concept. Detta för att se om det var möjligt att extrahera nyckelinformation kopplat till prostatacancer från löpande text. Vårt arbete inleddes med en litteraturstudie och intervjuer för att kartlägga vårdprocessen för prostatacancer. Vi lade fokus på att förstå vad som händer i de olika stegen samt hur data lagras. Parallellt med det här började vi implementera vårt program. För att kunna träna maskininlärningsalgoritmen använde vi data från journaltexter kopplade till prostatacancer. Vi valde ut tre olika värden som skulle extraheras från texterna. I det första steget tränades klassificerare för att ta bort irrelevanta texter (som ej är kopplade till prostatacancer). Därefter implementerades en maskininlärningsmetod för att hitta och extrahera de valda värdena. Detta kallas för named entity recognition (NER). Vi testade olika modeller för både klassificerare och NER för att se vilken som gav bäst resultat.

Klassificeringen blev bäst då vi använde en modell vid namn random forest med F-score på 0,978. För NER blev resultatet bäst då vi använde språkmodellen BERT som var förtränad på svensk text. Då blev F-score 0,922. För att bygga vår slutliga arkitetektur kombinerades en klassificerare med tre olika NER, se bild nedan. Detta gav en F-score på 0,911. Jämfört med andra projekt inom samma område är detta ett väldigt bra resultat.