

Kandidatexjobb: Lyrics Writing Assistant

Introduktion

Poesi, eller lyrik överlag, är någonting som anses vara en mänsklig konstform. Lyrik baseras oftast på känslor, med en sensmoral som huvudmål, och står ofta som ett bevis på en människas kreativa tankebanor. De kan berätta om romanser, tragedier, drömmar eller resor. En dator är, utom all rimlig tvivel, fullständigt oförmögen att förstå något av dessa ting. Men skulle en dator kunna vara till hjälp när en poet är fast i skrivkramp?

Redan idag finns mjukvara för att assistera vid lyrikproduktion[2], men många av dessa agerar enbart praktiska uppslagsverk för att hitta synonymer och ord som rimmar och spara dem som hjälp när de kan komma till att behövas. De saknar aktiva försök att hjälpa till med meningsbyggnaderna.

Tanken om det går att träna en dator för att komplettera lyriska strofer är spännande. Mycket forskning görs för att få datorer att kunna skriva, chatta och assistera vid framställning av formellare texter. Eftersom lyrik är en konstform så är kriterierna för vad man behöver för assistans en annan.

Problemformulering

Vi ämnar undersöka vad det är som är viktigt att ett sådant program baserar sina förslag på för att det ska anses vara användbart. Vad är det som spelar störst roll för att man ska få bra hjälp med att skriva allt från ett enkelt julklappsrim till en livsavgörande kärleksförklaring?

Är det bästa att man håller den exakta diktstrukturen? Är det bästa att rimmen är felfria? Är det bästa att förslagen är baserade på tidigare vanligt förekommande ord? Kanske är det en viktig kombination av dessa som ger bäst resultat. Vi ämnar i så fall söka oss till den optimala kombinationen.

Tillvägagångssätt

Vi kommer att börja vår undersökning med litteraturstudier för att identifiera viktiga aspekter inom språkteknologi, fonetik och angränsande ämnen som är av värde för studien.

Sedan följer en fas med implementation av en programvara som kan producera text. Mjukvaran grundar sig på två typer av verktyg. Det ena är ett fonetiskt lexikon, det vill säga en ordbok där varje ord presenteras både i klartext och i dess fonetiska form. Det andra verktyget är en så kallad språkmodell - en statistisk modell som kan uttrycka troliga ordföljder baserat på material den är tränad på. En frekvensanalys hos så kallade *n-gram*[3] är en relativt enkel, men samtidigt

effektiv sådan språkmodell. Diktstrukturer (haiku, limerick) är också någonting som kan tänkas komma till användning för att definiera hur en dikt FÅR se ut över huvud taget.

Vår mjukvara kommer försöka föreslå hur en dikt eller låtttext ska fortsättas baserat på det som användaren redan har matat in. Programmet kommer kunna basera sin output på olika typer av texter, som tidningstexter, dikter, sångtexter etc. eller en kombination av flera olika typer. Som en avancerad utökning vore det intressant att se om man kan tillverka ett program som klarar av att göra *ungefärliga* matchningar, för att på så sätt bredda möjligheterna till att hitta rätt ord för sammanhanget. Till exempel ungefärlig strängmatchning[4] kommer då komma bra till hands.

Därpå följer en empirisk undersökning som syftar till att svara på vår frågeställning. I undersökningen ska vår programvara användas och på olika sätt modifieras för att kunna jämföra olika algoritmer eller datakällor. Förslagsvis kan man be folk testa applikationen när de ska skriva julklappsrim, och se vilken kombination algoritmer som visade sig ge det bästa rimmet.

Referenser

1. **Elementär fonetik**, Eva Lundström-Holmberg, Peter af Trampe (1987)
2. **Master Writer**. Hjälpreda med bl.a. rimlexikon, ordbok, synonymer
(<http://masterwriter.com/>)
3. **n-gram**, en sannolikhetsmodell för sekvenser, baserat på Markovmodeller
(<http://en.wikipedia.org/wiki/N-gram>)
4. **The SecondString project**. Ett bibliotek för ungefärlig strängmatchning
(<http://secondstring.sourceforge.net/>)

Tidsplan

24 februari Program för att behandla data och bygga språkmodeller ska vara färdig

10 mars Mjukvaran ska vara klar för testning

17 mars Vi ska ha gjort våra empiriska studier och tester

12 april ska en första version av uppsatsen vara inlämnad

25-26 april Presentation av uppsatsen

TBD Slutversion av uppsats ska lämnas in