Hörsaalübung 15.07.2020

1. **Syntax**
   Gegeben sei der folgende Satz (i):

   (i) Unter dem Augustusplatz liegt eine große Tiefgarage.

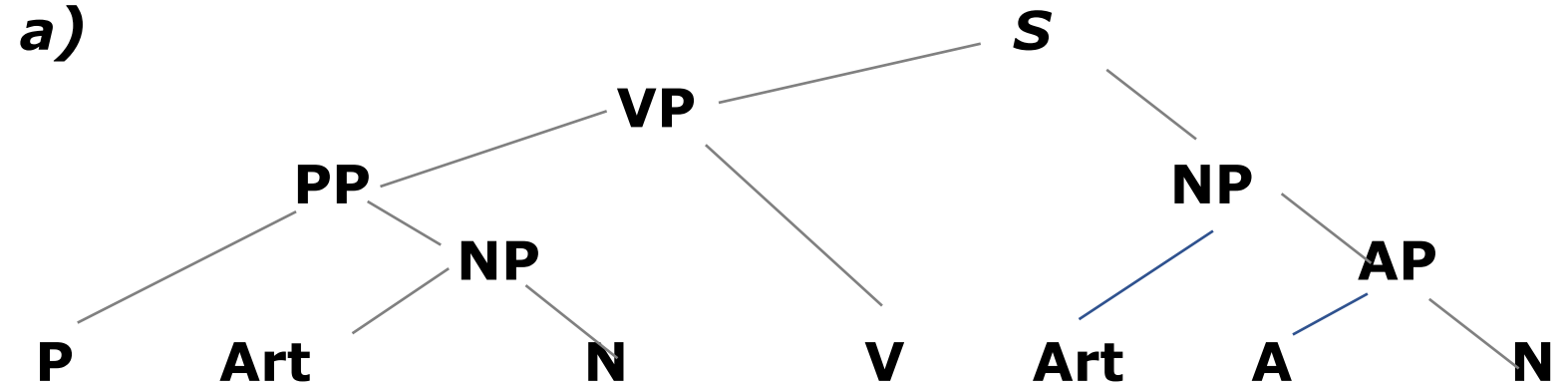
   (a) Analysieren Sie den Satz (i) mit Hilfe einer **kontextfreien Phrasenstruktur-grammatik.**

   (b) Was versteht man unter **Topikalisierung**? Erläutern Sie den Begriff am Beispiel des Satzes (i).

   (c) Skizzieren Sie, wie das Problem der Topikalisierung für das Deutsche mit Hilfe von **Transformationen** im Rahmen einer Phrasenstrukturgrammatik allgemein gelöst werden könnte.

   (d) Ergänzen Sie diesen Satz um eine syntaktische Konstruktion, die im allgemeinen Fall **nicht** mit Hilfe einer *regulären Grammatik* analysiert werden kann.

a)



Unter dem Augustusplatz liegt eine große Tiefgarage

c)

Strukturbeschreibung: NP V PP
1 2 3

Strukturveränderung: 1/NP 2/V 3/PP → 3 2 1

d) NP → Art N PP
PP → P NP

2. Zipf’sches Gesetz

(a) Formulieren Sie das sog. **Zipf’sche Gesetz** bezogen auf die **Sprache** Deutsch. Welchen Wert hat die Zipf’sche Konstante fürs **Deutsche**?

(b) Wie kann die Zipfverteilung von Wortformen für die automatische Extraktion relevanter Terme einer Fachsprache verwendet werden?

(c) In welchen Bereichen weicht bei der Analyse realer Daten der tatsächliche Funktionsgraf in einer doppelt logarithmischen Skalierung meist von dem nach dem Zipfschen Gesetz vorhergesagten ab? Wie lassen sich diese Abweichungen erklären?

(d) Im Verhältnis zur Anzahl der **tokens** in einem Text der Sprache Deutsch: Wie hoch wird vermutlich die Anzahl unterschiedlicher **types** sein?

3. Probabilistisches Sprachmodell – Teil 1

(a) Am lokalen Ende eines fehlerbehafteten, binären Kommunikationskanals Y werde eine “1” beobachtet: Y=1. Von der Quelle am fernen Ende des Kanals ist “a priori” bekannt, dass sie Einsen mit der Wahrscheinlichkeit p(X=1)=0.20 sendet. Der Kanal ist durch die Übertragungsmatrix bedingter Wahrscheinlichkeiten

$$p(Y=1|X=1)=0.90 \quad p(Y=1|X=0)=0.25$$
$$p(Y=0|X=1)=0.10 \quad p(Y=0|X=0)=0.75$$

vollständig charakterisiert. Welches Symbol X*=? hat die Quelle höchstwahrscheinlich gesendet? Berechnen Sie mit Hilfe des Satzes von Bayes.

Wir berechnen mit dem Satz von Bayes:

X=0: p(X=0|Y=1) = p(Y=1|X=0)*p(X=0)* = 1/4 * 4/5 = 1/5
X=1: p(X=1|Y=1) = p(Y=1|X=1)*p(X=1)* = 0.9 * 1/5 = 0.9 * 1/5

Lösung: X = 0 da 0.9 < 1 und somit p(X=0|Y=1) > p(X=1|Y=1)

4. Probabilistisches Sprachodell – Teil 2

(a) Gegeben seien die folgenden Sätze einer Sprache L:

S1 *Der Tenor singt ein Lied.*
S2 *Der Tenor trinkt ein Bier.*
S3 *Die Sopranistin trinkt ein Wasser*.

Berechnen Sie die **Wahrscheinlichkeiten** des Auftretens folgender Wortfolgen nach dem Trigramm-Modell (ohne Satzanfangs- und –endemarkierungen):

S4 *Der Tenor trinkt ein Bier*.
S5 *Der Tenor trinkt ein Wasser.*
S6 *Der Tenor trinkt ein Lied.*

5. Strukturalistischer Ansatz

(a) Definieren und erläutern Sie die folgenden Begriffe:
- Wortform
- Wort.

(b) Definieren und erläutern Sie den Begriff des
- lokalen und
- globalen Kontextes einer Wortform.

(c) Definieren und erläutern Sie den Begriff
- einer syntagmatischen und
- einer paradigmatischen Relation zwischen Wortformen einer Sprache.

(d) Definieren Sie den Begriff
- der Kookkurrenzen einer Wortform und
- geben Sie die Formel für die Berechnung von Kookkurrenzen mit dem Dice-Koeffizienten an.

5. Strukturalistischer Ansatz

(e) Definieren und erläutern Sie die folgenden Begriffe:
- Wortform: **diambiguiertes Token (flektierte Form eines Wortes)**
- Wort: **Äquivalenzklasse zusammengehöriger Wortformen**

(f) Definieren und erläutern Sie den Begriff des lokalen und globalen Kontexts:
- lokal: **Umgebung einer WF ohne diese**
- global: **Alle lokalen Kontexte einer WF**

(g) Definieren und erläutern Sie die Begriffe syntagmatisch und paradigmatisch:
- syntagmatisch: **gemeinsamen Auftreten**
- paradigmatisch: **gemeinsamer Kontext**

(h) Definieren Sie den Begriff
- der Kookkurrenzen einer Wortform: **statistisch signifikantes gemeinsames Auftreten**
- **Formel Dice:** $2\ N_{a,b} / N_a + N_b$

6. Markov Modelle und Tagging

(a) Geben Sie die Definition eines Hidden Markov Modells,
(b) Was sind die Markov-Eigenschaften?
(c) Beschreiben Sie, wie natürliche Sprache durch ein Markov Modell beschrieben werden kann.
(d) Geben Sie die beiden Markov-Ketten in Abb.1 an:

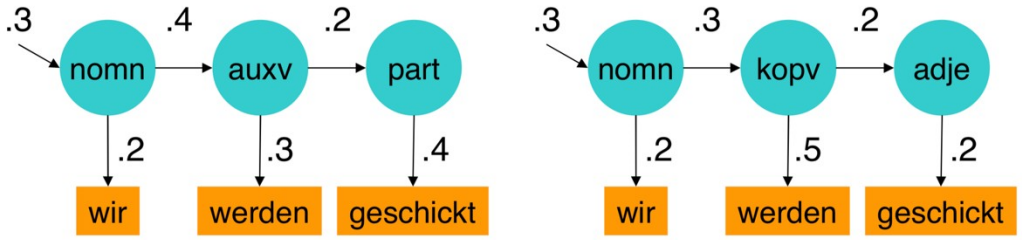


Abb.1

(e) Welcher Prozess wird in den beiden obigen Markov-Ketten ausgeführt ?

6. Markov Modelle und Tagging

(a) Was sind die Markov-Eigenschaften?

Begrenzter Horizont
der Wert von X_t hängt nur vom Vorgängerzustand X_{t-1} ab

Zeitinvarianz
der Wert des Folgesymbols hängt nicht vom Zeitpunkt t ab, i.e.,Übergangs- und Emissionswahrscheinlichkeiten bleiben über die Zeit konstant

(b) Wie kann natürliche Sprache durch ein Markov Modell beschrieben werden?,

Sprache ist eine lineare Sequenz von Symbolen.
Zufallsvariablen sind sprachliche Einheiten (z.B. Buchstaben, Wörter).
Diese sind nicht unabhängig, da nicht auf jedes Wort jedes beliebige ande re Wort folgen kann.

(c) Geben Sie die beiden Ketten eines HMM in Abb.1 an:

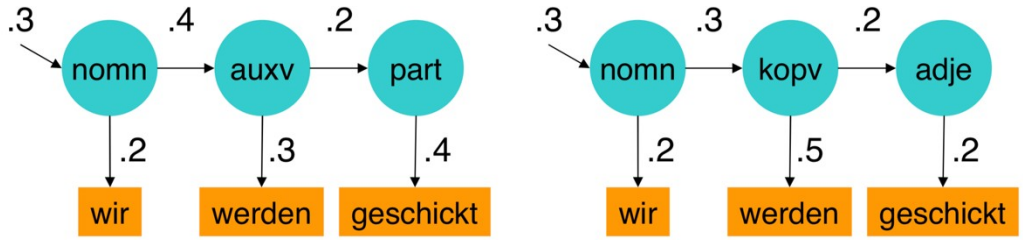


Abb.1

(i) p(start) * p(nomn | start) * p(auxv | nomn) * p(part | auxv) * p(nomn: „wir“) * p(auxv: „werden“) * p(part: „geschickt“)

(ii) p(start) * p(nomn | start) * p(kopv | nomn) * p(adje | kopv) * p(nomn: „wir“) * p(kopv: „werden“) * p(adje: „geschickt“)

(d) Als welchen Prozess könnte man diese Ketten interpretieren?

POS-Tagging, beobachtbar ist nur die Wortfolge *wir werden geschickt*, und es gibt zwei mögliche Ketten von Zuständen (POS-Tags), präferiert wird in der Regel die mit der höchsten Wahrscheinlichkeit.

(e) Geben Sie die Definition eines Hidden Markov Modells (HMM) an.

ein HMM ist ein Quadrupel (z^0, Z, S, E)
wobei z^0 der Startzustand ist, Z ist eine Menge von Zuständen, S ist Alphabet von Eingabe/Ausgabesymbolen und E ist eine Menge eine Menge von Kanten / Übergängen.

HMM: Quadrupel (z^0, Z, S, E)
Z: Zustandsmenge
Z= $\{z_1, \ldots z_n\}$ z^0: Startzustand
S: Alphabet von Ein/Ausgabesymbolen S= $\{s_1, s_2, \ldots, s_m\}$
E: Menge von Kanten bzw. Übergängen E= $\{e_1, e_2, \ldots, e_k\}$
$e_i = (z_i, z_j, s_k, p)$
E= $\{e_1, e_2, \ldots, e_k\}$
$e_i = (z_i, z_j, s_k, p)$
z_i Anfangszustand

z_j Zielzustand

s_k Symbol

p Übergangswahrscheinlichkeit

Mehrere Übergänge von einem Zustand aus mit demselben Symbol sind erlaubt. Deshalb kann die zeitliche Kette der Zustände nicht Mehrere Übergänge von einem Zustand aus mit demselben Symbol durch die Kette der Symbole bestimmt werden – „*hidden*“.

7. Topic Modelle

(a) Was ist die Idee generativer Dokumenten- und Topikmodelle?,

(b) was ist das Ziel der Latent Semantic Analysis (LSA),

(c) LSA: was kann man in /an den durch Dekomposition erhaltenen Teilmatrizen der (Ausgangs) Term-Dokument-Matrix ablesen?,

(d) ist Probabilistic Semantic Indexing generativ? Bitte erläutern,

(e) ebenso: ist Latent Dirichlet Allocation (LDA) generativ?,

(f) wie wirkt sich in LDA das ‚Tuning' des alpha-Parameters aus?

7. Topic Modelle

(a) Was ist die Idee generativer Dokumenten- und Topikmodelle?,

Ein Dokument ist “charakterisiert” durch eine spezifische Verteilung von latenten (nicht beobachtaren) Topiks. und jedes Topik ist eine spezifische Verteilung von Wörtern. Die Reihenfolge der Wörter jedoch ist unerheblich:

Bag-of-word assumption
Reihenfolge der Wörter wird nicht berücksichtigt
Ein Dokument entspricht einem „Sack“ voller Wörter
Auch: Korpus entspricht „Sack“ voller Dokumente
Für jedes Wort (type) wird Frequenz gespeichert

Annahme:
Information über Art und Anzahl von Wörtern reichen aus um Rück schlüsse auf die Struktur von Text zu ziehen

(b) was ist das Ziel der Latent Semantic Analysis (LSA),

Dimensionsreduktion der Term-Dokument-Ausgangsmatrix (welche Dokumente sind einander semantisch ähnlich?).

(c) LSA: was kann man in /an den durch Dekomposition erhaltenen Teilmatrizen der (Ausgangs) Term-Dokument-Matrix ablesen?,

Grundlage bildet eine Wort-Dokument Frequenzmatrix **C**
diese wird per Singulärwertzerlegung in drei Matrizen **U**, **D** und $\mathbf{V}^T$ zerlegt
alle bis auf n höchsten Singulärwerte werden auf 0 gesetzt (Matrix **D**)
ursprüngliche Matrix wird rekonstruiert (hat nun geringeren Rang).

Dokumente sind durch Zeilenvektoren in V repräsentiert, und Dokumentenähnlichkeit ergibt sich aus dem Vergleich von Zeilen in der Matrix VD.

Wörter werden repräsentiert durch Zeilenvektoren in U , und
Wortähnlichkeit wird bestimmt durch die Zeilenähnlichkeit in UD.

U basiert auf den Eigenwerten und Eigenvektoren von $\mathbf{C}\mathbf{C}^{\mathbf{T}}$ und schließlich auf dem Gram-Schmidt Orthonormalisierungsprozess.

$\mathbf{V}^{\mathbf{T}}$ basiert auf $\mathbf{C}^{\mathbf{T}}\mathbf{C}$ und wird analog zu **U** berechnet.

Die Diagonalen in **D** sind die Singulärwerte von **C**, geordnet nach Größe. Zur Dimensionsreduktion können z.B. die drei größten Eigenwerte genommen wer den.

(d) Ist Probabilistic Latent Semantic Indexing (pLSI) generativ? Bitte erläutern.

ja: der Prozess ist wie folgt:
Jedes Wort wird aus einer einzigen Klasse / Topik generiert, und verschiedene Wörter in einem Dokument können aus verschiedenen Klassen / Topiks werden. Jedes Dokument ist repräsentiert als eine Mischung dieser Verteilung und kann letztlich auf eine Wahrscheinlichkeitsverteilung einer festen Anzahl von Topiks reduziert werden.
Kurz: wähle ein Dokument *d* mit P(d), wähle eine latente Klasse *z* mit P(z | d) und **generiere** ein Wort *w* mit P(w | z).

$$P(d, w) = P(d)P(w|d),$$
$$P(w|d) = \sum P(w|z)\ P(z|d)$$

pLSI geht von vermischten Verteilungen und einem Modell der latenten Klas sen (Topiks) aus
Ordnet jeder Beobachtung (Term) eine latente (=nicht direkt beobachtbar) Vari able (Klasse / Topik) zu

(e) Ebenso: ist Latent Dirichlet Allocation (LDA) generativ?

Ja

Dokumente werden generiert auf der Basis einer Verteilung von (latenten) Topiks, wobei jeder Topik eine Verteilung über Wörter ist.

Wählen der Topicverteilungen $\theta \sim Dir(\alpha)$
Wählen der Wortverteilungen $\varphi \sim Dir(\beta)$

Für jedes Wort w_n der n Wörter in Dokument d

Wählen von Topik $z_n \sim$ Multinomial (θ_d)

Wählen von w_n aus $P(w_n|\varphi_{zn})$

Im Wesentlichen werden 2 Schritte ausgeführt:
1) p(topic z | document d)
2) p(word w | topic z)

Durch Gibbs-Sampling werden für n Iterationen die Wahrscheinlichkeiten durch Approximationen verändert, bis idealerweise ein stabiler Zustand eintritt / oder lediglich nur noch ein kleine Divergenz vorhanden ist.

(f) Wie wirkt sich in LDA das ‚Tuning' des alpha-Parameters aus?

bei kleinem alpha erhält man wenige, dafür aber deutlich ausgeprägte Topiks in einem Dokument, größeres alpha strebt zu einer immer gleichmäßigeren Verteilung von (vielen) Topiks im Dokument.