Hidden Markov Models in Anwendungen

Dr. Vera Demberg

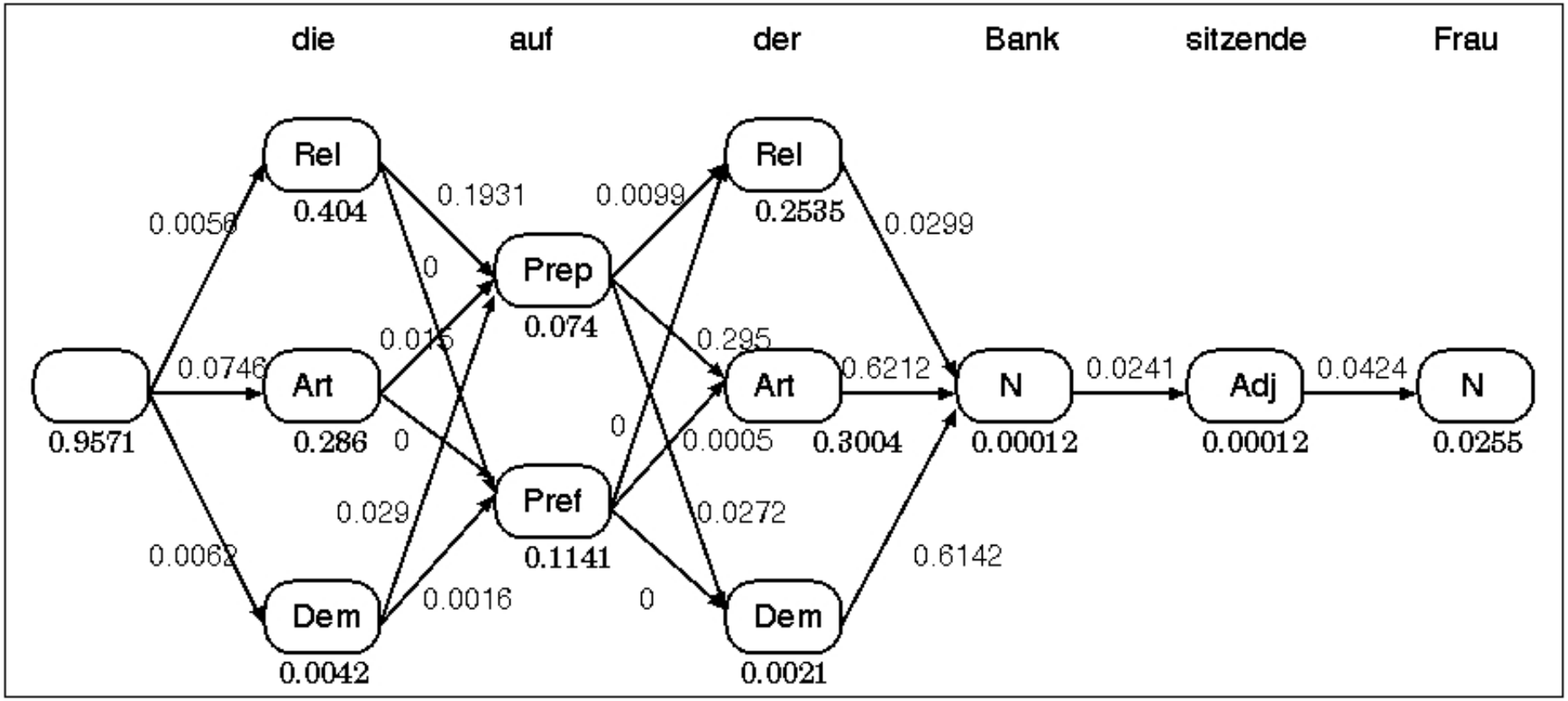
Universität des Saarlandes

31. Mai 2012

Table of Contents

Wiederholung: Was ist ein Hidden Markov Model?

Beispiel HMM für POS tagging



(Grafik entnommen von Skript Dr. Martin Volk, Zürich)

Hidden Markov Model vs. Markov Model

Terminologie: Hidden Markov Model vs. Markov Model

Wenn es um POS tagging geht sprechen wir manchmal über Hidden Markov Models und manchmal über Markov Models – was denn nun?

- Markov Models in Training auf POS-tag annotierten Daten: Die POS-tag Schicht ist sichtbar, daher nicht "hidden".
- Hidden Markov Model in Tagging: Wir behandeln die POS tags als versteckte Variable, da wir die Modelle auf nicht POS tag annotierten Texten laufen lassen.

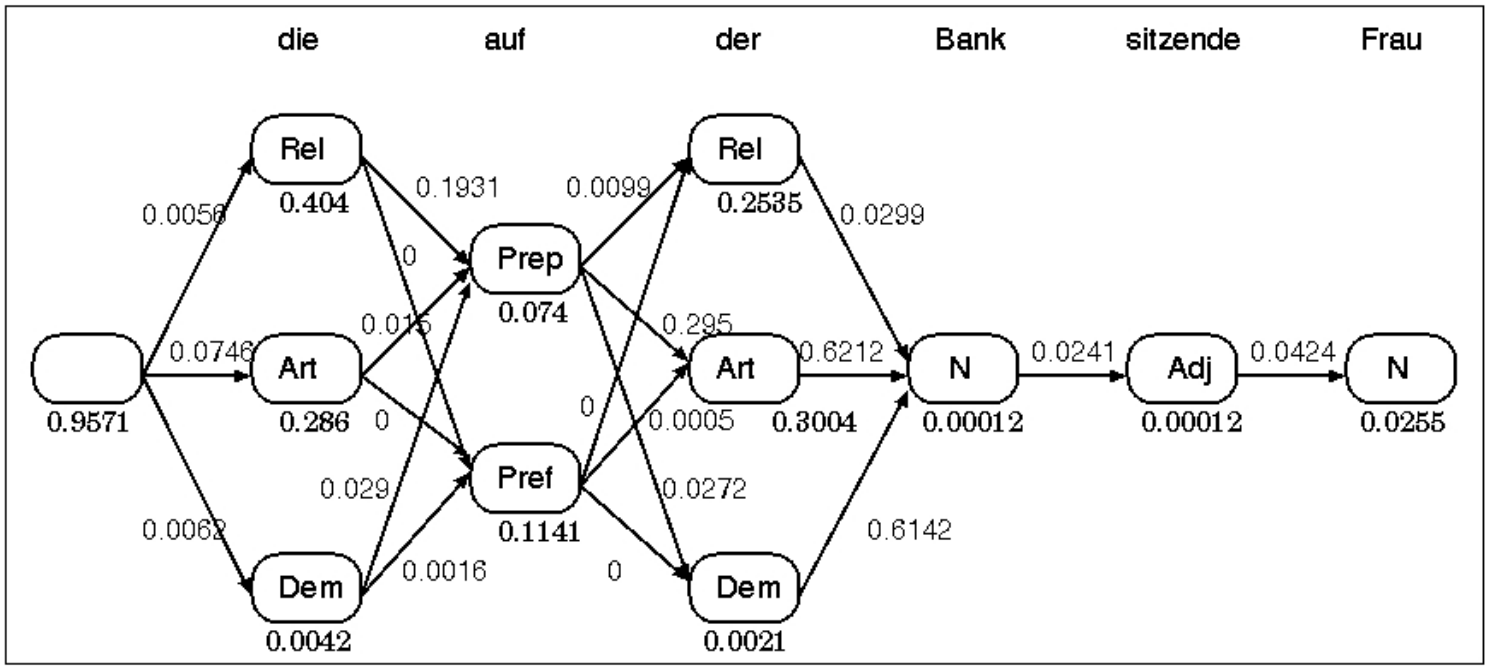
Anwendungen von Hidden Markov Models

Für welche Aufgaben werden Hidden Markov Modelle verwendet?

- Part-of-Speech Tagging
- Spracherkennung
- Text-to-speech / grapheme-to-phoneme conversion
- Word Alignment in Statistical Machine Translation
- Topic Segmentation
- Information Retrieval
- Bedeutungsdisambiguierung
- Handwriting / Character Recognition

POS tagging

- Worte sind beobachtete Ereignisse
- Wortarten sind die versteckten unterliegenden Ereignisse
- (mehr Details zu state-of-the-art Taggern im im 2. Teil dieser Vorlesung)



Spracherkennung

- Beobachtet?
- Verborgen?

Aber wie kriegen wir das Sprachsignal in den HMM?

Vgl. Gales, M. and Young, S. (2008). *The application of hidden Markov models in speech recognition*.

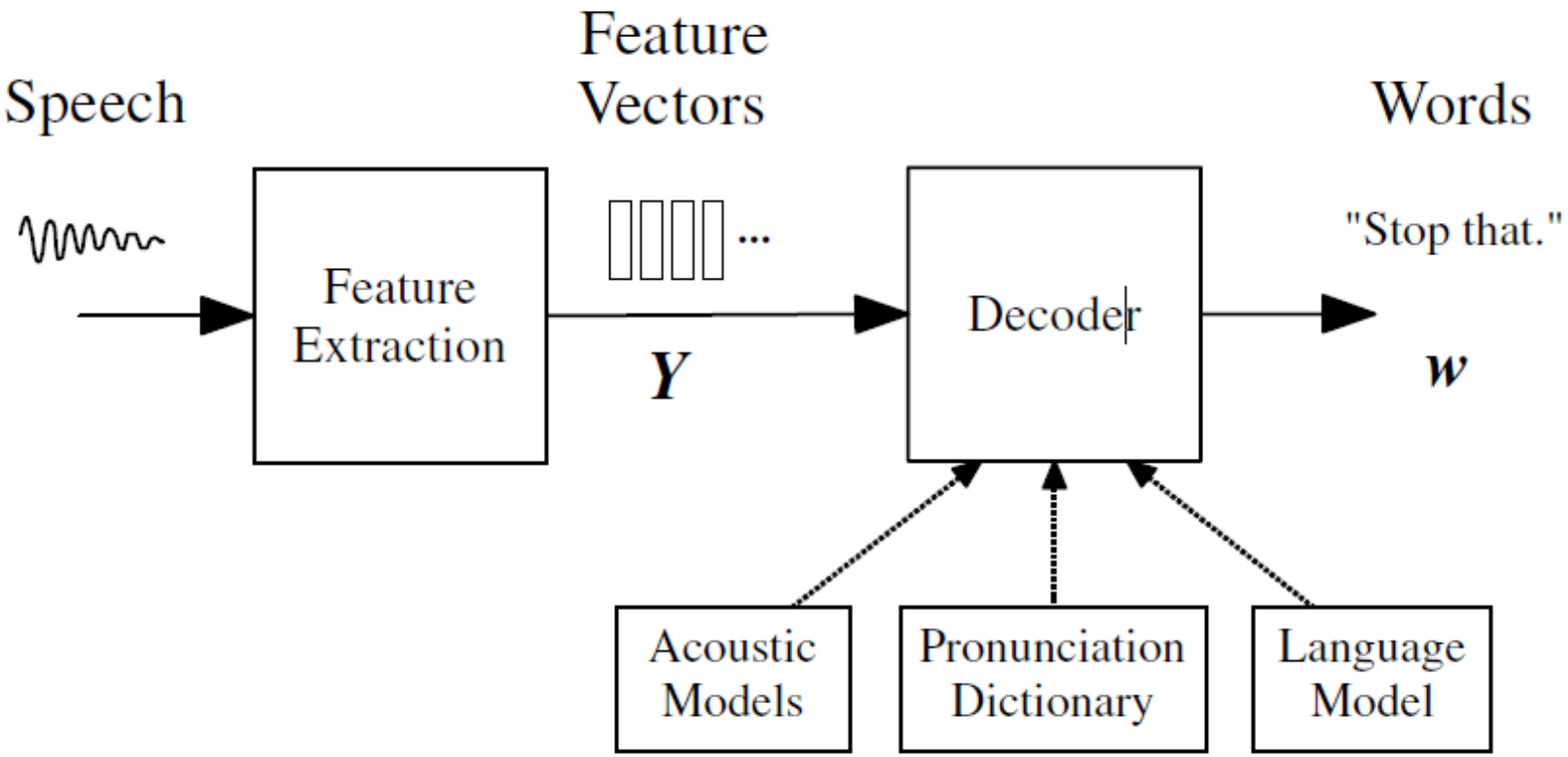
Spracherkennung

- Beobachtet? **akustisches Sprachsignal**
- Verborgen? **Woerter**

Aber wie kriegen wir das Sprachsignal in den HMM?

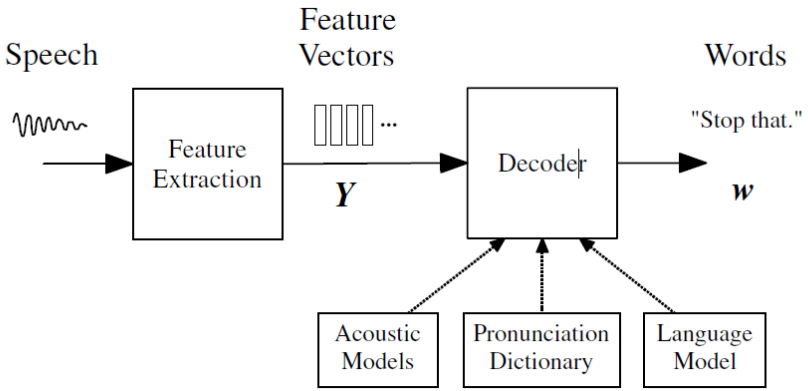
Vgl. Gales, M. and Young, S. (2008). *The application of hidden Markov models in speech recognition*.

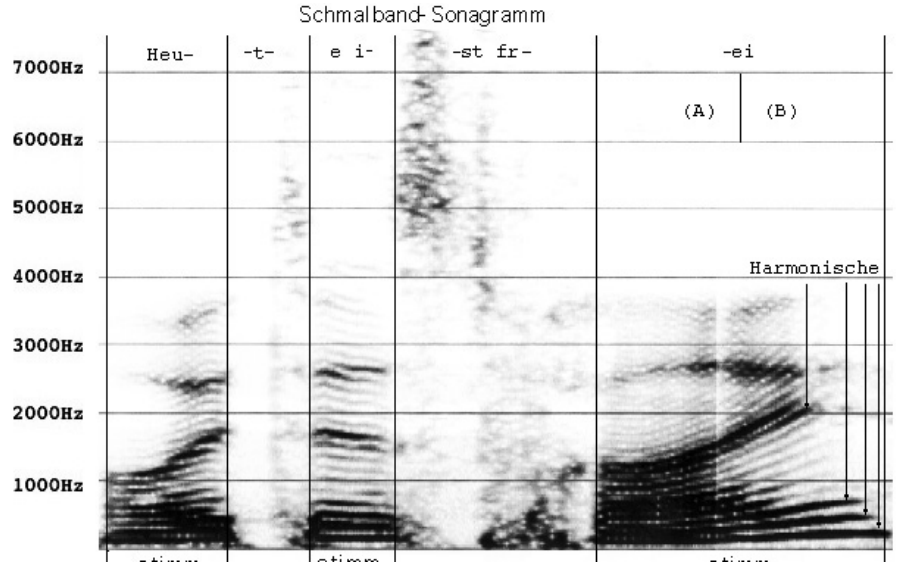
Spracherkennung



(aus Gales, M. and Young, S., 2008)

Spracherkennung





Features aus Sprachsignal:

- zerscheiden in 10ms-Schnipsel, Analyse zur spektralen Information

Modelle:

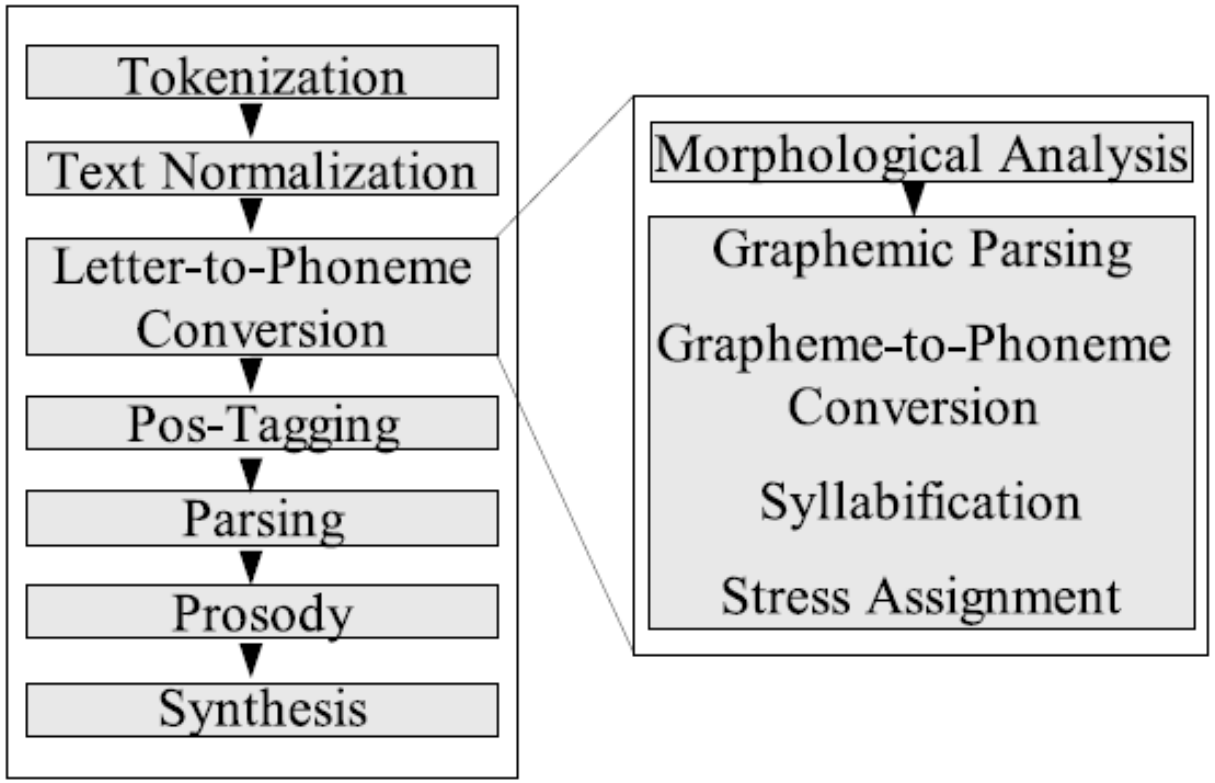
- Akustisches Model: Welche spektrale Charakteristik gehoert zu welchem Laut? (/a/, /i/, /s/, /f/), stark kontextabhaengig!
- Aussprachewörterbuch: Wie kommen Phoneme in Wörtern vor?
- Sprachmodell? In was für Sequenzen kommt ein Wort normalerweise vor?

Sprachgenerierung aus Text

- Beobachtet?
- Verborgen?

Sprachgenerierung aus Text

- Beobachtet?
- Verborgen?



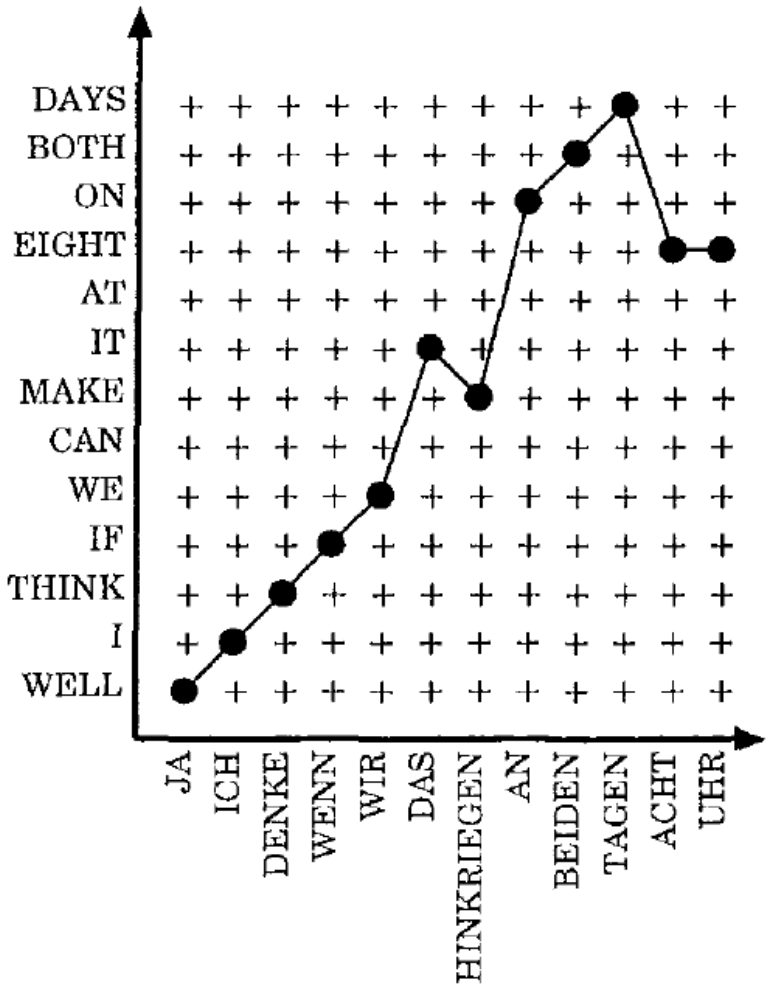
Text-to-speech

HMMs können für unterschiedliche Teilaufgaben benutzt werden:

- Silbenanalyse
  Beobachtet: Buchstabenfolge (+ Morphemgrenzen)
  Verborgen: Silbengrenze
- Wortbetonung
  Beobachtet: Buchstabe + Silbengrenze
  Verborgen: Betonte Silbe oder nicht?
- Graphem-zu-Phonem Konvertierung
  Beobachtet: Buchstabe + Silbengrenze + Betonung
  Verborgen: phoneme (rough $\rightarrow$ /r u f/)

Word Alignment in Maschineller Übersetzung

- Beobachtet: Wortfolgen in verschiedenen Sprachen
- Verborgen: Alignierung der Wortfolgen
- Intuition: Wortverteilung im Satz nicht zufällig sondern in Clustern
- Wahrscheinlichkeiten
  - Alignierungsposition gegeben Alignierungsposition des vorigen Worts
  - Übersetzungswahrscheinlichkeit gegeben Alignierungsposition



(vgl. Vogel, Hey, Tillmann: "HMM - Based Word Alignment in Statistical Translation")

Text Segmentierung nach Themen

Beobachtet?
Verborgen?

Wie viele Woerter in Betracht ziehen fuer jede Themen-Entscheidung?

- kleines Fenster: nicht genug Inhaltsworte um gute Entscheidung zu treffen
- grosses Fenster: schlechte Granularitaet.
- Einfacher HMM beste Ergebnisse bei ca. 100 Woertern.

(vgl. Blei und Moreno "Topic Segmentation wit an Aspect Hidden Markov Model")

Text Segmentierung nach Themen

Beobachtet? Woerter
Verborgen? Themengrenzen

Wie viele Woerter in Betracht ziehen fuer jede Themen-Entscheidung?

- kleines Fenster: nicht genug Inhaltsworte um gute Entscheidung zu treffen
- grosses Fenster: schlechte Granularitaet.
- Einfacher HMM beste Ergebnisse bei ca. 100 Woertern.

(vgl. Blei und Moreno "Topic Segmentation wit an Aspect Hidden Markov Model")

Table of Contents

HMMs – das Kleingedruckte

In der Praxis gibt es noch eine Reihe zusaetzlicher, wichtiger Details, die implementiert werden muessen, damit ein Hidden Markov Model gut funktioniert.

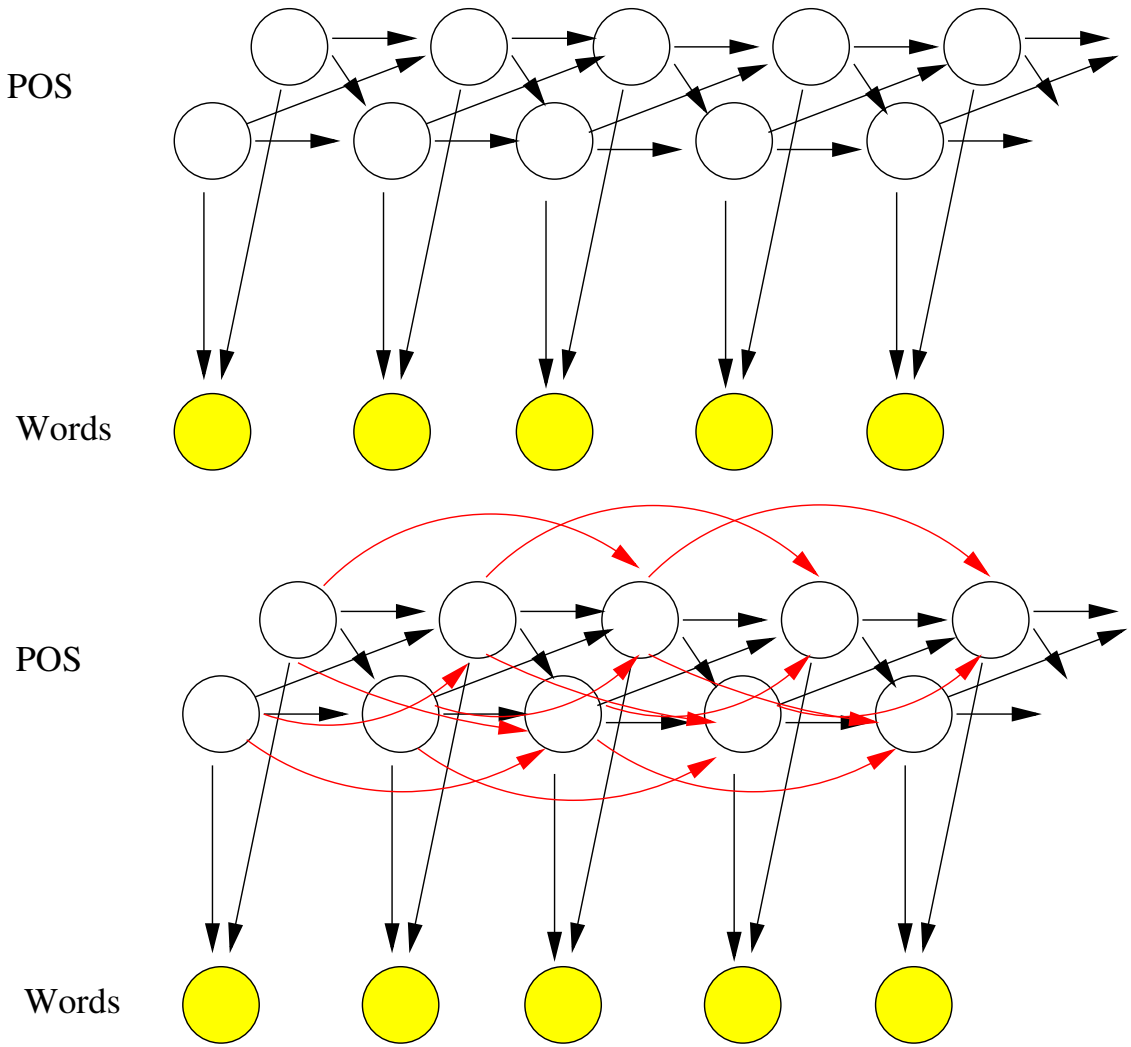
- Reicht es wirklich aus, die Wahrscheinlichkeit eines POS tags nur bezogen auf das vorausgehende POS tag abzuschätzen?
- Was machen, wenn wir ein Wort (oder eine POS tag-Folge) nicht in den Trainingsdaten gesehen haben?

Higher-order Markov Models

1st order HMM
$P(POS_i|POS_{i-1})$

2nd order HMM
$P(POS_i|POS_{i-1}, POS_{i-2})$

POS-tagging HMMs verwenden normalerweise trigrams (2nd order).



Wahrscheinlichkeitsabschätzung

Benötigte Wahrscheinlichkeiten für 2nd order HMM:

- $P(POS_i | POS_{i-2}, POS_{i-1})$
- $P(word_i | POS_i)$

Beispiel

dazu brauche wir z.B. die folgenden Frequenzen:

$\frac{freq(DT,NN,ADV)}{freq(NN,ADV)}$ und $\frac{freq(often,ADV)}{freq(ADV)}$

- keine Chance alle Wörter im Training zu sehen
- viele POS-tag Sequenzen nicht gesehen
- ungesehenen Ereignissen wollen wir aber nicht Wahrscheinlichkeit 0 zuweisen.

Unbekannte Wörter

- Wahrscheinlichkeitsmasse reservieren
- z.B. alle seltenen Woerter im Trainingskorpus durch spezielle Markierung (OOV = out of vocabulary) ersetzen und neue unbekannte Wörter als OOV abschätzen
- Bessere Idee: Wortart anhand von Wortendung und Groß- / Kleinschreibung raten
  - -able $\rightarrow$ ADJ
  - -ness $\rightarrow$ NN
  - -ment $\rightarrow$ NN
  - -ing $\rightarrow$ VB
- selbst mit Groß-/Kleinschreibung und Endungen können bekannte Worte leichter getaggt werden:
  - F-score bekannte Worte (deutsch): 97.7%
  - F-score unbekannte Worte (deutsch): 89.0%

Stichworte: **Smoothing und Backoff** – mehr dazu in ein paar Wochen.

Unbekannte Wörter

- Wahrscheinlichkeitsmasse reservieren
- z.B. alle seltenen Woerter im Trainingskorpus durch spezielle Markierung (OOV = out of vocabulary) ersetzen und neue unbekannte Wörter als OOV abschätzen
- Bessere Idee: Wortart anhand von Wortendung und Groß- / Kleinschreibung raten
  - -able $\rightarrow$ ADJ
  - -ness $\rightarrow$ NN
  - -ment $\rightarrow$ NN
  - -ing $\rightarrow$ VB
- selbst mit Groß-/Kleinschreibung und Endungen können bekannte Worte leichter getaggt werden:
  - F-score bekannte Worte (deutsch): 97.7%
  - F-score unbekannte Worte (deutsch): 89.0%

Stichworte: **Smoothing und Backoff** – mehr dazu in ein paar Wochen.

Unbekannte POS tag Sequenzen

Beispiel: Ungesehene POS-tag Sequenz

`PTKNEG $, PPER`
(wie in *Vermischen reicht ihnen* **nicht, sie** *nutzen die Elektronik*)

- Wahrscheinlichkeit durch kleineren Kontext abschätzen besser als gar nichts sagen können.
- wenn $P(PPER|PTKNEG\ \$,)$ nicht gesehen, nehme $P(PPER|\$,)$
- wenn $P(PPER|\$,)$ nicht gesehen, nehme $P(PPER)$
- damit es eine Wahrscheinlichkeitsverteilung wird: **Interpolation**

$$P(t_3|t_1,t_2) = \lambda_1 P(t_3) + \lambda_2 P(t_3|t_2) + \lambda_3 P(t_3|t_1,t_2)$$

Auch hierauf kommen wir in späteren Vorlesungen noch zurück.

Unbekannte POS tag Sequenzen

Beispiel: Ungesehene POS-tag Sequenz

`PTKNEG $, PPER`
(wie in *Vermischen reicht ihnen* **nicht, sie** *nutzen die Elektronik*)

- Wahrscheinlichkeit durch kleineren Kontext abschätzen besser als gar nichts sagen können.
- wenn $P(PPER|PTKNEG\ \$,)$ nicht gesehen, nehme $P(PPER|\$,)$
- wenn $P(PPER|\$,)$ nicht gesehen, nehme $P(PPER)$
- damit es eine Wahrscheinlichkeitsverteilung wird: **Interpolation**

$$P(t_3|t_1,t_2) = \lambda_1 P(t_3) + \lambda_2 P(t_3|t_2) + \lambda_3 P(t_3|t_1,t_2)$$

Auch hierauf kommen wir in späteren Vorlesungen noch zurück.

Flexible Kontexte

Eine alternative Möglichkeit ist die Nutzung von unterschiedlich langen Kontexten.

- Warum eigentlich nur Trigramme? Manchmal hilft es vielleicht, wenn wir ein Tag das noch weiter weg ist, kennen.

Beispiel: Lange Abhängigkeit

Peter **rechnet** heute sein Geld **nach**.

$$P(APPR|PPOSAT, NN) > P(PTKVZ|PPOSAT, NN)$$

(vgl. Peter bringt heute *sein Geld nach* Hause)

Würde der HMM Tagger dies falsch taggen?

- Idee: flexible Kontextlänge, nützliche Länge kann gelernt werden.
- Entscheidungsbaum: lernt, wie viel Kontext für welche Entscheidung nötig ist.

Entscheidungsbäume besprechen wir Anfang Juli.

Flexible Kontexte

Eine alternative Möglichkeit ist die Nutzung von unterschiedlich langen Kontexten.

- Warum eigentlich nur Trigramme? Manchmal hilft es vielleicht, wenn wir ein Tag das noch weiter weg ist, kennen.

Beispiel: Lange Abhängigkeit

Peter **rechnet** heute sein Geld **nach**.

$$P(APPR|PPOSAT, NN) > P(PTKVZ|PPOSAT, NN)$$

(vgl. Peter bringt heute *sein Geld nach* Hause)
Würde der HMM Tagger dies falsch taggen?

- Idee: flexible Kontextlänge, nützliche Länge kann gelernt werden.
- Entscheidungsbaum: lernt, wie viel Kontext für welche Entscheidung nötig ist.

Entscheidungsbäume besprechen wir Anfang Juli.

Flexible Kontexte

Eine alternative Möglichkeit ist die Nutzung von unterschiedlich langen Kontexten.

- Warum eigentlich nur Trigramme? Manchmal hilft es vielleicht, wenn wir ein Tag das noch weiter weg ist, kennen.

Beispiel: Lange Abhängigkeit

VVINF / VVFIN (infinites vs. finites Verb):
Wir **wollen** 1994 die modernste Autofabrik in Deutschland **eröffnen**.
Was Amerikaner fuer die beste Produktionsmethode **halten**, ...

- Idee: flexible Kontextlänge, nützliche Länge kann gelernt werden.
- Entscheidungsbaum: lernt, wie viel Kontext für welche Entscheidung nötig ist.

Entscheidungsbäume besprechen wir Anfang Juli.

Flexible Kontexte

Eine alternative Möglichkeit ist die Nutzung von unterschiedlich langen Kontexten.

- Warum eigentlich nur Trigramme? Manchmal hilft es vielleicht, wenn wir ein Tag das noch weiter weg ist, kennen.

Beispiel: Lange Abhängigkeit

VVINF / VVFIN (infinites vs. finites Verb):
Wir **wollen** 1994 die modernste Autofabrik in Deutschland **eröffnen**.
Was Amerikaner fuer die beste Produktionsmethode **halten**, ...

- Idee: flexible Kontextlänge, nützliche Länge kann gelernt werden.
- Entscheidungsbaum: lernt, wie viel Kontext für welche Entscheidung nötig ist.

Entscheidungsbäume besprechen wir Anfang Juli.

Training ohne annotierte Daten

Eine weitere Möglichkeit: im Training wirklich "verborgene" POS tags verwenden.

- Prinzipielle Idee: weise allen Wörtern alle möglichen POS tags und zufällige Wahrscheinlichkeiten zu und beobachte, welche Zuweisungen am besten generalisieren[1].
- Verfahren: Expectation maximisation (EM) für HMMs: forward-backward Algorithmus
- In der Praxis liegt zumindest Information aus einem Lexikon vor, welche Wortarten jedes Wort haben kann.
- Wenn Information ueber moegliche Wortarten fuer jedes Wort vorliegt, ist der Suchraum sehr viel kleiner.

[1] Wenn man gar keine Trainingsdaten hat, kann man nur unbennante POS tags lernen und nachtraeglich labeln.

Table of Contents

HMMs für POS tagging in verschiedenen Sprachen

- ein Hidden Markov Model ist im Prinzip sprachunabhängi
- Tagging Schwierigkeit hängt u.a. von der Größe des Tagsets ab
- benutzte Features (Groß/Kleinschreibung, Wortendungen) sind sprachabhängig
- andere Sprachen haben u.U. weniger Resources oder ein grösseres Spärlichkeitsproblem (z.B. durch mehr unterschiedliche Wortformen, freiere Wortstellung)

Features für deutsches POS tagging

- Englisch: Suffixliste für unbekannte Wörter
- Deutsch: auch Präfixe können informativ sein bzgl. der Wortart (z.B. “ver-”, “ent-”, “ge-”)
- Großschreibung (nicht am Satzanfang) enthält Information bzgl. der Wortart.
- Spärlichkeitsproblem für ungesehene Wörter gemindert durch automatisches Tagging von unannotierten Texten, die diese Wörter enthalten.
- Fscore mit Standardsettings auf Deutsch: 96.05%
  mit Anpassung: 97.53%
  (beide Ergebnisse beziehen sich auf Helmut Schmid’s TreeTagger, für Details siehe Schmid 1995)

Zusammenfassung

- Hidden Markov Modelle sind auf sehr viele verschiedene Probleme in der CL anwendbar.
- In der Praxis: oft higher order HMMs (z.B. Trigramme statt Bigramme)
- Flexible Kontextgröße mit Entscheidungsbaum
- Behandlung von unbekannten Ereignissen, Smoothing, Backoff
- Ergebnisse können verbessert werden durch Zufügung linguistischer Information

Weitere Vorlesungen und Übungen

- von mir gehalten ab 21. Juni
- ich werde Sie bitten, bestimmte Kapitel aus Manning & Schütze (*Foundations of Statistical Natural Language Processing*) **vor** den Veranstaltungen zu lesen.
- Genauere Hinweise zeitnäher per Email.
- Haben Sie sich auf HISPOS angemeldet? Nur noch Zeit bis 4. Juni.