Spracherkennung

Auswertung der Kooperation mit Fraunhofer IKTS und BTU Cottbus

Daniel Sobe, 15.06.2022

Arbeitspakete 2021

- Akustisches Modell (aufgezeichnete Sprache → Phonemkette):
  - Übergang auf ein obersorbisches Modell
  - Sprecheranpassung und Anpassung an den akustischen Pfad
- Wörterbuch (Phoneme → Wortkette):
  - Verbesserung der Ausspracheregeln
- Sprachmodell (Wörter → erlaubte Wortgruppen):
  - Technologieverbesserung für die Erstellung großer Modelle
- Aufzeichnungen:
  - Anleitung zur Satzauswahl und Aufnahmebedingungen

AUFZEICHNUNGEN

Satzauswahl und Aufnahmebedingungen (1)

- Beschaffung und Bearbeitung eines Textkorpus
- Auswahl phonetisch ausgewogener und reichhaltiger Sätze
- Aufnahmen mit verschiedenen Sprechern
- https://github.com/ZalozbaDev/speech_recognition_corpus_creation
  - ausführlicher Bericht auf Englisch
  - Zusammenfassung auf Deutsch

Satzauswahl und Aufnahmebedingungen (2)

- Bearbeiteter Textkorpus kann weiterverwendet werden:
  - Phonetisches Lexikon
  - Sprachmodell

- Beispiele zur Erstellung eines phonetischen Lexikons aus einem Textkorpus unter Zuhilfename eines Regelwerkes ebenfalls in diesem Github-Repository

WÖRTERBUCH

Phonetisches Lexikon – Regeln (1)

- Jedem Buchstaben eines Wortes wird ein Laut zugeordnet („kanonische Aussprache“)
- Zusätzliche Regeln beschreiben Abweichungen von dieser Zuordnung
- Grundsatz bei Spracherkennung:
  - So viele Lautmodelle wie nötig, so wenig wie möglich
  - Worte mit unterschiedlichen Bedeutungen müssen mit unterschiedlichen Lauten modelliert werden (Tschechisch: byt – Wohnung, být – sein)

Phonetisches Lexikon – Regeln (2)

- Beispiele „kanonische Aussprache“:
  - C ts
  - Č tS
  - Ć tS
  - D d
  - E E
  - CH x
- Beispiele für „Ausnahmeregeln“:
  - _D_K t zrědka, doprědka, srědki, přehladka, wuslědkow …
  - _E_DŹ e chcedźa, srjedź, njeskedźbliwa, swjedźenskim …
  - E_J_I * stejimy, meji, čejim
  - I_CH_ C serbskich, młódšich, nichtó, jich ...
  - Ausnahmeregel „E → e“ wird z.B. überprüft ob überhaupt nötig (keine „Sinnveränderung“)

Phonetisches Lexikon – Regeln (3)

- Resultat:
  - ANTIKSKICH a n t i k s k i C
  - ANTIKSKICH a n t i k s k i x
  - ANTIKSKICH Q a n t i k s k i C
  - ANTIKSKICH Q a n t i k s k i x
- Weiterverwendung für:
  - Phonetische Beschriftung von Aufnahmen (unverändert)
  - Erkennungslexikon (manuell korrigiert, ggfls Dialekte hinzugefügt)
- Beispiele: https://github.com/ZalozbaDev/speech_recognition_corpus_creation

Phonetisches Lexikon – Statistik

- Wenn ein genügend großes phonetisches Lexikon zur Verfügung steht, kann es mit Hilfe von statistischen Werkzeugen erweitert werden
  - phonetisaurus predict --model=models/g2p.fst --nbest=5 ĆOPŁOBĚŁU
    - ĆOPŁOBĚŁU tS O p w O b I w u
    - ĆOPŁOBĚŁU tS O p w O b I u
    - ĆOPŁOBĚŁU tS O p w O b e u
    - ĆOPŁOBĚŁU tS O p w O b e i u
    - ĆOPŁOBĚŁU tS U p w O b I w u
- https://github.com/ZalozbaDev/speech_recognition_statistical_g2p_modeling

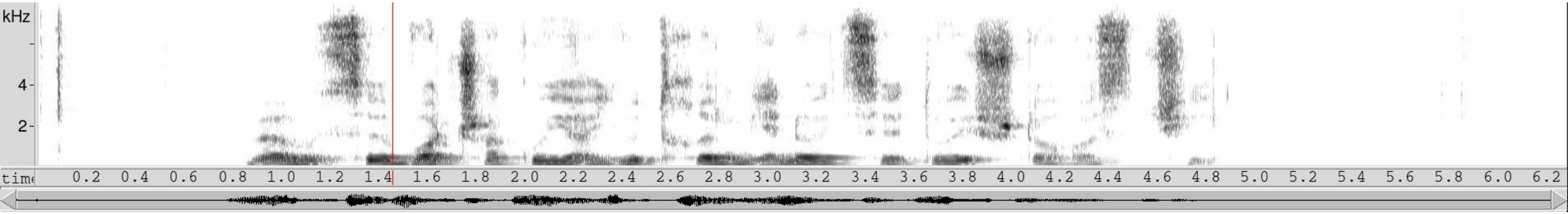
AKUSTISCHES MODELL

Phonetische Beschriftung (1)

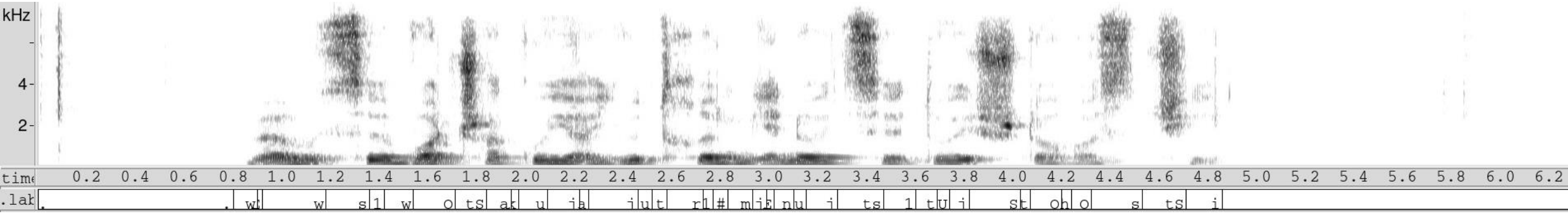
- Training akustischer Modelle benötigt phonetisch beschriftete Sprachaufzeichnungen
- Händisch erstellen: sehr aufwändig (und damit unrealistisch)
- Automatisiert erzeugen: mit möglichst wenig Fehlern, da nur stichprobenartige Kontrolle realistisch

Phonetische Beschriftung (2)

WE WSY WOČAKUJA JUTRY MJENUJCY TÓJŠTO HOSĆI



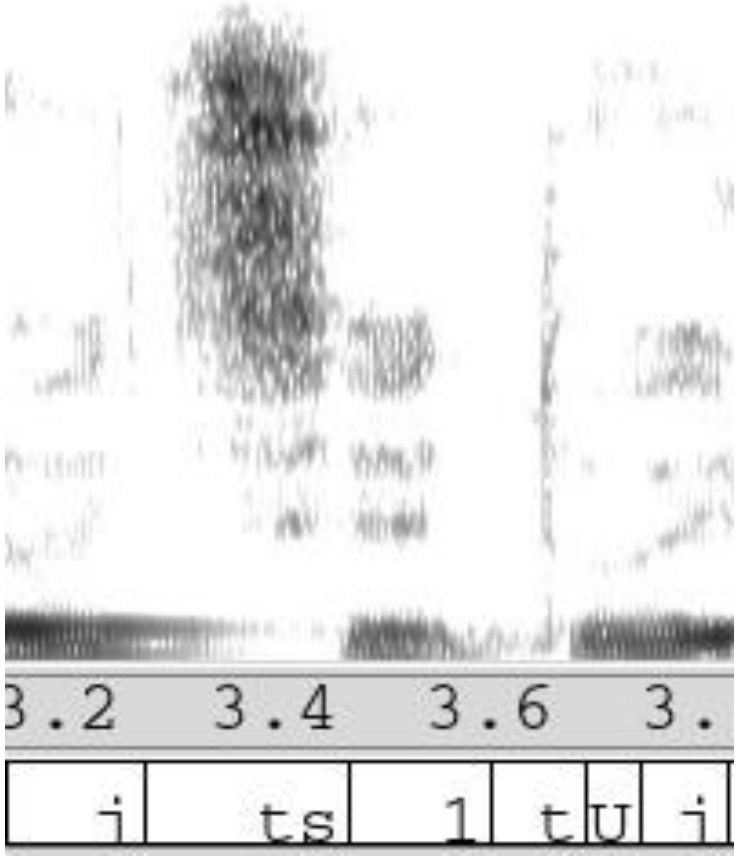
Gewünschtes Resultat



Phonetische Beschriftung (3)

MJENUJCY TÓJŠTO

Zoom in:



Phonetische Beschriftung (4)

- Zuhilfenahme eines „umetikettierten“ deutschen akustischen Modells
  - 'Z': ['z','S']
  - 'dZ': ['d','z','S']
  - 'e': ['e:']
  - 'ji': ['j','i:']
  - 'jn': ['j','n']
  - 's': ['s']
  - 't': ['t']
  - 'tS': ['t','S']
  - 'ts': ['t','s‘]

Phonetische Beschriftung (5)

- Akustisches Modell + (bekannter) Text + Wörterbuch

  → Automatisierte phonetische Beschriftung

- Keine Beschränkung auf eine bestimmte Sprache

- Aber: Phoneme müssen mit Hilfe der deutschen Modelle unterscheidbar sein (nicht notwendigerweise „erkennbar“)

Training

- Aktuell verfügbare Modelle können mit bis zu 50..60 Stunden aufgezeichneter Sprache mehrerer Sprecher in Studioqualität trainiert werden
- Technologische Grenzen der verwendeten Monophonmodelle:
  - Nur Aufnahmen guter Qualität
  - Modelle sind nach 60 Stunden Sprachmaterial „voll“
- Training eines guten Basismodelles möglich, aber kein Universalmodell

Adaption

- Anpassung des Basismodells an Sprecher und Mikrofon
- Aufzeichnung von z.B. 40 Sätzen unter „Feldbedingungen“
- Deutliche Verbesserung der akustischen Erkennung
- Anwendung z.B. für Diktiersysteme oder persönliche Assistenten
- https://github.com/ZalozbaDev/speech_recognition_acoustic_model_training

SPRACHMODELL

Word classes (1)

- Sprachmodell beschränkt Wortschatz auf „erlaubte Äußerungen“
- Je mehr Einschränkungen, desto bessere Spracherkennung
- Aber: Jede Äußerung muss in die Grammatik aufgenommen werden
- Ausgangspunkt: Bearbeiteter Textkorpus

Word classes (2)

- Problem: Komplexität wächst
  - GRM.1: <PERCENT> :_005_ PJEĆ:PJEĆ
  - GRM.1: <PERCENT> :_010_ DŹESAĆ:010
  - GRM.1: <PERCENT> :_015_ PJATNAĆ:015
  - GRM.1: <PERCENT> :_015_ PJATNAĆE:015
  - GRM.1: <PERCENT> :_020_ DWACEĆI:020
  - GRM.1: <PERCENT> :_020_ DWACĆI:020
  - GRM.1: <PERCENT> :_020_ DWACCI:020
  - GRM.1: <PERCENT> :_025_ PJEĆADWACEĆI:025
  - GRM.1: <PERCENT> :_025_ PJEĆADWACĆI:025
  - GRM.1: <PERCENT> :_025_ PJEĆADWACCI:025
  - ...
- Zahlen, Datum, Uhrzeiten in allen Varianten nicht mehr realistisch darstellbar

Word classes (3)

- Lösung: Erstellen von (verschachtelten) „word classes“
- Grammatik kann mit Platzhaltern einfach beschrieben werden
- GRM: <BRIGHTNESS> <LAMP> <NUM1-99> PROCENTOW

Word classes (4)

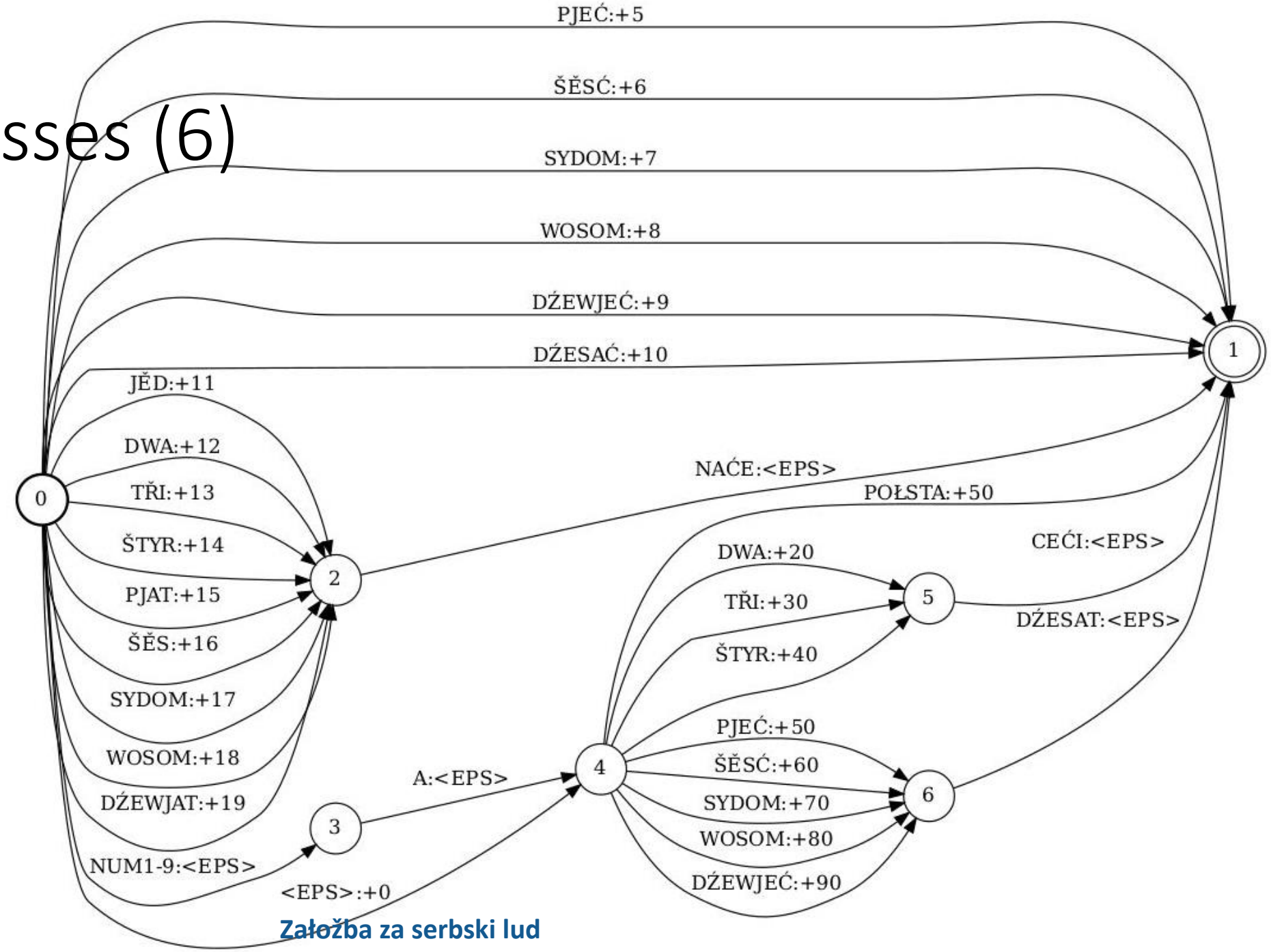
- NUM1-99.txt:
  - 0 1 jedyn +1
  - 0 1 dwaj +2
  - 0 1 tři +3
  - 0 1 štyri +4
  - 0 1 NUM5-99 <eps>
  - 1

Word classes (5)

- NUM5-99.txt:
  - 0 6 pjeć +5
  - …
  - 0 6 dźesać +10
  - 0 1 jěd +11
  - 0 1 dwa +12
  - …
  - 0 1 dźewjat +19
  - 1 6 naće <eps>
  - 0 2 NUM1-9 <eps>
  - 2 3 a <eps>
  - 0 3 <eps> +0
  - 3 4 dwa +20
  - 3 4 tři +30
  - …
  - 4 6 ceći <eps>
  - 3 5 pjeć +50
  - …
  - 5 6 dźesat <eps>
  - 3 6 połsta +50
  - 6

Word classes (6)

- NUM5-99.txt:



Word classes (7)

- NUM1-9.txt:
  - 0 1 jedyn +1
  - 0 1 dwaj +2
  - 0 1 tři +3
  - 0 1 štyri +4
  - 0 1 pjeć +5
  - 0 1 šěsć +6
  - 0 1 sydom +7
  - 0 1 wosom +8
  - 0 1 dźewjeć +9
  - 1

Statistisches Sprachmodell (1)

- Probleme mit großen Sprachmodellen:
  - Können nicht mehr manuell erstellt werden
  - Benötigen immer mehr Rechenzeit
- Kompromiss:
  - Unwahrscheinliche Äußerungen werden entfernt
- Vorgehensweise:
  - Erstellung eines statistischen „n-gramm“ Sprachmodells aus dem bearbeiteten Textkorpus
  - Der Textkorpus kann vorher für die Verwendung von „word classes“ modifiziert werden

Statistisches Sprachmodell (2)

- Kleine Textkorpora: Bigramme
  - Sprachmodell „kennt“ nur ausgewählte „Nachfolger“ eines Wortes:
    - dobre
      - ranje
      - polěpšenje
      - wjedro
      - ~~smykaće~~
- Große Textkorpora: Trigramme
  - Sprachmodell „kennt“ nur ausgewählte „Nachfolger“ der letzten 2 Worte:
    - sweča so
      - hasnje
      - swěći
      - ~~polěpši~~

Statistisches Sprachmodell (3)

- Mögliche Anwendung: Diktierfunktion

- Einschränkung: fehlende direkte Zuordnung von „Aktionen“, wie z.B. bei einem „smart home“
  - Kann teilweise durch den Einsatz von „word classes“ kompensiert werden

- https://github.com/ZalozbaDev/speech_recognition_language_modeling

AUSBLICK

Ausblick

- ~~Veröffentlichung aller Ergebnisse aus 2021~~
- Niedersorbisch
- Geplante Kooperationen:
  - Sorbisches Institut: Phonetik / Linguistik für Ober- und Niedersorbisch
  - RCW: Kombination Spracherkennung mit sotra.app
  - FHG und BTU: Anwendungen mit großem Sprachmodell
- Hilfe bei der persönlichen Benutzung und Anpassung des „digidom“-Prototyps
- Weitere Aufzeichnungen

Ende / Fragen