



Kriterien zur Auswahl eines Sprachkorpus

Zusammenfassung des Berichtes

Daniel Zoba, 13.01.2022



Gliederung (1)

- Einleitung
- Textkorpuserstellung
 - Textkorpusbeschaffung
 - Textkorpusnachverarbeitung
 - Textkorpusanalyse
 - Vokabulardefinition
 - Grammatikerstellung
- Aufnahmesatzauswahl
 - Phonologische Verteilung
 - Auswahl phonetisch vielfältiger Sätze
 - Werkzeuge



Gliederung (2)

- Sprachkorpuserstellung
 - Detaillierte Sprachkorpusspezifikation
 - Sprecherakquise
 - Aufnahmeprotokoll
 - Aufnahmesoftware
 - Sprachsignalqualität
 - Qualitätssicherung der Aufnahmesitzung
 - Umgang mit den erfassten & aufgezeichneten Daten
 - Rechtliche Aspekte
 - Datenschutz



Textkorpuserstellung (1)

- Textkorpusbeschaffung:
 - Domänendefinition: Welche Anwendung will ich erstellen?
 - Use-cases definieren: Welche Funktionsweise erwarten die Nutzer von der Anwendung?
 - Methoden zur Korpusbeschaffung:
 - Automatisierte Webseitenbeschaffung und –auswahl
 - Use-Case-Szenarien manuell erstellen
 - Textdatenerweiterungsmethoden:
 - Manuelle Erweiterung (Experten)
 - Einbeziehung von Texten fremder Domänen zur Grammatikerstellung
 - „Wizard-of-Oz“ Simulation des Systems mit nicht eingeweihten Nutzern
 - Vorteile / Nachteile / Empfehlung



Textkorpuserstellung (2)

- Textkorpuserstellung:
 - Reine Textdateien im UTF-8 Encoding erstellen, ggfls manuell kontrollieren
 - Fremdwörter, Rechtschreibfehler usw. korrigieren/entfernen
 - „Normalisierung“:
 - Nummern, Datumsangaben, Abkürzungen, Akronyme („PDF“) usw. korrekt „ausformulieren“
 - 1 = eins
 - 3.4.2021 = dritter april zweitausendeiundzwanzig / dritter vierter zweitausendeiundzwanzig
 - PDF = peh deh eff
 - 200€ = zweihundert euro
 - VI = vau ie / sechs (je nach Kontext)
 - Groß- und Kleinschreibung vereinheitlichen: Jedes Wort sollte nur in einer Variante enthalten sein
 - Zielstellung: Korpus für Erstellung von Vokabular und Transkriptionen geeignet



Textkorpuserstellung (3)

- Textkorporusanalyse
 - Typische Analyseparameter:
 - Worthäufigkeit
 - Wortanordnungen (mehr als nur Vorgänger / Nachfolger)
 - Konkordanzen (wichtige Wörter und Phrasen)
 - Wortgruppenhäufigkeit
 - Eigennamen
 - uvm.
 - Zielstellung: Sammlung von Informationen und Eigenschaften zur Entwicklung des Sprachmodells und der Applikation für diese Domäne



Textkorpuserstellung (4)

- Vokabulardefinition:
 - Händische Erstellung
 - Automatisiert aus dem normalisierten Korpus:
 - Wortliste aus Korpus erstellen
 - Liste sortieren
 - Duplikate entfernen
 - Zielstellung: Vokabularliste kann zur Erstellung des Aussprachelexikons verwendet werden



Textkorpuserstellung (5)

- Wortklassen:
 - Worte derselben semantischen Bedeutung (Datum, Uhrzeit, Eigennamen) werden in Wortklassen zusammengefasst
 - Sprachmodelle benutzen Wortklassen lediglich als Platzhalter/Label
 - Zielstellung: Deutliche Reduktion der Komplexität eines Sprachmodells
 - Offene Wortklassen: z.B. Eigennamen (quasi nie komplett)
 - Geschlossene Wortklassen: z.B. Wochentage (ändern sich nicht)
- Wortklassen auf morphologischer Basis sind für Sprachen wie Ober- und Niedersorbisch empfehlenswert
- Wortklassen müssen von Anfang an definiert werden, da die Textkorpusverarbeitung davon abhängt
- Wortklassen können manuell oder automatisiert aus den vorhandenen Textkorpora erstellt werden



Textkorpuserstellung (5)

- Grammatikerstellung (UASR):
 - Grammatik besteht aus Regeln für
 - Lexikon (1:1 Abbildung orthographischer String auf phonetische Transkription)
 - „LEX: ČAS tSas“
 - Varianten möglich: „LEX: bisher blsh(e|E:)(r|6)“ → 4 Varianten kompakt beschrieben
 - Grammatik
 - „Normale“ (endliche) Regeln, linkes Symbol ist ein benannter Zustand
 - GRM: (S) :_TELLTIME_ <TIME> (F)
 - Kontextfreie Regeln, linkes Symbol ist kein terminaler Zustand
 - GRM: <TIME> KAK:KAK JE:JE NA:NA ČASU:ČASU
- Erstellung von Dialogen wird unterstützt (z.B. Kommandowort)



Aufnahmesatzauswahl

- Phonologische Verteilung
 - Idealerweise ist jede mögliche Kombination eines Phonemes mit Vorgängern und Nachfolgern („Triphonen“) gleich häufig enthalten
- Auswahl phonetisch vielfältiger Sätze
 - Graphem- und Phonemliste sowie Ausspracheregeln („exceptions“) benötigt
 - Auswahl von Sätzen aus dem normalisierten Korpus



Sprachkorpuserstellung (1)

- Detaillierte Sprachkorpusspezifikation
 - Möglichen maximalen Re-use sicherstellen, da Korpuserstellung teuer
 - Alle Metadaten müssen penibel aufgezeichnet werden
 - Merkmale des Korpus
 - Methoden der Qualitätssicherung
 - Aufnahmebedingungen
 - Sprechereigenschaften
 - ...
- Sprecherakquise
 - Anzahl, Geschlecht, Alter, Dialekt, Bildung, Muttersprachler, ...
 - Sprachstil: Vorlesen, Fragen beantworten, Vorgefertigtes Vokabular, freies Sprechen



Sprachkorpuserstellung (2)

- Aufnahmeprotokoll
 - maschinenlesbar, so ausführlich wie möglich
 - Datum, Uhrzeit
 - Aufnahmeumgebung
 - Sprecheridentifikation (nicht der Name)
 - Parameter des Aufnahmeequipments
 - ...
- Aufnahmesoftware
 - „BAS SpeechRecorder“ frei verfügbar, quasi open-source
 - Kann mit Mühe geforkt werden, keine kollaborative Entwicklung



Sprachkorpuserstellung (3)

- Sprachsignalqualität:
 - Amplitude, Aussteuerung, Hintergrundgeräusche, Raumakustik ... beachten
- Qualitätssicherung der Aufnahmesitzung
 - Kontinuierlicher Prozess während der gesamten Erstellung
 - Sprecherüberwachung (Monitoring) während der Aufnahme
 - z.B. Korrektur bei Aussprachefehlern
 - Nachträgliche Überprüfung (Validation) von Aufnahmen
 - Es gibt nicht den „einen richtigen“ Prozess



Sprachkorpuserstellung (4)

- Umgang mit den erfassten & aufgezeichneten Daten
 - Sprecher-Identifikationen
 - Identifizierung geheimhalten, aber ID immer gleich lassen
 - vermeidet doppelte Aufzeichnungen
 - Ermöglicht späteres Entfernen der Aufnahmen eines Sprechers
 - Dateilisten
 - Werden für das spätere Training benötigt
 - Müssen ggfls verarbeitet / aufgeteilt werden:
 - Trainingsdaten vs Evaluierungsdaten (ein Sprecher nie in beiden Listen!)
 - Speicherplatz und Sicherheit
 - Datenmenge vorher kalkulieren und Speicherplatz organisieren
 - Sicherstellen dass Datentransfer kein Bottleneck wird (z.B. Sprecher sollen nicht warten)
 - Backups
 - Zugriffssicherheit (Verschlüsselung, Abschließen, ...)



Sprachkorpuserstellung (5)

- Rechtliche Aspekte
 - Urheber- und Nutzungsrechte
- Datenschutz
 - Alle Daten speziell sichern, die zur Identifikation eines Sprechers dienen könnten
 - Namen, Adressen, Telefonnummern...



Ende