Bilaga 5: Nytt språk svenska

Wikispeech –

en användargenererad talsyntes på Wikipedia

Innehållsförteckning

Introduktion

Det här dokumentet beskriver vad som behövs för att bygga den svenska rösten. Motsvarande gäller för engelska och arabiska, även om behoven för exempelvis lexikon kan skilja sig åt, men detta beslutas under arbetet med att identifiera/inventera resurser.

Nyckel:

Grön = Walking skeleton
Gul = MVP
Vit = Möjlig vidareutveckling

1 Intresse

Kommunicering av intresse

2 Identifiera existerande resurser

Identifiera existerande resurser:

1. Fonemuppsättning
2. Övriga tagset (ordklasser mm)
3. Övriga specifikationer och definitioner (ISO-språktaggar, etc)
4. Uttalslexikon
5. g2p-resurser
6. Taldatabas för syntes, med uppmärkning

3 API-anpassningar

Eventuella anpassningar av API och infrastruktur som behövs för att hantera språket ifråga

4 Komponenter

4.1 NLP

4.1.1 Textprocessning

1. Enkel tokenisering: Splitta på mellanslag + skiljetecken
2. Siffergenerering: Konvertering siffror => ord
3. Siffergenerering: Ordningstal
4. Siffergenerering: Enkla fall av romerska siffror, vanliga kungar osv
5. Vissa typer av datum
6. Normal indata - normal utdata: Om indata är någorlunda “normal” för språket i fråga, ska komponenten generera en uttalbar uppmärkning, och inte heller bli tyst, hänga sig eller liknande.
7. Frasering: Kort paus vid skiljetecken, längre paus vid meningsslut
8. Frasering: Reducerad betoning på funktionsord
9. Förkortningar: Hantera (expandera) vanliga förkortningar som “t.ex.”, “osv”, “mm”, “dvs” …

4.1.2 Uttalskomponent

1. Fylla lexikondatabasen med innehåll från fritt tillgängliga källor
2. g2p-regler
3. Modifiera uttalslexikonet efter behov (ta bort överflödig data, lägga till ev. uppmärkning som saknas)
4. Normal indata - normal utdata: Om indata är någorlunda “normal” för språket i fråga, ska komponenten generera en uttalbar uppmärkning, och inte heller bli tyst, hänga sig eller liknande.
5. Validering
6. Sammansättningskomponent för okända ord

4.2 Syntes

1. Använd existerande röst (marytts)
2. Normal indata - normal utdata.
3. Använd annan röst
4. Bygg ny röst
5. Anpassa existerande röst

4.2.1 Taldatabas

Behövs inte i första skedet för svenska.

1. Använd tidigare inspelad databas för att bygga ny röst
2. Spela in ny taldata (studio eller via wikipedia)

5 Installation

1. Manuell installation av utvecklare (dokumentation)
2. Manuell konfiguration på server (dokumentation)

6 Konfigurering

1. Manuell konfiguration av gemenskap (dokumentation)

7 Specifikationer

7.1 Intresse

steg 1:

Komponentnamn: Intresse
Beskrivning: Kommunicering av intresse
Existerande: ja
Att göra: Inget, svenska är redan specificerat i ansökan

7.2 Identifiera existerande resurser

steg 2, egenskap 1:

Komponentnamn: Fonemuppsättning
Beskrivning: Definiera fonemuppsättning
Existerande: SAMPA https://www.phon.ucl.ac.uk/home/sampa/swedish.htm
Att göra: Vi utgår från SAMPA och beslutar om anpassning vid behov

steg 2, egenskap 2:

Komponentnamn: Övriga tagset
Beskrivning: Identifiera vilka övriga tagset som behövs, och inventera/definiera dessa (ordklasser mm)
Existerande: delvis
Att göra: Identifiera vilka övriga tagset som behövs, och inventera/definiera dessa (ordklasser mm)

steg 2, egenskap 3:

Komponentnamn: Övriga specifikationer och definitioner
Beskrivning: Identifiera vilka övriga specifikationer och definitioner som behövs för språket ifråga -- språktaggar med mera
Existerande: delvis
Att göra: Identifiera vilka övriga specifikationer och definitioner som behövs för språket ifråga -- språktaggar med mera

steg 2, egenskap 4:

Komponentnamn: Uttalslexikon
Beskrivning: Inventera vilka resurser som finns för uttalslexikon
Existerande: delvis
Att göra: Inventera vilka resurser som finns för uttalslexikon

steg 2, egenskap 5:

Komponentnamn: g2p-resurser
Beskrivning: Inventera vilka resurser som finns för g2p
Existerande: delvis (även delvis samma som resurser för uttalslexikon)
Att göra: Inventera vilka resurser som finns för g2p

steg 2, egenskap 5:

Komponentnamn: Syntesdata
Beskrivning: Inventera vilka resurser som finns för syntesdata (taldatabaser, uppmärkning mm)
Existerande: delvis
Att göra: Inventera vilka resurser som finns för syntesdata (taldatabaser, uppmärkning mm)

7.3 API-anpassningar

steg 3:

Komponentnamn: API-anpassningar
Beskrivning: Eventuella anpassningar av API och infrastruktur som behövs för att hantera språket ifråga
Existerande: nej
Att göra: Identifiera och implementera+testa eventuella anpassningar av API och infrastruktur som behövs för att hantera språket ifråga

7.4 Komponenter

7.4.1 NLP

7.4.1.1 Textprocessning

Steg 4.1.1, egenskap 1:

Komponentnamn: Enkel tokenisering
Beskrivning: Enkel tokenisering som splittar på mellanslag och skiljetecken
Existerande: nej
Att göra: Skriva en enkelt tokeniseringskomponent som splittar på mellanslag och skiljetecken

Steg 4.1.1, egenskap 2:

Komponentnamn: Siffergenerering: Konvertera siffror till ord
Beskrivning: Komponent som konverterar vanliga siffror till ord
Existerande: nej
Att göra: Skriva och testa en komponent som konverterar siffror till ord. Avgränsa vilka siffror som ska hanteras

Steg 4.1.1, egenskap 3:

Komponentnamn: Siffergenerering: ordningstal
Beskrivning: Komponent som hanterar vanliga ordningstal
Existerande: nej
Att göra: Skriva och testa en komponent som hanterar vanliga ordningstal. Avgränsa vilka siffror som ska hanteras

Steg 4.1.1, egenskap 4:

Komponentnamn: Siffergenerering: romerska siffror
Beskrivning: Komponent som hanterar vanliga romerska siffror
Existerande: nej

Att göra: Skriva och testa en komponent som hanterar vanliga romerska siffror. Avgränsa vilka siffror som ska hanteras

Steg 4.1.1, egenskap 5:

Komponentnamn: Datum
Beskrivning: Komponent som hanterar vanliga datumtyper
Existerande: nej
Att göra: Skriva och testa en komponent som hanterar vanliga typer av datum. Avgränsa vilka datum/-format som ska hanteras

Steg 4.1.1, egenskap 6:

Komponentnamn: Normal indata - normal utdata
Beskrivning: Om indata är någorlunda "normal" för språket i fråga, ska komponenten generera en uttalbar uppmärkning, och inte heller bli tyst, hänga sig eller liknande.
Existerande: nej
Att göra: Definiera vad som är "normal" indata och implementera + testa hantering av detta (fallbacklösningar behövs i de fall man får in okänd data)

Steg 4.1.1, egenskap 7:

Komponentnamn: Frasering
Beskrivning: Fraseringskomponent som sätter ut pauser vid skiljetecken och meningsslut
Existerande: nej
Att göra: Skriva och testa en enkel fraseringskomponent som sätter ut korta pauser vid skiljetecken och lite längre pauser meningsslut**.**

7.4.1.2 Uttalskomponent

Steg 4.1.2, egenskap 1:

Komponentnamn: Fylla lexikondatabasen med innehåll från fritt tillgängliga källor
Beskrivning: Fylla lexikondatabasen med innehåll från fritt tillgängliga källor.
Existerande: delvis
Att göra: Preprocessning av tillgängliga uttalslexikon:

- Analysera indata och bedöma vad man kan använda
- Omformatera tillgängliga lexikon till det nya textformatet anpassat till databasen
- Validera lexikonen (format, tillåtna/obligatoriska fält, osv)
- Stoppa in uttalslexikonen i databasen

Testa lexikonen -- funktion och prestanda

Steg 4.1.2, egenskap 2:

Komponentnamn: g2p-regler
Beskrivning: Bygga g2p-regler
Existerande: Delvis. Det finns mjukvara för att bygga och senare anropa regler.

Att göra: Välja g2p-metod; Konvertera lexikondata till korrekt inputformat; Bygga och testa reglerna; Bygga och testa fallbacklösning (för svenska: alla a-z med diverse diakritiska tecken ska ge en transkription); “Montera” reglerna i Wikispeech

Steg 4.1.2, egenskap 3:

Komponentnamn: Modifiera lexikonet efter behov
Beskrivning: Automatiskt eller halvautomatiskt komplettera ev. uppmärkning som saknas i lexikon men som behövs i Wikispeech-systemet; eller rensa lexikonet på överflödig/felaktig data
Existerande: nej
Att göra: Inventering görs under “konvertera lexikon” ovan. Utifrån detta identifiera vad som ev. behöver kompletteras, och göra detta. Det kan exempelvis handla om att lägga till uppmärkning såsom sammansättningsuppdelning eller annat. Det kan också handla om att bedöma kvalitetsnivån i lexikonet

Steg 4.1.2, egenskap 4:

Komponentnamn: Normal indata - normal utdata
Beskrivning: Om indata är någorlunda “normal” för språket i fråga, ska komponenten generera en uttalbar uppmärkning, och inte heller bli tyst, hänga sig eller liknande.
Existerande: Nej
Att göra: Definiera vad som är “normal” indata. För svenska bör uttalskomponenten kunna hantera a-z med valfria diakritiska tecken. Siffror och andra icke-alfabetiska tecken ska processas tidigare i textprocessningen, men om g2p-modell saknas för exempelvis “ë” så bör man kunna konfigurera uttalskomponenten (i praktiken g2p:n) att välja uttal för “e” istället.

Steg 4.1.2, egenskap 5:

Komponentnamn: Validering
Beskrivning: Validering av transkriptioner
Existerande: Nej
Att göra: Automatisk validering av transkriptioner. Fonemen behöver valideras (bara tillåtna symboler får användas). Sedan behövs syntas/format-valdering (exempel: hur kan betoningar placeras ut, hur många vokaler/konsonanter kan det finnas i varje stavelse, hur placerar man ut stavelse-/morfemgränser, hur kan konsonanter kombineras i början/slutet på stavelser). Man kan också lägga till “sanity checks” där man jämför ortografi och transkription, och bedömer vad som är rimligt. Man kan också göra mer specifika regler som hanterar vanligt förekommande delsträngar och så vidare. Möjligheterna är i princip oändliga.

Steg 4.1.2, egenskap 6:

Komponentnamn: Sammansättningskomponent
Beskrivning: Sammansättningskomponent för okända ord
Existerande: Nej
Att göra: Bygga upp lexikon av för-/mellan-/efterled: ortografier för analys och ortografier + transkriptioner för generering av transkriptioner (kan innehålla fler

uppslag än analysdelen); Bygga regler för att analysera okända ord och hitta sammansättningsled; Bygga regler för hur man sätter ihop transkriptionerna (betoning, assimilation)

7.4.2 Syntes

steg 4.2, egenskap 1:

Komponentnamn: syntes
Beskrivning: Använd existerande svensk röst i marytts
Existerande: ja
Att göra: givet att driver finns för marytts: installation, testning, felrapportering

steg 4.2, egenskap 2:

Komponentnamn: syntes
Beskrivning: normal indata - normal utdata
Existerande: ja
Att göra: givet att driver finns för marytts: installation, testning, felrapportering

steg 4.2, egenskap 3:

Komponentnamn: syntes
Beskrivning: Använd annan svensk röst
Existerande: -
Att göra: givet att driver finns för marytts: installation, testning, felrapportering

7.4.2.1 Taldatabas

Behövs inte i första skedet, utan först senare, för att förbättra eller anpassa röst, eller bygga ny.

7.5 Installation

steg 5, egenskap 1:

Komponentnamn: Manuell installation av utvecklare
Beskrivning: Alla komponenter för svenska installeras i servermiljön.
Existerande: nej
Att göra: Installera nödvändiga komponenter

steg 5, egenskap 2:

Komponentnamn: Manuell konfiguration på server
Beskrivning: Konfigurera allt som behövs som ligger på serversidan, t.ex. vilka
Existerande: nej
Att göra: Konfigurera, koordinera med Wikimedia Foundations utvecklare.

7.6 Konfigurering

steg 6, egenskap 1:

Komponentnamn: Manuell konfiguration av gemenskap (dokumentation)

Beskrivning: Koordinera med den svenska Wikipedia-gemenskapen så att saker blir uppsatt korrekt i enlighet med gemenskapens önskemål.

Existerande: nej

Att göra: Kommunicera med gemenskapen och stötta lokal konfigurering.