Slutrapport

Transkribering av känsligt material

Innehållsförteckning

Inledning

Referenser

[1]

Definitioner

Begrepp	Beskrivning

Versionshistorik

Version	Datum	Ändrad av	Kommentarer
0.1	2023-12-22	Lena Strömbäck	Första utkast

Sammanfattning

Projektet syftar till att ta fram en lösning för transkribering av ljudfiler som innehåller känsligt material/sekretess eller känsliga personuppgifter. För känsliga personuppgifter är en lösning att drift sker lokalt på lärosätet och inte via en molnbaserad tjänst.

Bakgrund

Inom forskning finns ett stort behov av transkribering av ljudfiler innehållandes känsligt material/sekretess eller personliga uppgifter. Detta gör det svårt att nyttja molnbaserade lösningar för dessa transkriberingar. Det finns dock tekniska lösningar för att driftsätta en transkriberingslösning lokalt på lärosätet genom att nyttja containerbaserade transkriberingslösningar. Möjliga tjänster listas nedan, där vi valt att gå vidare med Azures lösning i detta projekt.

- Speech to text containers - Speech service - Azure AI services | Microsoft Learn
- The Azure AI Speech CLI - Azure AI services | Microsoft Learn
- Speech to text - OpenAI API

Avgränsningar

Projektet hade som mål att fram en fungerande lösning som driftsätts och testas på minst två lärosäten lokalt. Produktifiering och driftsättning av gemensam tjänst ingick inte som del av projektet.

Resultat

Vi har utvecklat en lösning som driftsatts i testmiljö vid Linköping och Umeå Universitet. Kod, dokument och instruktioner för att sätta upp lösningen finns lagrat på ITCFs GiTHub-repository. Detta dokument sammanfattar våra erfarenheter från projektet och användartester samt ger förslag på hur lösningen kan vidareutvecklas.

Kommenterad [MG1]: Menar vi samma instans, eller "baserad på samma kod"

Kommenterad [PH2R1]: I projektdirektivet/iden så är min uppfattning att vi menade samma instans. Detta är dock "på sikt"

Programkod och dokumentation

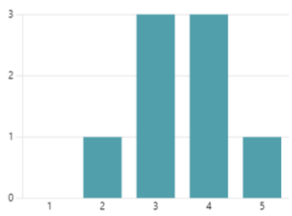
Programkod och dokumentation för den som vill använda lösningen finns tillgängligt på ITCFs GiTHub-repository https://github.com/ITCF-projects/Transcription.

Där finns också instruktioner för hur man gå tillväga för att sätta upp systemet i Docker, Kubernetes samt PODMAN samt en vägledning till uppsättning av autentisering.

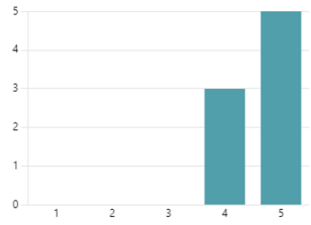
Driftsatta lösningar och resultat från användartestning

Tjänsten Transcribe har sedan den 18:e december 2023 funnits i testdrift vid Linköping och Umeå universitet. I båda fallen är det en testversion som varit tillgängliga för utvalda användare för test av funktionalitet och transkriberingskvalitet.

Vid Linköpings Universitet körs systemet i dagsläget med Podman på en virtuell Linux server. Vid Umeå Universitet körs systemet med Docker i WSL på en befintlig virtuell WS2022 server. Hittills har 4 forskare vid Linköpings Universitet och 4 forskare vid Umeå Universitet fyllt i vår utvärderingsenkät.



De forskare som testat tycker generellt att kvalitén på det som transkriberats är bra. Positiva omdömen är till exempel att transkriberingen fångade alla ord, att det blir korrekt svenska, och går att få ett sammanhang. Någon säger att de är positivt överraskad, men att det transkriberade kräver en genomgång efteråt. Mer negativa kommentarer är att de saknar information om talare och hade önskat en ny rad vid varje röst, att texten delas konstigt i flera delar och att det blir svårt när fler pratar samtidigt. De påpekar också att systemet ibland gissar ord som blir felaktiga.



De allra flesta tycker att systemet är enkelt att komma igång med. De är relativt nöjda med formatet för indatafiler och utdatafiler. Dock finns flera önskemål på hur utdatafiler kan bearbetas vidare för att enklare kunna analyseras av forskarna.

Hälften av forskarna anger att de skulle vilja använda ett system som detta. Resterande säger att de eventuellt vill använda det. Det som generellt är avgörande är om de anser att systemet sparar tid genom att de får ett första utkast av transkriberat material som sedan kan bearbetas vidare.

Fortsatta planer vid de deltagande lärosätena:

För de lärosäten som ingått i projektet finns följande planer för driftsättning av tjänsten:

Linköping

Vid Linköpings Universitet planerar vi att kunna gå i drift med en enklare version av systemet under 2024. Aktiviteter som måste genomföras innan det är möjligt är:

- Beslut om förvaltningsorganisation och systemägare,
- Justeringar av befintlig lösning:
  - Undersöka driftsäkerhet bland annat PODMAN
- Krav för drift för känsliga persondata:
  - Dokumentation, riskanalys och konsekvensbedömning av tjänsten
  - Tekniska justeringar baserat på analysen.
- En enkel möjlighet till fakturering av användning vilket kräver en administratörsinloggning.

Förutom detta finns ytterligare funktionalitet som vi anser är önskvärd för att förenkla för våra användare och förenkla administrationen av systemet. Vi kommer att undersöka vilka av dessa som är lämpliga att genomföra innan en första driftsättning av systemet.

- Utökad funktionalitet:
  - Bättre information till användaren om hur lång tid transkriberingen kan ta genom information om antingen hur många som ligger för i kön eller uppskattade väntetider.
  - Notifiering via mail om när jobbet är klart.
  - Anpassade utdataformat för att forskaren enklare ska kunna göra vidare bearbetning av transkriberingen.
- Vidareutveckling för att kunna debitera forskargrupper
  - Roller som möjliggör administratörsinloggning
  - Faktureringsinformation och statistik som enkelt kan laddas ner och utgöra bas för debitering till användare.
  - Profilinformation för användare som möjliggör att de kan spara faktureringsinformation och även se statistik över sina kostnader och användning.

Umeå

- Utredning av förvaltningsorganisation (vem ska vara mottagare)
  - Driftsmiljöanalys
  - Utvärdering av alternativ
    - (Sunet tal-till-text (amberscript)) legalitet känslig personinformation.
    - Whisper
  - Framtagande av formaliadokument (Informationsklassning, Datahanteringsplan, RSA, bevarandeplan et.c.)
- Vidareutveckling av tjänst
  - Förtydliga enligt utfall från acceptanstest.
  - Förbättrad felhantering
    - Transkribering blir “Completed” trots avstängda transkriberingscontainrar.
    - Transkribering stannar för alla vid vissa transkriberingsfel.
  - Införa tjänsteövervakning och healthchecks.
  - Export av loggar till central loganalys
  - Status-sida eller uppladdningsstatus “Underhåll pågår”
- Behov betalningsmodell och uppdelning på kostnadsställe.
  - Initialt tas kostnad centralt, vid skenande kostnader analyseras framkomlig väg.

- Behov Umu-look and feel?
    - Eventuellt vissa css-ändringar, men inget stort planerat.

Luleå

- Driftsätt i testmiljö för att kunna presentera.
- Utred behov från verksamheten.
- Utredning av förvaltningsorganisation (vem ska vara mottagare).

Göteborg

- Göteborg har en transkriberingslösning baserad på Whisper och följer detta projekt.

Förslag på vidare aktiviteter för att få en lösning som kan produktifieras och förvaltas på flera lärosäten

Kommenterad [PH3]: Målet i projektplan : "Ge förslag på vidare aktiviteter för att lösningen skall kunna produktifieras och förvaltas som gemensam tjänst för lärosäten. "

Projektgruppen har identifierat att varje lärosäte som vill driva denna tjänst ställer delvis olika krav beroende på driftmiljö och krav för hantering av känsliga data. Därför finns behov av olika varianter av lösningen. Som en del av leveransen finns dokumenterat hur detta kan ske, bland annat hur lösningen kan anpassas till olika containerlösningar.

En möjlig framtidslösning är att ett eller ett fåtal universitet erbjuder denna tjänst och tillhandahåller den för andra lärosäten. Detta kräver en genomgång av vilka lagar som är applicerbara och vilka avtal som behöver skrivas mellan lärosätena. Förutom detta skulle en sådan lösning ställa ett antal ytterligare tekniska krav:

- Lösningen måste anpassas för inloggning via SWAMID
- Högre krav på roller i systemet med bland annat behov av administrationsfunktioner.
- En automatiserad kostnadshantering
- Högre krav på skalbarhet vid högre belastning för att undvika långa svarstider.

Ytterligare en möjlighet är att mindre lärosäten får hjälp med driften. Det kan ske antingen genom en paketlösning som de får hjälp att köra på sina egna servrar eller möjligen med teknik för så kallad confidential computing där driften sker på ett annat lärosäte men på ett sätt som garanterar att ingen kan få tillgång till data.

Utökad funktionalitet som identifierats av projektgruppen

Under arbetet har vi identifierat följande funktionalitet som skulle kunna intressant för framtida versioner av systemet:

- Om man vill tillåta användare utanför det egna lärosätet eller en gemensam tjänst för flera lärosäten behöver inloggning via BankID eller SWAMID undersökas.
- Vi har för närvarande implementerat gränssnittet enbart på engelska, men det kan vara önskvärt att erbjuda det på svenska också. Andra förbättringar av gränssnittet skulle kunna vara att tillåta uppladdning med drag and drop, att kunna ge en beskrivning av sin fil för att lättare identifiera dem.
- Den nuvarande lösningen bygger på att varje transkribering består av en textfil med tillhörande termlista. Man kan tänka sig att det skulle kunna underlätta för användaren att i stället göra ett paket av flera filer som skickas som ett jobb till transkriberingsmotorn.
- Det skulle gå att förenkla hanteringen av termer. Ett exempel är att kunna ladda upp en termlista för att forskaren enkelt ska kunna spara sina listor mellan olika transkriberingar. Ett

annat är att starta om ett jobb med en modifierad termlista om man tror att det kan förbättra kvalitén på transkriberingen.
- Om användaren debiteras för transkribering finns en del utökad funktionalitet som vore önskvärd. Här har vi till exempel diskuterat, förvalt kostnadsställe, statistik över transkriberingar och kostnader, möjlighet att få beräknad tid och kostnad innan transkriberingen startas.
- Om det blir längre kö och transkriberingstider är det önskvärt att användaren får information om när transkriberingen är klar via e-mail.
- Vi har också diskuterat utökad hantering av ljudfiler, till exempel möjlighet att lyssna på och ladda ner dem. Man kan också tänka sig möjlighet till konvertering av indatafiler för att underlätta för användaren.
- Man kan också tänka sig att material raderas automatiskt vid nedladdning eller att användaren har möjlighet att påverka den tid som materialet finns kvar på servern.

Utökad funktionalitet som identifierats vid användartester

Från de användartester som gjorts kan vi i huvudsak se följande önskemål på ytterligare funktionalitet som gör det enklare att bearbeta resultatet vidare:

- I de flesta fall transkriberas material som bygger på intervjuer eller andra diskussioner med två eller flera av talarna. Därför skulle det vara användbart med talaridentifiering av de textavsnitt som finns i resultatet. Detta är sannolikt möjligt med användning av funktioner från andra delar av MS paket, men det är oklart hur mycket arbete det innebär.
- I många fall bryts texten där talaren gör en paus och inte i slutet av en mening vilket möjligen kan åtgärdas.
- Titta på om vi det går att processa resultatet på ett sätt som gör det enklare att bearbeta vidare i Word eller de analysprogram som används av forskarna.

Plan för gemensam vidareutveckling

Plan för hur vidareutveckling vid ett universitet ska göras tillgänglig för alla universitet i Sverige. Koden görs tillgänglig under licensen CC BY-NC-SA, dvs eventuella tillägg ska nämna ursprungskällan, delas under samma villkor och vi tillåter bara icke kommersiell användning. Vi förordar att den som använder koden också matar tillbaka sin vidareutveckling till repositioriet så att den kan komma alla till nytta.

Kommenterad [MG4]: Finns publikt repo tillgängligt i ITCF:s regi. Vilka förhållningsregler gäller där? Funkar det för detta syfte, vilka alternativ är bättre lämpade?

Kommenterad [PH5R4]: Tanken är att skapa upp ett repo för ITCF och lägga koden där.

Projektgruppen ser behov av en fortsatt gemensam utveckling då det finns ett flertal punkter identifierade där en gemensam vidareutveckling av koden skulle vara önskvärd. De punkter som vi framförallt ser som aktuella för ett sådant projekt skulle vara:

- Uppsättning av produktionsmiljö för säker drift av tjänsten. Här har vi idag gjort en första utredning men mer arbete behövs för att sätta upp systemen och underlätta för andra som vill sätta upp systemet i sin egen driftsmiljö.
- Fortsätta utvärdera transkriberingskvalitéten och göra en jämförelse med andra löningar. Vi har sett vissa begränsningar i Microsofts lösning gällande ordlista och felhantering vid längre pausar. Intressant att göra det möjligt att välja transkriberingslösning inom samma system.
- Utveckla roller för administration för att underlätta för att underlätta för debitering och statistik i systemet.

- Efterbearbetning av resultatet för att forskare enkelt ska kunna arbeta vidare med dem i de verktyg som de föredrar.
- Ytterligare utveckling som utvecklar kvalitén. Bland annat undersöka möjligheten till identifiering av talare i en dialog.

Sammanfattande erfarenheter

Projektet har resulterat i testuppsättningar av en transkriberingsmotor vid Linköpings och Umeå Universitet. Det har varit givande att samarbeta och utbyta erfarenheter mellan lärosätena både gällande tekniska detaljer i projektet och mer allmänna diskussioner om olika lösningar för drift av system.

Våra erfarenheter var att det var relativt enkelt att komma igång med och använda Microsofts lösning. De forskare som testat verkar också i stort set nöjda med resultatet. Vi ser dock behov av vidareutveckling både gällande transkriptionsfunktionaliteten och gällande funktionalitet som underlättar drift och administration av systemet.