Korpusbasierte Sprachverarbeitung

LVA SS 05 am IMKAI

1) **Warum will man stochastische Verarbeitung?**

- Verbesserung von Kompetenz-Grammatiken mit Performanzaspekten der Sprache
- Ist datenorientiert: Korpora beinhalten einen Ausschnitt einer Sprache zusammen mit der Performanzinformation
- effizientere Verarbeitung als der Grammatikansatz
- robuster im Sinne von Fehlertoleranz: Grammatiken zu streng, Fehler werden immer gemacht und der stochastische Ansatz hilft dabei

2) **Wozu dienen korpusbasierte Ansätze in der Sprachverarbeitung?**

Um die Kompetenz- und Performanzinformation in die statistischen Modelle zu integrieren.

- Performance Data: Korpora (Korpus Daten spiegeln den Sprachgebrauch wieder)
- Competence Data: Korpusannotierung (Annotierung leitet die statistische Generalisierung)
- Statistisches Modell: muss so entworfen sein, dass sie die relevanten linguistischen Informationen und die Auftrittshäufigkeit optimal kombiniert

3) **Was sind ihre Vorteile (siehe Kompetenz versus Performanz)? Warum macht man korpusbasierte Sprachverarbeitung in der Computerlinguistik?**

Korpusbasierte Ansätze sind effizienter und robuster gegenüber den Kompetenzgrammatiken. Die Grammatiken sind streng und deuten kleine Fehler nicht mehr als Teil einer erzeugten Sprache -> mit Hilfe der Performanzinformation kann man damit umgehen, wie die Wirklichkeit der Sprache ist -> somit fehlertoleranter. Korpusbasierte Ansätze produzieren akzeptablere Ergebnisse (höhere accuracy). Weiters werden bei den Grammatiken nur ein Teil der Äußerungen betrachtet -> das führt natürlich zu schlechteren Ergebnissen. Mit Hilfe der Performanzinformation können auch sprachliche Ambiguitäten leichter aufgelöst werden.

4) **Was sind die Schwierigkeiten?**

- Sparse data problem: man benötigt relativ große Mengen an Daten, um Lernen zu können
- großer Aufwand einen Korpus zu erstellen (siehe golden standart)
- statistische Methoden müssen empirisch getestet werden, um Ausreißer beim Lernen zu verhindern

5) **Was sind solche Anwendungen? (3-4 Beispiele wie zB Tagger, Parser, Speech recognition)**

Tagger, Parser, Speech recognition, Language translation, Information retrieval, Collocation Identification, Text categorization, Text summarization, Text-to-Speech synthesis

6) Basis der Wahrscheinlichkeitstheorie: Was ist das Wahrscheinlichkeitsmaß?
Das Wahrscheinlichkeitsmaß P ist eine Funktion von Ereignissen im Raum Ω oder einem Subraum von Ω auf eine Menge von reellen Zahlen in [0,1] welche die folgenden Eigenschaften besitzen:

- $0 \leq P(A) \leq 1$ für alle A aus Ω
- $P(\Omega) = 1$
- $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$

7) **Was ist der Bayessche Satz / bedingte Wahrscheinlichkeit?**
Bayessche Formel: $P(T_+ \mid K) = \frac{P(T_+ \wedge K)}{P(K)} = \frac{P(K \mid T_+) \cdot P(T_+)}{P(K)}$
Die bedingte Wahrscheinlichkeit beschreibt die Wahrscheinlichkeit eines Ereignisses unter der Annahme, das ein anderes Ereignis bereits eingetreten ist.

8) **Wieso verwendet man das maximum-likelihood Prinzip?**
Als **Maximum-Likelihood-Schätzung** bezeichnet man in der Statistik eine Parameterschätzung, die nach der Maximum-Likelihood-Methode berechnet wurde. In der englischen Fachliteratur ist die Abkürzung MLE (*maximum likelihood estimate*) dafür sehr verbreitet. Eine Schätzung, bei der Vorwissen in Form einer *a priori* Wahrscheinlichkeit einfließt wird **Maximum-A-Posteriori-Schätzung** (MAP) genannt.

In der korpusbasierten Sprachverarbeitung wird die MLE eingesetzt, weil man von der Grundannahme ausgeht, dass die Ereignisse unabhängig voneinander sind. Somit macht die maximum-likelihood Methode am meisten Sinn, sucht man hiermit doch das Ereignis mit der größten Wahrscheinlichkeit aus.

9) Wie verläuft die Behandlung von ungesehenen Ereignissen?
Unbekannte Wörter könnten über Groß/Kleinschreibung sowie morphologische Information identifiziert werden. So könnte man z.B. Regeln einbauen, welche großgeschriebene Worte, welche nicht am Satzanfang stehen, als Nomen klassifizieren. Auch Verben könnten grundsätzlich über ihre Endung identifiziert werden. Man kann sie aber auch als „unknown“ taggen.

10) Was bedeutet PoS-Tagging?
PoS-Tagging steht für Part of Speech Tagging, mit dieser Methode will man Wortstrings Wortartenstrings zuordnen. Das Tagset (also die Wortkategorien) besteht hierbei meist zwischen 50 und 100 tags.

11) **Wie ist ein HMM definiert?**
Markov Models können verwendet werden, um die Wahrscheinlichkeit einer linearen Abfolge von Ereignissen zu modellieren. Ein MM entspricht einem nichtdeterministischen endlichen Automaten. Die Wahrscheinlichkeit, den aktuellen Knoten zu erreichen, ergibt sich aus der a-priori Wahrscheinlichkeit des Knotens mal dem Produkt der bedingten Wahrscheinlichkeiten der zuvor traversierten Knoten, also formal:

$$P(X_1 = x_{i_1}, \cdots X_t = x_{i_t}) = P(X_1 = x_{i_1}) * P(X_2 = x_{i_2} | X_1 = x_{i_1}) * \cdots * P(X_t = x_{i_t} | X_{t-1} = x_{i_{t-1}})$$

Bei visible Markov Models sind die Zustände bekannt, sie werden zum Training verwendet (annotierter Korpus), bei hidden Markov Models sind sie unbekannt, wie beim Tagging wenn nur eine Wortsequenz beobachtet wird.

12) **Wie funktioniert POS mit HMM (nur Prinzip)?**
HMMs werden beim POS-Tagging eingesetzt, um die Wahrscheinlichkeit einer korrekten Zuordnung zu maximieren. Hierbei finden beim Training der Forward-Backward Algorithmus und beim Tagging der Viterbi-Algorithmus Anwendung.

13) Was sind Parameter in der Verarbeitung, was liest man aus?
Die Parameter sind die Wahrscheinlichkeiten, dass ein Wort einer bestimmten Kategorie angehört. Man liest die Kategorie mit der Maximalwahrscheinlichkeit aus.

14) Welche Korpustypen gibt es? Was sind deren Merkmale, sprich welche Labels sind annotiert?

- Corpus Type I: Text Type

  Wünsche: large samples, große Anzahl Domains der Texte (Zeitung, Lyrik, ...), Domains müssen gekennzeichnet sein

  - Balanced Corpora
    Versch. Text Genres, Größe porportional zur Verteilung eines Text types
    Problem: Darstellung relevanter Information

  - Pyramidal Corpora
    Große Mengen einiger weniger repräsentativer Genres oder wenige Samples vieler unterschiedlicher Genres
    Problem: wenige Daten (sparse data problem)

  - Opportunistic Copora
    Typischer Korpus, „take what you can get“

- Corpus Type II: Kind of Annotation
  Raw: tokenized, cleaned (keine control chars, Markierung von Text Type, headlines und paragraphs) -> SGML markup

  PoS-tagged: tokenized Text der mit der syntaktischen Kategorie auf Wortebene (PoS) annotiert wurde

  - Treebanks
    PoS tagged Text mit syntaktischer Struktur, parse grammar, automatisch geparsed, parse trees werden von Menschen korrigiert, falls parsing fehlschlägt wird word strings notfalls manuell annotiert
  - Linguistically Interpreted Corpora
    penibel annotiert mit versch. Arten von linguistischer Information vielschichtige Repräsentation, Annotation datengetrieben und deklarativ

- Corpus Type III: Usage

  - Training Corpora
    Verwendet fürs Lernen, groß, annotiert
  - Testing Corpora
    Verwendet fürs Evaluieren der statistischen Modelle, klein, Referenzannotation

15) **Warum werden annotierte Korpora anstatt nicht annotierter Korpora zum Lernen von Sprachmodellen benutzt?**

Disambiguieren von word strings -> Parsing, collocation identification, Text categorization/summarization, Text-to-Speech synthesis, ...

16) Wie sieht ein POS Korpus aus?

Ein Korpus beinhaltet:
- PoS Tags
- Phrasal Tags
- Function Tags
- den Wörtern selbst welche durch die Tags annotiert sind

17) Auf welchen Ebenen kann man annotieren?

Man kann auf den folgenden Ebenen annotieren:
- Syntaktische Kategorie der Wortebene
- Wortbedeutung
- Syntaktische Struktur
- Semantische Interpretation
- Morphosyntaktische Features

Sprich es gibt:
- syntaktische Tags
- phrasale Tags
- funktionale Tags
- strukturelle Tags

18) Stochastisches Parsing: Was ist CFG?

Eine kontextfreie Grammatik G ist ein Quadrupel (V_N, V_T, S, R).
V_N ist eine endliche Menge Nichtterminalsymbole
V_T ist eine endliche Menge Terminalsymbole
S ist ein Startsymbol
R ist eine Menge von Produktionsregeln der Form:
X -> ß
Mit X Nichtterminal und ß beliebige Kette aus Terminalen und Nichtterminalen.

Eine CFG ist in Chomsky Normalform, wenn sie die nach folgenden Typen von Regeln aufgebaut ist:
Xi -> Xj Xk
Xi -> w
(X Nonterminale, w Terminal)
Sie ist in erweiterter Chomsky Normalform, wenn sie diese Regeln oder die folgende beinhaltet:
Xi -> Xj

19) **Was ist eine SCFG? Wie wird aus einer CFG eine SCFG?**

Die stochastische kontextfreie Grammatik ist ein Quintupel bestehend aus denselben Mengen und einer Funktion P von R nach [0,1], die den verschiedenen Produktionsregeln Wahrscheinlichkeiten zuweist, wobei die Summe dieser Wahrscheinlichkeiten für gleiches X immer 1 ergeben muss.

Demnach genügt es, die Produktionsregeln mit Wahrscheinlichkeiten zu gewichten (die sich für gleiches X auf 1 aufsummieren müssen), um aus einer CFG eine SCFG zu machen.

20) Was ist ein Baum? Wie kann man Grammatiken lexikalisieren?
Ein **Baum** ist ein azyklischer gerichteter Graph, der genau einen Knoten ohne direkten Vorgänger hat (die **Wurzel**). Den direkten Vorgänger eines Knotens nennen wir **Vater**, einen direkten Nachfolger **Sohn**. Knoten mit demselben Vater sind **Brüder**. Knoten ohne Söhne heißen **Blätter**.

21) **Was ist eine Derivation?**
Eine **Ableitung (auch Derivation)** bezeichnet eine Folge von Produktionen, die anhand einer formalen Grammatik ein Wort einer formalen Sprache erzeugt. Umgekehrt kann ein Wort (der *Eingabetext*) auch anhand der Grammatik zerlegt werden, um diejenige Ableitung zu erhalten, die dieses Wort produziert. Diesen Vorgang nennt man auch parsen.
Wenn bei einer Ableitung immer das am weitesten links stehende Nichtterminalsymbol einer formalen Grammatik ersetzt wird, so spricht man von Linksableitung.

22) **Was sind Parsebäume?**
Ein Parsebaum eines Eingabestrings von Terminalsymbolen, welcher durch eine Kontextfreie Grammatik G erstellt wird, muss die folgenden Charakteristika besitzen:
- Seine Wurzel ist S (das Startsymbol der Grammatik)
- Alle Blattknoten sind Elemente aus V_T (Terminalsymbole)
- Alle inneren Knoten sind Elemente aus V_N (Nichtterminale)
- Ist ein Knoten X Vater von anderen Knoten $X_1,\ldots, X_k$, so existiert in G eine Produktionsregel der Form $X \rightarrow X_1,\ldots, X_k$

23) Was ist eine Dependenzgrammatik?
Die Dependenzgrammatik untersucht die hierarchische Struktur eines Satzes ausgehend von Abhängigkeiten. Dependenz ist Abhängigkeit in dem Sinne, dass ein Wort (das regierte) von einem anderen Wort (das regierende) abhängt.

24) Was tut man um lexikalische Information mit einzubeziehen, d.h. wie kommt man von einem Baum auf Dependenzbestimmungen?
Um lexikalische Information miteinzubeziehen identifiziert man für jeden Vaterknoten ein Head-element. Daraus ergeben sich die Head-modifier Paare der Dependenzgrammatik, welche den Kontext miteinbeziehen.

25) Was sind die Ähnlichkeiten zwischen HMM mit tagging und chunking?
Sowohl PoS-tagging als auch Chunk-tagging basieren auf HMM, ein Tagger kann also sowohl für tagging als auch für chunking verwendet werden und benutzt hierbei nur unterschiedliche Tagsets.

26) Welche Teile benutzt man fürs Lernen einer CFG?

XXX