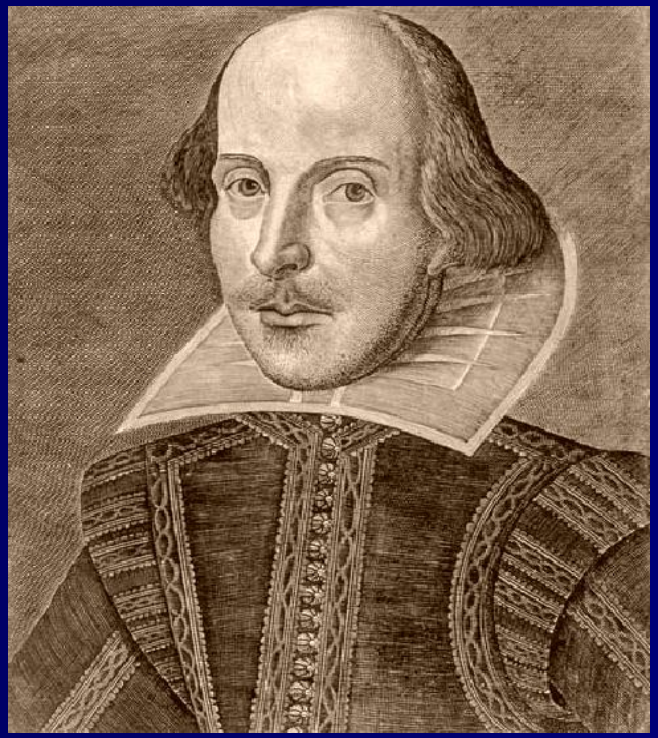
Shakespeare's Monkeys

Spracherkennung mittels Markov-Modellen



Übersicht I

- Idee
  - Shakespeare's Monkeys
  - Buchstabenhäufigkeiten
- Funktionsprinzip
  - Markov-Modell
  - Häufigkeitsvergleich mittels MSE
  - Training und Erkennung

Übersicht II

- Implementierung
  - Einschränkungen/Filterung
  - Umsetzung in Perl/HTML
  - Training (3 Sprachen)
- Erkennung
  - Erkennungsraten
  - Demonstration

Idee

Shakespeare's Monkeys

- Affen an Schreibmaschinen schreiben Werk Shakespeares mit gewisser (wenn auch sehr geringer) Wahrscheinlichkeit
- Aufeinanderfolgende Buchstaben sind mehr oder weniger zufällig → Sprachbesonderheiten (Häufigkeit) unberücksichtigt → dauert lange

Buchstabenhäufigkeiten I

- Bestimmte Buchstaben kommen in bestimmten Sprachen häufiger vor als andere (im Deutschen z.B. das E)
- Idee: „Virtuelle Affen“, die beim Tippen Buchstabenhäufigkeiten berücksichtigen
- Betrachtung einzelne Buchstaben: Ordnung 0

Buchstabenhäufigkeiten II

- Verbesserung: Häufigkeiten von Buchstabenübergängen betrachten. Z.B. ‚Q' → ‚U' im Deutschen wahrscheinlicher als ‚Q' → ‚X'
- Häufigkeit → Wahrscheinlichkeit
- Betrachtung Buchstabenübergänge: Ordnung 1 (bei einzelnen Buchstaben)

Buchstabenhäufigkeiten III

- Weitere Verbesserung: Buchstabenfolgen berücksichtigen. Z.B. ‚AB' → ‚ER' im Deutschen wahrscheinlicher als ‚AB' → ‚XY'

  ‚XJ' → ‚QY' im Deutschen so gut wie unmöglich (sehr unwahrscheinlich)
- Buchstabenanzahl = Ordnung (hier 2)

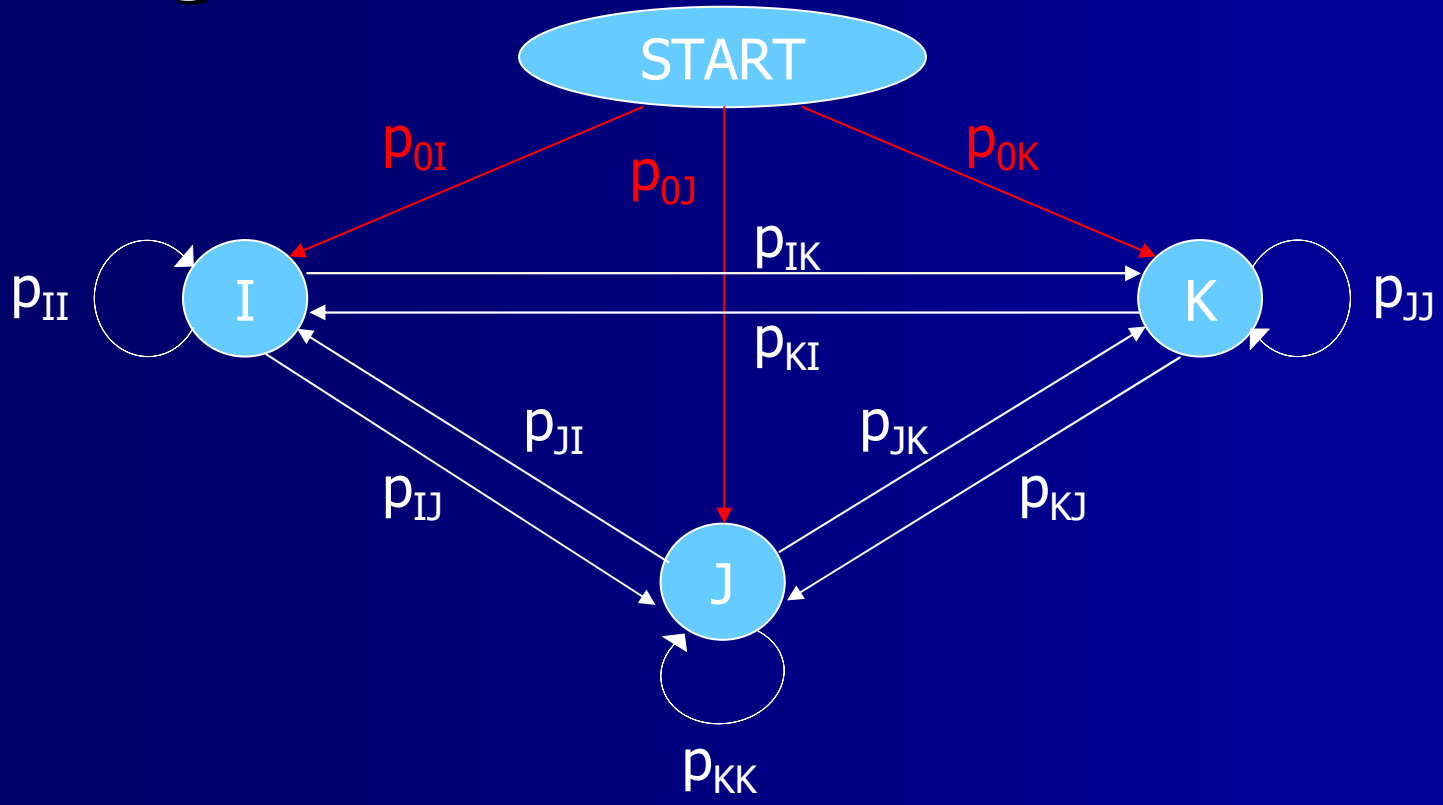
Funktionsprinzip

Markov-Modell I

- Buchstabenübergänge = Zustandsübergänge → Markov-Modell
- Zustandsübergänge mit vorgegebener Wahrscheinlichkeit
- Wahrscheinlichkeiten werden aus Trainingstexten ermitelt
- Beispiel: Vereinfachung: Alphabet IJK

Markov-Modell II

Ordnung 1:



Markov-Modell III

- Speicherung der Wahrscheinlichkeiten der Zustandsübergänge in einer Matrix

$$\begin{pmatrix} p_{II} & p_{IJ} & p_{IK} \\ p_{JI} & p_{JJ} & p_{JK} \\ p_{KI} & p_{KJ} & p_{KK} \end{pmatrix}$$

Häufigkeitsvergleich I

- Unterschiedliche Übergangswahrscheinlichkeiten je Sprache → Vergleich gespeicherte und analysierte Wahrscheinlichkeiten
- Beste Übereinstimmung → erkannte Sprache (geringste Abweichung der Übergangswahrscheinlichkeiten)

Häufigkeitsvergleich II

- Methode: MSE (Mean Squared Error, mittlerer quadratischer Fehler):

$$MSE_s = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i(s))^2$$

- Y_i = Wahrscheinlichkeit im Text
- X_i = Wahrscheinlichkeit in Sprache s

Häufigkeitsvergleich III

- Sprache mit kleinstem MSE → erkannte Sprache
- Einschränkung: Fehler wird nur über im angegebenen Text vorkommende Wahrscheinlichkeiten aufsummiert, nicht über alle. Z.B. ‚ABCDEF'. Nur Wahrscheinlichkeit ‚AB' → 'CD' und ‚CD' → ‚EF' wird verglichen

Training und Erkennung

- Training: Häufigkeiten von langen Texten der jeweiligen Sprache werden analysiert und die Wahrscheinlichkeiten gespeichert
- Erkennung: Häufigkeiten eines gegebenen Textes analysieren und per MSE mit gespeicherten Wahrscheinlichkeiten vergleichen

Implementierung

Einschränkung/Filterung

- Reduktion auf Alphabet mit 27 Zeichen (Großbuchstaben und Leerzeichen)
- Entfernung aller anderen Zeichen aus dem Text
- Spezialfälle: ä → ae, ö → oe, ü → ue
- Weitere Filterung möglich

Umsetzung in Perl

- Definition eines mehrdimensionalen Arrays (Workaround)
- Berechnung Indizes ähnlich einem Zahlensystems mit Basis 27

  A = 0, B = 1,... Z = 25, Space = 26
- Ordnung bestimmt Größe des Arrays
- ‚BE': ‚B' = 1, ‚E' = 4 → $1*27^0+4*27^1$

Training (3 Sprachen)

- Pro Sprache ein Modell mit gespeicherten Wahrscheinlichkeiten
- 3 Sprachen mit je 100.000 Zeichen trainiert (je mehr Zeichen, desto genauer die Wahrscheinlichkeiten)
- Deutsche Texte aus Projekt Gutenberg (z.B. Faust) entnommen

Erkennung

Erkennungsraten I

- Je länger der angegebene Text, desto höher Erkennungsrate (Wahrscheinlichkeiten repräsentativer)
- Höhere Ordnungen des Modells erhöhen Rechenzeit und Speicherbedarf unverhältnismäßig
- Texte mit Ordnung 2: >90% erkannt

Erkennungsraten II

- Fremdwörter und Fachausdrücke senken die Erkennungsrate (Wahrscheinlichkeiten nicht repräsentativ)
- Werden repräsentative, typische deutsche Texte (z.B. Faust) zum Trainieren verwendet, werden bereits bei 5000 Zeichen >80% erkannt

Demonstration

- Deutscher Text
- Englischer Text
- Italienischer Text

- Textgenerierung

Shakespeare's Monkeys

Danke für Eure Aufmerksamkeit!