Skript und Begleitmaterial

Statistische Methoden in der Sprachverarbeitung

Sommer 2004

Stefan Langer

CIS, Universität München

stefan.langer@cis.uni-muenchen.de

Vorbemerkung

Dieses Skript entstand vom Sommersemester 2001 bis zum Sommersemester 2004 im Rahmen des Statistik-Einführungskurses am Centrum für Informations- und Sprachverarbeitung der Universität München. Es basiert teilweise auf Manning/Schützes F*oundations of statistical natural language processing*, teilweise auf anderen im Skript genannten Quellen.

Zahlreiche Studentinnen und Studenten haben durch ihre Anmerkungen zur Korrektur und Verbesserung des Skripts beigetragen.

Im Sommersemester 2004 hat Aleksandra Wasiak als Tutorin für das Seminar, eine Reihe von Fehlern und Inkonsistenzen berichtigt. Das Kapitel zur Wortbedeutungsdisambiguierung stammt in erster Linie von ihr.

Sprechstunde Mo 11-12 und jederzeit nach Vereinbarung, Zi. B105 Tel: 2180 9707

1. Statistik und Computerlinguistik

1.1. Allgemeines

Nachfolgend zur Einführung eine kurze Übersicht über die Rolle der Statistik in der Computerlinguistik - bezogen auf die einzelnen linguistischen Teildisziplinen.

Dabei wird unterschieden zwischen:

- Deskriptiver und explorativer Statistik (Datenbeschreibung)
  - Beschreibung von Daten
  - Darstellung von Daten
  - Entdeckung von Strukturen und Besonderheiten in den Daten
- Analytischer Statistik - induktiver Statistik
  - Schlussfolgerungen aus Daten
  - Einbeziehung der Wahrscheinlichkeitstheorie (Stochastik)

1.2. Statistik in der Syntax

Syntax ist die Beschreibung der Struktur natürlichsprachlicher Syntagmen (Sätze und ihrer Konstituenten). Teilbereiche der Syntax bez. auf die Computerlinguistik sind: Tagging, Parsing, Generierung syntaktisch korrekter Sätze, Erfassung und Erkennung von Mehrwortlexemen.

Deskriptive Statistik

- Welche syntaktischen Kategorien sind wie häufig in einem Text / einem Korpus?
- Welche Kategorienfolgen sind wie häufig? Es könnten sich beispielsweise bei der Analyse eines Korpus mit 10 000 laufenden Wörtern folgende Häufigkeiten für Folgen von zwei Kategorien ergeben:

  DET N: 300

  ADJ N : 32

  V N : 3

...

Ebenso können statistisch erfasst werden:

- Satztypen; Konstruktionstypen;

Induktive Statistik

Welche syntaktische Kategorie ist wahrscheinlich - sollte gewählt werden (Wortartentagging). Im folgenden Beispiel kann die Wortform *irre* prinzipiell eine Verbform sein (von *irren*) oder eine Adjektivform (von *irr*). Neben symbolischen Methoden (grammatischer Analyse), erlauben hier einfache deskriptive Statistiken über wahrscheinliche / unwahrscheinliche Wortartenabfolgen (DET V A ist in beliebigen Trainingskorpora inexistent bis selten, DET A N häufig) die Bestimmung der richtigen Kategorie (hier Adjektiv).

der irre Professor

DET V/A N

Welche syntaktische Analyse ist die richtige?

Statistisches Parsing - Ambiguitätsresolution.

Sie beobachtet den Mann mit dem Fernrohr

(Fernrohr ist Instrument zum sehen vs. Mann, der gesehen wird, hat ein Fernrohr)

Die statistische Beobachtung, dass *Fernrohr* im Zusammenhang mit dem Verb *beobachten* in 90% Teil einer instrumental-PP ist, erlaubt es, hier zu folgern, dass "mit dem Fernrohr" nicht zur NP "den Mann" gehört, sondern Adverbial ist.

1.2.1 Statistik in der Semantik

Die Semantik befasst sich mit der Beschreibung der Bedeutung lexikalischer Einheiten (Wortsemantik/ lexikalische Semantik) und mit den Regularitäten zur Ermittlung der Bedeutung komplexer Einheiten (Satzsemantik).

Deskriptive Statistik

- Welche Bedeutung eines Worts/einer Wortform ist häufiger/weniger häufig?

  Maus (Tier/Computerteil)
- Kontextspezifische Bedeutungshäufigkeit - in welchem lexikalischen oder syntagmatischen Kontext taucht welche Bedeutung häufiger auf?

  Er scrollt mit der Maus (Lesart Computerteil)

  Im Keller piepst eine Maus (Lesart Tier)
- Textspezifische Bedeutungshäufigkeit - in welchen Bedeutungen taucht ein Lexem in bestimmten Textsorten (z.B. Märchen vs. Pressetext) / Textsammlungen (z.B. Computermagazinartikeln) auf?

- Welche Bedeutungsklassen tauchen in bestimmten Textsammlungen häufig auf (etwa: Computerkomponten in einer Sammlung von Texten auf einem Computermagazin; Tiere in einer zoologischen Fachzeitschrift)?
- Kompositaanalyse. Im folgenden sind typische semantische Klassen von Erstglieder zu *Blatt* nach Ihrem Mutual Information Wert geordnet; weiter oben die stärker assozierten Klassen:

  PFLANZE : MI : 1.5 (Rosenblatt)

  DRUCKWERK : MI : 1.0 (Manuskriptblatt)

Induktive Statistik

Die Schlussfolgerungen aus semantischen Datenbeschreibung werden z.B. verwendet zur Bedeutungsdisambiguierung (etwa im Fall *Maus*). Disambiguierung von Wortbedeutungen ist für zahlreiche Anwendungen unverzichtbar, so etwa im Bereich der maschinellen Übersetzung. Statistische semantische Regularitäten können auch verwendet werden in der thematischen Zuordnung von Texten (Textkategorisierung) oder in der Kompositaanalyse.

1.2.2 Statistik in der Morphologie

Die Morphologie beschäftigt sich mit dem Aufbau von Wortformen aus kleineren, noch bedeutungrelevanten Einheiten (Morphemen).

Deskriptive Statistik

- Vollformen-Grundformen-Reduktion. Im folgenden Beispiel ist die Reduktion zu *Hau* wesentlich unwahrscheinlicher als die zweite:

  Haus - Hau

  Haus - Haus
- Häufigkeit von Affixen und Ihre morphologische Zuordnung

geXt =(99%) = X+Part.Perf.

- Das Beispiel aus dem Gebiet Semantik (s.o.) zur Kompositasegmentierung, hat natürlich auch mit Morphologie zu tun.

Induktive Statistik

Die Anwendungen obiger Daten sind relativ einsichtig:

- Statistiken über die Häufigkeit von Grundformen für eine Vollform können zur richtigen Grundformenreduzierung herangezogen werden
- Analyse unbekannter Formen

gecancelt = cancel + Part.Perf.

weil

geXt =(99%) = X+Part.Perf.

- Statistiken über Kompositaregularitäten können zur korrekten Analyse von Komposita herangezogen werden

1.2.3 Textlinguistik und Textklassifikation

Die Textlinguistik ist ein sehr weites Gebiet. Sie beschäftigt sich mit dem satzübergreifenden Regularitäten in Texten, Texteigenschaften (Wortartenverteilung) und Textklassifikation (Genre, Themengebiete).

Deskriptive Statistik

- Bestimmung von themengebietspezifischem Wortschatz/Konstruktionen
- Welche Wörter / Wortarten sind häufig in bestimmten Textsorten (z.B. Homepages vs. wissentschaftliche Artikel)
- andere sortenspezifische Texteigenschaften (Wortlängen, Wortarten)

Induktive Statistik

- Anaphernresolution (Bezug von Pronomina)
- Erkennung von Textsorten, Themengebieten etc.;

1.2.4 Lexikographie

Die Lexikographie beschäftigt sich mit der Praxis der Wörterbucherstellung, d.h. mit der Erfassung von Wörtern und Mehrwortlexemen und ihren Eigenschaften.

Deskriptive Statistik

Für die Lexikographie lassen sich morphologische, syntaktische, semantische und textlinguistische Daten (s.o.) verwenden. Einige Beispiele für statistisch ermittelbare lexikalische Eigenschaften sind:

- Stilebenen
- andere textsortenspezifische Eigenschaften
- Häufigkeiten/Gebräuchlichkeit/historische Entwicklung

Induktive Statistik

Statistische Methoden werden in der Lexikographie v.a. verwendet um lexembezogene Daten aus Textkorpora zu ermitteln. Z.T. werden die statistisch akquirierten Daten nur als Vorauswahl für eine weitere manuelle Klassifikation verwendet.

- automatische oder semiautomatische Ermittlung von textsorten-/fachgebietsspezifischen Eigenschaften von Lexemen;
- Zur Erkennung und Extraktion von Mehrwortlexemen gehört die Analyse von Kookurrenzen - d.h. des statistisch signifikanten gemeinsamen Auftretens von Wörtern und Wortformen, und darauf aufbauende Schlussfolgerungen über die Zusammengehörigkeit von Wörtern und Wortformen. Die Ermittlung von bestimmten Typen von Mehrwortlexemen wird im Skript noch genauer beschrieben (s.u.).

1.3. Anwendungsbereiche für Statistik in der CL

Hier einige Anwendungen, in denen statistische Methoden eine Rolle spielen (können) - eigentlich sind es weitgehend alle Anwendungsbereiche der CL in denen die Statistik inzwischen eine Rolle spielt.

Rechtschreibkorrektur und Grammatikkorrektur

Rechtschreib- und Grammatikkorrektur sind Anwendungen, die jedem Computerbenutzer aus Textverarbeitungsprogrammen bekannt sind. Rechtschreibkorrektur wird aber auch in zahlreichen anderen Anwendungen (z.B. Internet-Suchmaschinen) eingesetzt.
Verbesserung der Rechtschreib- und Grammatikkorrektur:

*in der **s**chule wird gelacht ...*

Das Wort s/Schule kann Nomen oder Verb sein - nicht aber im Kontext Präp-Artikel - hier ist es immer Nomen. Deshalb ist die Kleinschreibung hier falsch

er sagte, er Schule zur Zeit Arbeitslose Computerlinguisten.

Auch hier ist Groß- und Kleinschreibung erkennbar falsch.

Textgenerierung und Wortvorschlagssysteme

Textgenerierung ist die Erzeugung von Texten aus Inhaltsdaten; z.B. aus einer semantischen Formel, einer formellen Beschreibung, einem Datensatz einer Datenbank.
Überprüfung der Plausibilität einer von einer Grammatik generierten Wortfolge;
Kommunikationshilfen mit **Wortvorschlagssystemen** - hier können einfache Wortfolgestatistiken brauchbare Fortsetzungen vorschlagen:

Ich möchte dir einen W... (Witz erzählen)

Spracherkennung

Spracherkennung ist die Umwandlung gesprochener Sprache in einen elektronischen Text. Die automatische Transskription gesprochener in geschriebene Sprache beruht schon seit langem in erster Linie auf statischen Methoden:

- Hidden-Markov-Modelle zur Laut-Phonem-Zuordnung
- Sprachmodelle zur Festlegung wahrscheinlicherer Wortfolgen

Textklassifikation

Textklassifikation ist die Einordnung von Texten in vordefinierte Kategorien irgendwelcher Art.
Sprachenidentifikation (s. die meisten Suchmaschinen) - Ermittlung der Wahrscheinlichkeit mit der ein Text einer Sprache zugeordnet werden kann.
Genreklassifikation (Texttyp)
Filter. z.B. Pornofilter, wie in den meisten Suchmaschinen verfügbar

- Ermittlung des relevanten Vokabulars
- Zuordnungswahrscheinlichkeit zur einer Sammlung von Referenzdokumenten

Inhaltliche Klassifikation wie z.B. Scirus (www.scirus.com) - automatische Erkennung eines wissenschaftlichen Fachgebiets

Textretrieval

Textretrieval (Information Retrieval) befasst sich mit dem Auffinden spezifischer Textdokumente in einer Dokumentensammlung. Die bekanntesten Textretrievalsysteme sind sicher die Internetsuchmaschinen.

Ranking - Ermittlung der Relevanz eines Dokuments bez. der Suchanfrage, abhängig von der Häufigkeit und Position des / der Suchterme/s, von der Länge des Dokuments ...
Vorklassifizierung von Texten nach verschiedenen Kriterien (Sprache, Domäne, Typ) (siehe Textklassifikation)
Vorverarbeitung der Anfrage (Query)

Maschinelle Übersetzung und Alignierung mehrsprachiger Korpora

Statistische Algorithmen beruhen auf

- Zuordnungen in bilingualen Korpora
- Eigenschaften der Einzelsprachen

Lexikalische Zuordnung: Welches Wort ist wahrscheinlich eine Übersetzung eines anderen Wortes;
Disambiguierung: Welche Bedeutung eines polysemen Wortes liegt vor

Bruchstück und Satz-Zuordnung: Welche Phrase / welcher Satz ist wahrscheinlich eine Übersetzung eines anderen;
Generierung: Welche Übersetzung ist ein wahrscheinlicherer Satz in der Zielsprache?

1.4. Übungen

Wie stellen Sie sich die Anwendung statistischer Methoden in der maschinellen Übersetzung vor? Welche Daten werden verwendet? Wozu wird die statistische Auswertung eingesetzt?

2. Statistisches Rüstzeug

2.1. Grundbegriffe der Statistik

Grundgesamtheit (a. Population)

Menge aller statistischen Einheiten, über die man Aussagen gewinnen will.

In der Computerlinguistik können dies z.B. sein: Grapheme, Phoneme, Wörter, Wortfolgen, Texte.

Die Grundgesamtheit kann endlich sein (Bsp.: Phoneme, Grapheme) oder unendlich (Bsp.:Texte).

Stichprobe; repräsentative Stichprobe

Eine Stichprobe ist eine Teilmenge der Grundgesamtheit; sie sollte, um Rückschlüsse auf die Gesamtheit zu erlauben, möglichst **repräsentativ** sein.

LINGUISTIK:

Beispiel I

Wenn man Aussagen über die Häufigkeitsverteilung von Wörtern in allen standarddeutschen Texten machen will, sollte man als Stichprobe nicht ausschließlich Rezepttexte verwenden; eine solche Stichprobe ist nicht repräsentativ.

Beispiel II

Wenn man die durchschnittliche Wortlänge deutschen Nomina in einem Wörterbuch (ohne Komposita) berechnen will, erhält man bei einer repräsentativen Stichprobe einen Wert von ca 8,8. Würde man als Stichprobe alle Bezeichnungen für Vögel und Säugetiere heranziehen, wäre diese sicher nicht repräsentativ; hier ergibt sich nämlich ein Wert von 6,1!

Merkmale

Eigenschaften der statistischen Einheiten der Grundgesamtheit. Merkmale können bestimmte Ausprägungen/Werte annehmen.

LINGUISTIK: Untersucht man Wörter, könnten interessante Merkmale etwa sein: Genus von Wörtern, Wortart, Wortlänge, Worthäufigkeit in Texten ...; das Merkmal Genus hat die Ausprägungen/Werte maskulin, feminin und neutrum; die Wortlänge in Buchstaben hat als Werte die natürlichen Zahlen.

Diskrete vs. stetige Merkmale

Merkmale können diskret (abzählbar) oder stetig (beliebig viele Zwischenwerte) sein.

LINGUISTIK: die Länge eines Textes in Zeichen ist diskret. Die Tonhöhe eines Lautes in Hertz ist dagegen stetig.

Bei sehr feiner Abstufung eines diskreten Merkmals spricht man von quasi-stetigen Merkmalen (z.B. Textlänge bei längeren Texten). Andererseits lassen stetige Merkmale lassen sich durch Gruppierung der Werte in Intervallen auf diskrete Merkmale abbilden (z.B. die Tonhöhe in Hertz in die Intervalle 0-100, 100-200, 200-300 Hertz).

Merkmalstypen

Nominalskalierte Merkmale sind solche, deren Ausprägungen Kategorien sind, die nicht sinnvollerweise quantitativ interpretiert werden können.

LINGUSTIK: *Genus* ist ein solches Merkmal. Es hat im Deutschen drei Ausprägungen: *maskulin, feminin und neutrum*. Diese Werte lassen sich nicht sinnvoll in einer aufsteigenden oder absteigenden Reihe anordnen.

Ordinalskalierte Merkmale sind solche, bei denen die Ausprägungen skalar geordnet werden können, aber die Abstände nicht gleichermaßen interpretiert werden können.

LINGUISTIK: Grammatikalitätsurteile lassen sich auf einer Skala abbilden (*ungrammatisch= 0, schlecht = 1, bedingt akzeptabel = 2, voll akzeptabel = 3*), aber die Abstände zwischen den Werten sind nicht einheitlich (es lässt sich damit hier auch nicht sagen - obwohl der "Nullpunkt" *ungrammatisch* vorhanden ist, das *bedingt akzeptabel* doppelt so grammatisch ist wie *schlecht*).

Intervallskalierte Merkmale: Hier können die Abstände skalar interpretiert werden, es gibt aber keinen sinnvollen Nullpunkt.

LINGUISTIK: Kein linguistisches Beispiel gefunden. Nicht-linguistisches Beispiel ist die Temperatur in Grad Celsius (es ist nicht sinnvoll, zu sagen, 20 Grad sind doppelt so warm wie 10 Grad Celsius).

Verhältnisskaliert: Hier können die Abstände skalar interpretiert werden und es gibt es einen sinnvollen Nullpunkt:

LINGUISTIK: Beim Merkmal *Länge in Zeichen eines Wortes* kann man sinnvollerweise von "doppelt so lang" sprechen - ein Wort mit zwanzig Buchstaben ist doppelt so lang wie ein Wort mit 10 Buchstaben.

Eine Skala für intervall- und verhältnisskalierte Merkmale wird auch als Kardinalskala bezeichnet.

Quantitative Merkmale vs. qualitative Merkmale

Qualitative bzw. kategoriale Merkmale sind Größen, die endlich viele Ausprägungen besitzen und nominalskaliert, oder maximal ordinalskaliert sind. Ordinalskalierte qualitative Merkmale lassen sich z.T. auch quantitativ interpretieren.

LINGUSTIK: *Genus*, *Grammatikalitätsurteile* sind qualitative Merkmale. Letztere lassen sich allerdings auch quantitativ interpretieren. Nach einer Abbildung auf die Zahlen 0-3 lassen sich für Grammatikalitätsurteile durchaus Mittelwerte bilden (quantitative Interpretation); es ist allerdings etwas fraglich, wie aussagekräftig sie sind.

2.1.1 Übungen

1) Sie wollen eine vergleichende Untersuchung der Wort- und Satzlänge deutscher seriöser Zeitungen mit deutschen Boulevardblättern vornehmen:

a) Was ist ihre Grundgesamtheit?

b) Wie kommen Sie an eine möglichst repräsentative Stichprobe?

c) Welche Merkmale untersuchen Sie, welchen Typ haben diese Merkmale?

2) Sie interessieren sich nun für die Fontgröße in Überschriften. Beantworten Sie hier dieselben Fragen.

2.2. Grundbegriffe der Deskription und Exploration von Daten

Urliste, Rohdaten

Reine Daten, d.h. Merkmalswerte für Untersuchungseinheiten.

LINGUSTIK: Die Rohdaten einer Untersuchung über die Wortlänge wären die Zuordnungen Wort - Merkmal also etwa:

Die:3

Rohdaten:8 ...

Absolute und relative Häufigkeit

Die bloße Zählung eines Merkmals in den Rohdaten nennt man absolute Häufigkeit (notiert als f). Das Verhältnis dieser Häufigkeit zur Gesamtszahl von Daten N ist die relative Häufigkeit (h). Es gilt also:

$$h = f/N$$

LINGUISTIK: Für o.g. Untersuchung kann an den Rohdaten abzählen, wie häufig welche Wortlänge ist. Die absolute Häufigkeit von Wörtern mit einer bestimmten Wortlänge lässt sich durch bloßes Abzählen ermitteln, die relative Häufigkeit ergibt sich aus der absoluten Häufigkeit und der Menge aller Wörter.

Dokumentfrequenz (document frequency)

Gilt für Dokumentensammlungen (z.B. Sammlung von Webseiten, Emails). Die Dokumentfrequenz eines sprachlichen Ausdrucks - etwa einer Wortform - ist die Zahl der Dokumente, in denen der Ausdruck vorkommt (egal wie oft). Die Dokumentfrequenz eines Terms ist vor allem für Anwendungen im Bereich des Information Retrieval wichtig.

2.2.1 Mittelwerte

Arithmetisches Mittel

Das arithmetische Mittel ergibt sich aus der Aufsummierung aller numerischen Werte in den Rohdaten und der Division durch die Anzahl aller Werte. Die Ermittlung des arithmetischen Mittels ist nicht möglich bei nominalskalierten Merkmalen (z.B. Genus), da diese sich nicht quantitativ interpretieren lassen, und ist bei ordinalskalierten Daten vorsichtig zu handhaben.

LINGUSTIK: Für das Beispiel Wortlänge wird aus der Urliste das arithmetische Mittel folgendermaßen berechnet:

z_i sei die Wortlänge des Wortes w_i, n sei die Menge alter Wörter

Dann ist der arithmetische Mittelwert x_{ar}:

$$x_{ar} = \frac{(z_1 + z_2 \ldots z_n)}{n}$$

Das arithmetische Mittel reagiert empfindlich auf Ausreißer. Eine Möglichkeit, diese Ausreißer zu eliminieren, ist die Ermittlung des getrimmten arithmetischen Mittels, bei dem Extremdaten nicht berücksichtigt werden.

LINGUSTIK: Wenn wir etwa für einen Text von fünf Wörtern, von denen alle 4 Zeichen lang sind, bis auf eines, das 24 Zeichen lang ist, den Mittelwert der Wortlänge berechnen, ist dieser 8. Lässt man 20% extremsten Daten weg (also hier die Werte für das längste Wort), ergibt sich ein arithmetisches getrimmtes Mittel von 4.

Median

Ein weiterer Mittelwert ist der Median. Hier werden Ausreißer nicht stark gewichtet, d.h. Der Median ist robust und resistent gegen Extremwerte. Er ist der Wert, für den gilt, dass die Hälfte der Daten oberhalb, und die andere Hälfte der Daten unterhalb liegt.

Man ordnet die Werte nach Größe

$w_1 >= w_2 >= w_3 .... >= w_i$

Dann ist der Median der in der Mitte liegende Wert (bei ungerader Anzahl von Daten) bzw. der Mittelwert der beiden in der Mitte liegenden Werte (bei gerader Anzahl von Daten).

LINGUSTIK: Im obigen Beispiel zu Wortlänge wäre der Median x_{med} 4. Der Ausreißer nach oben für nur ein Wort spielt keine Rolle:

$$\downarrow$$

$$24 >= 4 >= 4 >= 4 >= 4$$

Modus

Der Modus ist die Ausprägung mit der größten Häufigkeit.

LINGUISTIK: Im vorliegenden Beispiel ist der Modus x_{mod} 4.

Ein Beispiel

Das folgende Diagramm zeigt die Wortlänge aller Nomina im CISLEX, die nicht Komposita sind.

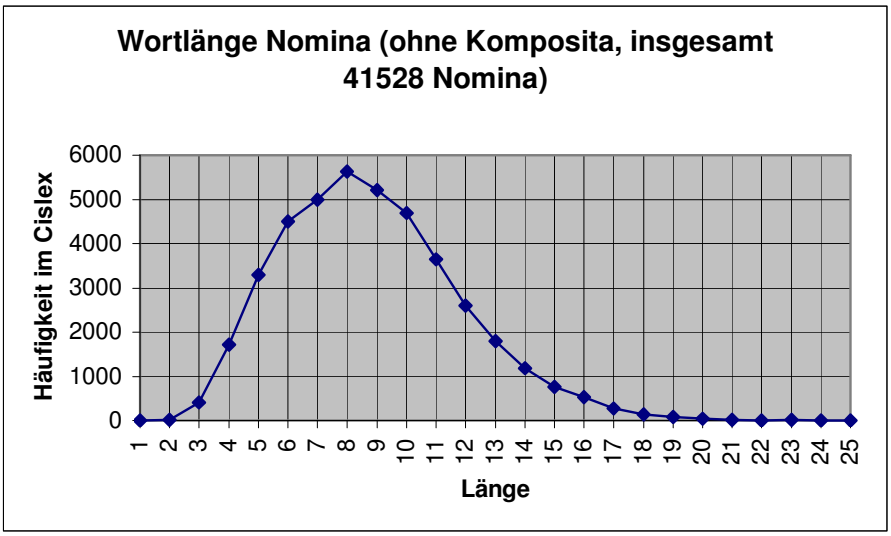

Abbildung 2-1: Wortlängen von Nomina im CISLEX

Arithmetisches Mittel:	8,83
Median :	9
Modus:	8

2.2.2 Eigenschaften von Datenverteilungen

Allgemein

Bei gleichen Mittelwerten können die Daten unterschiedliche Verteilungen aufweisen. So ist z.B. die durchschnittliche Wortlänge für 10 Wörter immer 10 Zeichen in allen folgenden, völlig unterschiedlichen Fällen:

1. alle 10 Wörter sind 10 Zeichen lang
2. 5 Wörter sind 5, 5 Wörter 15 Zeichen lang
3. 9 Wörter sind 9, ein Wort ist 19 Zeichen lang.

Unimodal - bimodal - multimodal

Weist eine Verteilung einen Gipfel auf, spricht man von einer unimodalen Verteilung, bei zwei bzw. mehr Gipfeln von einer bimodalen bzw. multimodalen Verteilung. Das Beispiel der Nominallänge in Abbildung 2-1 zeigt eine typische unimodale Verteilung.

Lageregeln - Symmetrie

Symmetrie bezeichnet die Eigenschaft, dass die Werte in gleicher Weise nach oben bzw. unten vom arithmetischen Mittel abweichen - im obigen Beispiel sind 1) und 2) symmetrische Verteilungen. Zeichnet man ein Diagramm bzw. eine Kurve, ist diese Figur symmetrisch.

Symmetrische Verteilung: $x_{ar} \approx x_{med} \approx x_{mod}$

Linkssteile Verteilung: $x_{ar} > x_{med} > x_{mod}$

Rechtssteile Verteilung: $x_{ar} < x_{med} < x_{mod}$

Betrachten wir Abbildung 2-1 und die dort zu findenden Mittelwerte, widersprechen sich Diagrammaussehen (spricht für linkssteile Verteilung) und Mittelwerte. Dies liegt daran, dass es hier keine Zwischenwerte gibt; würde die diskreten Werte durch eine stetige Verteilung approximieren, läge der Modus wohl etwas über 8.0, und wir hätten die Eigenschaften einer linkssteilen Verteilung vorliegen.

Standardabweichung und Varianz

Die Standardabweichung und die Varianz sind Maße für die Streuung einer Verteilung. Die Varianz berechnet sich folgendermaßen: man bildet die Summe aus den Quadraten der Abweichung aller Werte w vom arithmetischen Mittelwert x_{ar}, und teilt diese Summer durch die Anzahl der Werte.

$$s2 = \frac{((w_1 - x_{ar})^2 + (w_2 - x_{ar})^2 ... + (w_n - x_{ar})^2)}{n}$$

Manchmal wird statt durch die Anzahl der Wert n auch durch die Anzahl der Werte minus 1 (n-1) geteilt.

Die Standardabweichung s ist die Wurzel aus der Varianz.

LINGUISTIK: Für unser Beispiel Wortlänge w lässt sich die Varianz s^2 nach folgender Formel berechnen

$s^2 = (1/n)\ ((w_1-x_{ar})^2 + ... + (w_n-x_{ar})^2)$

Die Wurzel hieraus ist dann die Standardabweichung.

2.3. Übung

Schreiben Sie ein Programm, das für jede beliebige Wortliste (ein Wort pro Zeile) die durchschnittliche Wortlänge in Buchstaben untersucht. Das Programm soll:

1. die Wortliste einlesen
2. das arithmetische Mittel berechnen
3. die Varianz berechnen
4. die Ergebnisse ausgeben.

3. Grundbegriffe der Wahrscheinlichkeitstheorie

(v.a. nach Manning/Schütze: 40ff und Fahrmeir /Künstler/Pigeot/Tutz: 171ff)

Übersicht

Um entscheiden zu können, ob eine statistische Beobachtung zufällig entstanden ist, oder ob sie eine bestimmte Aussage erlaubt, ist es nötig, einen Begriff von der Wahrscheinlichkeit von Ereignissen zu entwickeln.

LINGUISTIK: Wir wollen aus einem Korpus Assoziationspaare extrahieren. Für ein bestimmtes Wortpaar, etwa Leviten + lesen erhalten f=20, wobei die Häufigkeit von Leviten 21 ist und die von lesen 750. Der Korpus hat 1 Million Wörter, unsere Fenstergröße sind 5 Wörter. Ohne Wahrscheinlichkeitsrechnung können wir hier bereits eine auffällige Häufung feststellen. Die Wahrscheinlichkeitsrechnung ermöglicht es uns nun aber, zu sagen, wie signifikant diese Abweichung ist, d.h. mit welcher Wahrscheinlichkeit dieser Wert zufällig entstanden sein könnte.

Zufallsvorgang, Zufallsexperiment und Ergebnismenge/Ergebnisraum (sample space)

Ein Zufallsvorgang führt zu einem oder mehreren möglichen Ergebnissen. Ein Zufallsexperiment ist ein Zufallsvorgang unter kontrollierten, wiederholbaren Bedingungen. Die Menge der möglichen Ergebnisse eines Zufallsvorgangs ist der Ergebnisraum (auch Ergebnismenge o. Stichprobenraum). Er wird notiert als Ω (großes Omega).

Ereignis, Elementarereignis

Eine Menge von möglichen Ergebnissen (ω_1 .. ω_n) eines Experiments ist ein Ereignis. Ein Ereignis ist damit stets eine Untermenge von Ω. Ω selbst heißt das **sichere Ereignis**, die leere Menge ist das **unmögliche Ereignis**. Eine Menge, die nur ein mögliches Ergebnis eines Experiments enthält, wird als **Elementarereignis** bezeichnet. Der Ereignisraum ist die Menge aller Teilmengen von Ω.

LINGUISTIK: Hier ein Zufallsexperiment: Wählen wir aus einem Text zufällig ein Wort aus und interessieren uns für die Wortart also SUBSTANTIV, VERB, ADJEKTIV oder ANDERE (die wir nicht weiter unterscheiden). Der Ergebnisraum Ω ist dann {SUBSTANTIV, VERB, ADJEKTIV, ANDERE}. Die Elementarereignisse sind dann {SUBSTANTIV}, {VERB}, {ADJEKTIV} oder {ANDERE}. Andere Ereignisse wären {SUBSTANTIV, ADJEKTIV} - zu interpretieren als: entweder SUBSTANTIV oder ADJEKTIV etc.

Bernoulli-Experiment (Bernoulli trial)

Ein Bernoulli-Experiment ist ein Zufallsexperiment mit zwei möglichen Ergebnissen. Die beiden möglichen Ergebnisse werden als "Treffer" bzw. "Niete" (success / failure) bezeichnet, bzw. auf die Zahlen 0 oder 1 abgebildet.

BEISPIEL: Das Werfen einer Münze ist ein Bernoulli-Experiment mit den Ergebnissen Kopf oder Zahl.

LINGUISTIK: Untersuchen wir einen Text auf die Frage hin, ob ein bestimmtes Wort an einer bestimmten Stelle auftritt, lässt sich dies ebenfalls als Bernoulli-Experiment beschreiben: das Wort ist da = Treffer / das Wort ist nicht da = Niete. Im Hinblick auf das Auftreten eines bestimmten Wortes lässt sich damit ein laufender Text als eine Folge von Bernoulli-Experimenten beschreiben.

Wahrscheinlichkeitsraum

Tripel aus Ω, F, P

- Ω ist ein Ergebnisraum;
- F ist eine Ereignismenge, genauer, die Menge aller möglichen Untermengen von Ω (Potenzmenge);
- P ist eine Funktion, die Ereignissen aus F eine reelle Zahl zwischen 0 und 1 zuordnet (eine Wahrscheinlichkeit).

Für P gelten folgende Axiome (Grundannahmen):

1. $P(\Omega) = 1$
2. $P(\emptyset) = 0$
3. Für disjunkte Ereignisse A_1-A_n aus F gilt, dass die Wahrscheinlichkeit der Vereinigungsmenge dieser Ereignisse gleich der summierten Wahrscheinlichkeit dieser Ereignisse A_i ist.

1) heißt, salopp gesagt, dass die Wahrscheinlichkeit, dass ein beliebiges Ergebnis (das sichere Ereignis) auftritt, immer 1 ist; 2) sagt, dass die Wahrscheinlichkeit, dass keines der möglichen Ergebnisse (das unmögliche Ereignis) eintritt, null ist.

Aus 1.) und 3.) folgt, dass die aufsummierte Wahrscheinlichkeit aller Elementarereignisse immer 1 sein muss. Außerdem folgt, dass aus der Wahrscheinlichkeit der Elementarereignisse alle anderen Wahrscheinlichkeiten berechnet werden können.

LINGUISTIK: Als Wahrscheinlichkeit eines Wortes w oder einer zugeordneten Kategorie k (z.B. Wortart) in einem Korpus können wir die relative Häufigkeit h(w) bzw. h(k) ansetzen - d.h. die Wahrscheinlichkeit, das man bei der zufälligen Auswahl eines Wortes im Korpus auf eine Instanz von w bzw. auf ein Wort mit Kategorie k trifft.

In unserem Beispiel können könnte eine Wahrscheinlichkeitsfunktion den Elementarereignissen folgende Wahrscheinlichkeiten zuordnen:

$P(\{S\}) = 0.3$ (SUBSTANTIV)
$P(\{V\}) = 0.2$ (VERB)
$P(\{A\}) = 0.1$ (ADJEKTIV)
$P(\{X\}) = 0.4$ (ANDERE)

Die Wahrscheinlichkeit des Ereignisses {V,S}, d.h. die Wahrscheinlichkeit, dass wir zufällig ein Verb oder ein Substantiv auswählen ist nun:

$P(\{V,S\}) = P(\{V\}) + P(\{S\}) = 0,2+0,3 = 0,5$

Analog lassen sich die Wahrscheinlichkeiten von allen anderen Ereignissen berechnen.

n.b. Die Wahrscheinlichkeiten die hier errechnet werden, sind die Wahrscheinlichkeiten, dass bei der rein zufälligen Auswahl eines Wortes aus dem Korpus, das Wort von der angegebenen Kategorie ist.

Laplace-Wahrscheinlichkeitsraum

Ein Laplace-Wahrscheinlichkeitsraum entsteht bei einem Laplace-Experiment: Alle Elementarereignisse eines Laplace-Experiments haben dieselbe Wahrscheinlichkeit.

LINGUISTIK: Es gibt wohl kaum Laplace-Wahrscheinlichkeitsräume in der statistischen Beschreibung sprachlicher Phänomene. Laplace-Wahrscheinlichkeitsräume ergeben sich z.B. durch Würfeln mit einem normalen Würfel (jede Augenzahl ist gleich wahrscheinlich) oder dem Werfen einer Münze.

Wahrscheinlichkeit von Bernoulli-Experimenten

Für ein Bernoulli-Experiment mit genau zwei möglichen Ergebnissen (s.o.) muss nur die Wahrscheinlichkeit für ein Elementarereignis angegeben werden – da es nur zwei Elementarereignisse gibt ({ERFOLG} oder {NIETE}) ergibt sich die Wahrscheinlichkeit des einen aus der Wahrscheinlichkeit des anderen, denn die Wahrscheinlichkeiten müssen sich auf den Wert 1 aufsummieren (die Wahrscheinlichkeit von $\Omega = 1$, s.o.). Ist die Wahrscheinlichkeit für {ERFOLG} = p, so ist die Wahrscheinlichkeit für {NIETE} = 1-p.

Kombinierte Wahrscheinlichkeiten von unabhängigen Ereignissen

Angenommen, zwei Ereignisse A und B sind unabhängig. Dann ist die Wahrscheinlichkeit von $P(A \cap B)$ – d.h. die Wahrscheinlichkeit, dass beide Ereignisse gemeinsam auftreten das Produkt der Einzelwahrscheinlichkeiten.

$$P(A \cap B) = P(A) * P(B).$$

Die Annahme der Unabhängigkeit gilt etwa der Fall bei Würfelexperimenten: Angenommen, Sie werfen zwei Würfel gleichzeitig. Dann ist die etwa die Wahrscheinlichkeit zwei Sechser zu würfeln $P(\{6\}) * P(\{6\}) = 1/6 * 1/6 = 1/36$. Genau dasselbe gilt, wenn Sie einen Würfel zweimal hintereinander werfen.

Bedingte Wahrscheinlichkeit (conditional probability)

Die Wahrscheinlichkeit für ein bestimmtes Ereignis A, gegeben Ereignis B.

Sie ist $P(A|B) = P(A \cap B) / P(B)$

Linguistik: Führen wir das Beispiel der Kategorienwahrscheinlichkeiten in einem Text fort: Nehmen wir an, die Wahrscheinlichkeit, bei zufälliger Auswahl einer Zweiwortfolge die Folge <S,V> zu ziehen, sei 0.15. Was ist nun die Wahrscheinlichkeit, eine Folge <S,V> zu erhalten, wenn als erster Teil der Kategorienfolge bereits ein Substantiv gezogen wurde?

Ereignis A: wir ziehen als zweite Kat. eine Verb, also $A = \{<A,V>,<S,V>,<X,V>\}$

Ereignis B: wir habe als erste Kat. ein Substantiv also: $B = \{<S,S>,<S,V>,<S,A>,<S,X>\}$

Dann ist $A \cap B = \{<S,V>\}$

$$P(\{<A,V>,<S,V>,<X,V>,<V,V>\}|\{<S,S>,<S,V>,<S,A>,<S,X>\}) = \frac{P(\{<S,V>\})}{P(\{<S,S>,<S,V>,<S,A>,<S,X>\})}$$

Folgende Vereinfachung ist nun möglich: die Wahrscheinlichkeit, dass eine Kategorie an erster oder zweiter Stelle einer Zweierfolge mit beliebigem anderen Element auftaucht, ist gleich der Wahrscheinlichkeit dieser Kategorie, also:

$$P(\{<S,S>,<S,V>,<S,A>,<S,X>\}) = P(\{S\})$$

Damit lassen sich in unsere Formel folgende Werte einsetzen

$$P(\{<S,V>\} / P\{S\} = 0{,}15/0.3 = 0.5$$

Unabhängigkeit

Ist $P(A) = P(A|B)$ dann spricht man davon, dass zwei Ereignisse unabhängig voneinander sind.

Nach obiger Formel für $P(A|B)$ gilt dann $P(A \cap B) = P(A) * P(B)$ - d.h. die Wahrscheinlichkeit, dass beide Ereignisse zusammen auftreten, ist gleich dem Produkt der Wahrscheinlichkeiten (s.o.).

LINGUISTIK: Die Annahme der Unabhängigkeit und der Vergleich der Werte bei angenommener Unabhängigkeit im Vergleich zum tatsächlichen Wert der bedingten Wahrscheinlichkeit kann sehr aufschlussreich sein. In obigem Beispiel ist die $P(\{V\}) = 0{,}2$, die Wahrscheinlichkeit von $\{<\_,V>\}|\{<S,\_>\}$ dagegen höher, d.h. die Wortfolge Substantiv-Verb ist häufig - beide Ereignisse sind korreliert. Die Wahrscheinlichkeit von $\{<S,V>\}$ ist ebenfalls viel höher als $P(\{S\}) * P(\{V\})$ (= 0,06).

Satz von Bayes

Es seien A_1 - A_k eine disjunkte Zerlegung von Ω.

$$P(A_j|B) = \frac{P(B|A_j)\,P(A_j)}{\sum (P(B|A_i) * P(A_i)} = \frac{P(B|A_j)\,P(A_j)}{P(B)}$$

Erläuterung: Der Satz von Bayes erlaubt es, die bedingte Wahrscheinlichkeit von A gegeben B aus den Wahrscheinlichkeiten B gegeben A und den einfachen Wahrscheinlichkeiten B bzw. A zu errechnen.

Der Satz von Bayes lässt sich herleiten aus der Formel für die bedingte Wahrscheinlichkeit (s.o.) und dem Satz der totalen Wahrscheinlichkeit, der es erlaubt, die Wahrscheinlichkeit für ein Ereignis B aus den bedingten Wahrscheinlichkeiten zu berechnen: $P(B) = \sum (P(B|A_i) * P(A_i)$

LINGUISTIK Angenommen, in einem Text sind 1 Prozent der Wörter Bezeichnungen für Personen. Wir setzen einen HIWI ein, der Personenbezeichnungen markieren soll. Wir finden leider heraus, dass er etwas schlampig arbeitet:

von 100 Personenbezeichnungen markiert er nur 90.

von 100 anderen Wörtern markiert er eine versehentlich als Personenbezeichnung.

Wie wahrscheinlich ist es nun, dass ein markiertes Wort wirklich eine Personenbezeichnung ist?

B steht für: ist markiert;

A_1 steht für: ist eine Personenbezeichnung;	P(A1) = 0.01
A_2 steht für: ist keine Personenbezeichnung;	P(A2) = 0.99
$B\|A_1$ steht für: ist Personenbezeichnung und markiert;	P(B|A1) = 0.9
$B\|A_2$ steht für: ist nicht Personenbez. aber markiert;	P(B|A2) = 0.01

Dann ist $P(A_1|B)$ nach folgender Formel zu berechnen

$$P(A_1|B) = \frac{P(B|A_1)\,P(A_1)}{P(B|A_1) * P(A_1) + P(B|A_2) * P(A_2)}$$

$$P(A_1|B) = \frac{0.9 * 0.01}{0.9 * 0.01 + 0.01 * 0.99} = 0,009/(0.009+0.0099) = 0.47$$

Die Wahrscheinlichkeit, bei einem markierten Wort tatsächlich eine Personenbezeichnung vor sich zu haben, ist damit geringer als 1/2.

Zufallsstichproben

Zieht man aus einer endlichen Grundgesamtheit mit N Einheiten, n Einheiten, spricht man von einer Zufallsstichprobe vom Umfang n.

LINGUISTIK: Wähle ich aus einem Wörterbuch von N Wörtern n Wörter zufällig aus, dann ist die resultierende Wortliste eine Zufallstichprobe vom Umfang N.

Kann jede Einheit aus der Grundgesamtheit mehrmals gezogen werden, spricht man von Ziehen mit Zurücklegen; kann jede Einheit nur einmal gezogen werden (wie etwa beim Lotto), dann spricht man von Ziehen ohne Zurücklegen.

Binomialverteilung

Die Binomialverteilung beschreibt die Wahrscheinlichkeit für eine bestimmte Anzahl des Auftretens eines Ereignisses beim Ziehen mit Zurücklegen (also bei gleich bleibender Wahrscheinlichkeit) in einem Bernoulli-Experiment.

Nehmen wir an, wir kennen die Wahrscheinlichkeit p für das Ergebnis {ERFOLG} eines Bernoulli-Experiments. Wie wahrscheinlich ist es nun, bei einer Zufallsstichprobe des Umfangs n genau x Treffer zu bekommen?

Hierfür gilt folgende Formel:

$$f(x) = \binom{n}{x} \bullet p^x \bullet (1-p)^{n-x}$$

Das $\binom{n}{x}$ (sprich: n über x) ist gleich n! / (x! * (n-x)!). Zur Erinnerung: n! (sprich n Fakultät) ist n*(n-1)*(n-2)...*1.

BEISPIEL: Die Wahrscheinlichkeit beim viermaligen Werfen einer fairen Münze genau zwei Treffer zu bekommen ist:

$$f(2) = \binom{4}{2} \bullet 0{,}5^2 \bullet (0{,}5)^2$$

Das ergibt 3/8.

LINGUISTIK: nehmen wir an, ein Wort hat die Auftretenswahrscheinlichkeit p. Mithilfe der obigen Formel lassen sich nun etwa folgende Fragen beantworten: Wie wahrscheinlich ist es, dass das Wort genau/mindestens/mehr als N mal im Text auftritt?

Fragen, die uns später beschäftigen werden sind: Wie groß muss eine Zufallsstichprobe sein, dass sie eine Aussage über die Grundgesamtheit mit einer bestimmten Sicherheit erlaubt?

LINGUISTIK: Wie viele Wörter muss man aus einem Korpus auswählen um eine halbwegs sichere Aussage zur durchschnittlichen Wortlänge im Korpus machen zu können?

4. Fragen der Textaufbereitung (z.T. nach Manning/Schütze, 123ff)

1.1. Dokumentenformate

4.1.1 Dokumentenformate: Formatierung

Gängige Formatierungsformate sind:

- nur Text: Der reine Text ohne irgendwelche Formatierungsmerkmale wie Fett/Kursiv/Fontgrößen etc. und ohne irgendwelche expliziten Strukturformatierungen (Überschrift etc.). Formatierung nur über Absatzmarken u.ä. Für Fragen des Zeichensatzes und der Zeichensatzkodierung s.u.
- HTML: die meisten Dokumente auf dem Internet, aber auch Intranetdokumente, sind in HTML verfasst. HTML bietet inhaltliche Formatierungstags (Überschriften, Absätze, hervorgehobener Text), Zeichenformatierung (Fett, Kursiv etc.). Sämtliche Tags in HTML sind standardisiert;
- XML bietet ein ähnliches Tagformat wie HTML, doch können Tags vom Benutzer selbst definiert werden (in einer DTD (Document Type Definition), die die Struktur des Dokuments festlegt). Ähnlich ist SGML (Standardized General Markup Language);
- RTF ist ein Microsoft-Austausch-Format,
- Alle Textverarbeitungsprogramme (WordPerfect, Word etc.) haben eigene Formate.

Informationsreiche Formate (HTML, XML etc.) können zur Vereinfachung der Auswertung in nur-Text überführt werden - in die umgekehrte Richtung ist eine solche Konvertierung nicht möglich.

4.1.2 Dokumentenformate: Annotierte Korpora

Ein vorliegendes Korpus kann bereits mit linguistischer (oder anderer) Information angereichert sein. Geschieht diese Informationsanreicherung auf Wortebene, spricht man von einem getaggten Korpus. Ein Tag kann unterschiedliche Informationen beinhalten. Häufig enthalten getaggte Korpora:

- syntaktische Information (v.a. Wortart)
- morphologische Information (Grundform s.u. Lemmatisierung)
- semantische Information (Bedeutungsklassen/Wortbedeutungen aus einem gegebenen Lexikon)

Ein getaggtes Korpus kann sein:

- manuell getaggt: d.h. die Tags wurden von Hand vergeben oder zumindest kontrolliert; je nach Typ von Information und Bearbeitern ist die Information dann recht zuverlässig

- automatisch getaggt: die Tags wurden durch ein statistisches oder regelbasiertes Verfahren vergeben. Die Zuverlässigkeit hängt ab von der Qualität des Tagging-Werkzeugs und dem Tag-Typ (Wortartentagger z.B. arbeiten zuverlässiger als semantische Tagger)

4.1.3 Dokumentenformate: Zeichensatz und Zeichensatzkodierung

Der Zeichensatz eines Dokuments sind die Zeichen, die das Dokument enthalten kann, also Buchstaben, Zahlen, Interpunktionszeichen und einige Sonderzeichen wie Währungszeichen, Paragraphenzeichen etc. Der Zeichensatz für westeuropäische Dokumente ist schon seit langer Zeit weitgehend standardisiert (erst kürzlich wurde allerdings das Euro-Zeichen eingeführt.
Die Zeichensatzkodierung bestimmt, durch welche Bytes bzw. Byte-Folgen jedes Zeichen des Zeichensatzes elektronisch dargestellt wird. Der meistverbreitete Standard hier ist die ASCII-Zeichensatzkodierung, die die 128 Zeichen einen Byte-Wert zuordnet. Fast alle anderen Zeichensatzkodierungen sind kompatibel zu ASCII: ASCII enthält nicht - neben vielen anderen Zeichen - die deutschen Umlaute (Ää etc.) und das 'ß'. Diese wurden auf verschiedenen Rechnertypen und Betriebssystemen durch verschiedene Werte dargestellt (DOS, Windows, Macintosh). Um ein deutsches Dokument darstellen zu können, ist es daher wichtig, den Zeichensatz zu kennen oder herauszufinden. Der derzeit gängigste Zeichensatz für das Deutsche ist Windows Codepage 1252, der weitgehend kompatibel ist zu ISO-8859-1 (ISO-Latin 1). Dieser Zeichensatz enthält fast alle Zeichen, die für westeuropäische Sprachen benötigt werden - nicht aber alle Zeichen für osteuropäische Sprachen mit lateinischen Zeichen.
Für die Darstellung aller Schriftzeichen aller Sprachen dieser Welt wurde der Zeichensatz UNICODE konzipiert. Er umfasst > 64 000 Zeichen und wird noch erweitert. Jedem Zeichen ist eine Zahl zugeordnet, wobei die Zahlen 0-256 kompatibel sind zu ASCII bzw. ISO-8859-1. Folgende Zeichensatzkodierungen sind für Unicode verbreitet:

- UCS-2: jedes Zeichen wird durch 2 Byte dargestellt. Allerdings sind auf diese Weise nur 65536 Zeichen darstellbar
- UCS-4: jedes Zeichen wird durch 4 Byte dargestellt. Dies ist sehr speicherplatzaufwendig
- UTF-8: jedes Zeichen wird durch 1 oder mehr Byte dargestellt (0-127 durch ein Byte, höhere durch mehrere Bytes, wobei das erste Byte die Zahl der folgenden Bytes, die zu dem Zeichen gehören, bestimmen). Dieses Verfahren hat den Vorteil, dass Texte in lateinischer Schrift nicht wesentlich länger werden als bei ein-Byte-Zeichensatzkodierungen.

In HTML-Dokumenten werden Umlaute und sonstige Sonderzeichen oft durch Zeichenkombinationen (HTML-Entities) dargestellt; 'ü' beispielsweise durch "ü".

4.2. Tokenisierung

Tokenisierung ist die Bestimmung von Tokens (i.d.R Wortformen) im Text. Selbst wenn man von einer sehr einfachen Definition des Begriffes Wort ausgeht (ein Wort steht zwischen zwei Blanks) stößt man auf Probleme:

- Als Blanks in diesem Sinne zählen nicht nur Leerzeichen, sondern sicher auch Satzzeichen, Gedankenstriche, Tabulatoren, Anführungsstriche u.ä.
- wie behandelt man Punkte (können Teil einer Abkürzung sein)?
- wie behandelt man Bindestriche?
    - Bindestriche in Wörtern wie Mathematik-Vorlesung
    - Bindestrich am Ende von Wörtern (Mathematik- und Statistikvorlesung)
- Wie behandelt man Klammerungen und andere Besonderheiten? (Computer)Linguistik; Dozent/in, DozentIn etc.
- Wie behandelt man Wörter mit Apostroph? ("*in wen'gen Fällen*" oder der englische Apostroph in "*Stefan's lecture*")?

In einigen Sprachen (z.B. Chinesisch, Japanisch, Koreanisch) werden Wörter zudem nicht systematisch durch Leerzeichen segmentiert.

4.3. Normalisierung

Normalisierung bedeutet die Rückführung eines Wortes auf eine Normalform. Dazu gehört:

- Normalisierung der Groß- bzw. Kleinschreibung. Dieses Problem stellt sich v.a. für Wörter am Text- und Satzanfang.
- Normalisierung von Bindestrichwörtern (mit Bindestrich oder getrennt/zusammengeschrieben)
- Normalisierung von Abkürzungen
- Normalisierung von am Zeilenende getrennten Wörtern. Manchmal ist nicht klar ob es sich um Trennstriche oder Bindestriche handelt. (Außerdem besonders schwierig im Deutschen (vor der Rechtschreibreform: Trippel-Konsonanten (*Schifffahrt*); *ck-kk*.

Eine Besonderheit des Deutschen sind Verben im abtrennbarem Partikel (*ab-trennen* etc.). Die Partikel kann bei Verb-Zweit-Stellung in einigem Abstand zum Stamm stehen - hier ist die Normalisierung besonders schwierig, besonders da viele der abtrennbaren Partikel homonym zu Präpositionen (*an*, *ab* ...) sind.

4.3.1 Lemmatisierung

Lemmatisierung ist eine extreme Form der Normalisierung: Im Text gefundene Wortformen werden auf eine kanonische Grundform zurückgeführt (z.B. Nominativ Singular für Nomina). Lemmatisierung kann über morphologische Regeln oder über ein Vollformenlexikon erfolgen - und ist nie hunderprozentig fehlerfrei.

4.3.2 Kompositasegmentierung

Für Sprachen, die, wie das Deutsche, Wörter in kompositionellen Bildungen zu einem Wort zusammenfügen können (z.B. Mathematikdozent = Dozent für Mathematik) kann eine Kompositasegmentierung die Ergebnisse statistischer Datenanalyse verbessern. Allerdings muss darauf geachtet werden, dass:
voll lexikalisierte, nicht-kompositionelle Komposita nicht segmentiert werden, da sich hier der Kopf allein (i. d. R das Letztglied) anders verhält als das Kompositum so gilt:

Mathematikdozent = Dozent (für Mathematik)

Himmelsschlüssel /= Schlüssel (für den Himmel)

15. JULI 2004: BLATT 27

5. Statistische Assoziationspaare / Kollokationen (weitgehend nach Manning/Schütze, 151ff)

5.1. Terminologie

Im folgenden Kapitel verwenden wir folgende Begriffe:

Wortpaar

Ein Wortpaar ist ein Paar aus zwei Wörtern bzw. Wortformen.

Bigramm

Ein Bigramm ist ein Wortpaar aus zwei im Text direkt aufeinanderfolgenden Wörtern oder Wortformen.

Assoziationspaar

Ein Assoziationspaar ist ein Paar aus zwei Wörtern bzw. Wortformen, die statistisch assoziiert sind. Weiter unten werden wir verschiedene Maße kennen lernen, um die Assoziiertheit zu berechnen.

Fenster

Im Zusammenhang mit der Extraktion von Wortpaaren aus Korpora verstehe ich unter einem Fenster den Bereich, innerhalb dessen zwei Wörter als Wortpaar angesehen werden. Im Falle der Extraktion von Bigrammen hat das Fenster die Größe zwei - d.h. es werden stets nur Wörter betrachtet, die direkt aufeinander folgen.

Kompositionalität

Semantische Eigenschaft eines Syntagmas: die Bedeutung des gesamten Ausdrucks ergibt sich systematisch aus der Bedeutung der Komponenten. So ist ein *kranker Raucher* ein Raucher der krank ist, ein *starker Raucher* aber ist nicht ein Raucher der stark ist - sondern ein Raucher der sehr viel raucht; d.h. der Ausdruck *starker Raucher* ist nicht voll kompositionell; das Nomen *Raucher* behält aber seinen Bedeutung, weshalb wir hier von einem **semikompositionellen** Ausdruck sprechen. Der Ausdruck *rotes Tuch* in der nicht-wörtlichen Bedeutung ist **nicht kompositionell**, da es sich dabei weder um ein Tuch, noch um einen roten Gegenstand handelt.

Kollokationen

Semikompositionelle Ausdrücke wie *starker Raucher* und Stützverbkonstruktionen wie *Kritik üben* werden als Kollokationen bezeichnet; dabei wird der semantisch nicht verschobene Begriff als Kollokant

(*Raucher*; *Kritik*) der semantisch reduzierte Ausdruck als Kollokat bezeichnet. Der Begriff der Kollokation hat allerdings unterschiedlichste Definitionen und wird teilweise in der Literatur auch für Assoziationspaare allgemein verwendet.

5.2. Übersicht

Statistische Assoziationspaare sind Paare von Wörtern oder Wortformen, die häufig gemeinsam auftreten. Die linguistischen Phänomene, die sich in Assoziationspaaren niederschlagen, sind in folgender Tabelle aufgelistet:

Phänomen	**Wortarten** (V: Verb, N: Nomen, A: Adjektiv, Ad: Adverb)	**Beispiel**
Selektionspräferenzen in freien Syntagmen		
• Verb-Argument	VN	*waschen Wäsche*
• Attribute in NPn	AN	*schöner Ausflug*
Semikompositionelle Bildungen (Kollokationen im linguistischen Sinn):		
• Stützverbkonstruktionen	VN	*üben Kritik*
• Funktionsverbgefüge	VN	*in Gang bringen*
• semantisch reduzierte Attribute in NPn	AN	*starker Raucher*
• Adverbiale		*klipp und klar*
Idiome (nicht-kompositionell)		
• verbal	VN, AN ...	*lesen Leviten; Kopf verdrehen*
• NPn	AN, NN	*rotes Tuch, Hinz Kunz*
• adverbial	AA, AdAd, AN ...	*Jacke wie Hose*
• satzwertig (Sprichwörter)	Alle	*Morgenstund Gold Mund*

Tabelle 1 : Linguistische Phänomene, die Assoziationspaaren zugrunde liegen

Zur Bestimmung von Assoziationspaaren werden Textkorpora statistisch ausgewertet. Dabei sind folgende Parameter relevant:

- Fenstertyp und -größe: Assoziationspaare werden innerhalb eines Fensters im Korpus ermittelt. Das Fenster hängt ab von der verfügbaren Information (z.B.: ist das Korpus geparst? sind Satzgrenzen markiert?) und vom Typ von Assoziationspaaren, die extrahiert werden sollen. Ein Fenster kann z.B. sein:
    - Satz, Teilsatz oder andere Syntagmen (z.B. NP, Verbalphrase);
    - N Wörter nach links bzw. rechts (z.B. 1 Wort nach links zur Ermittlung von Adjektiv-Nomen-Paaren oder 10 Wörter nach links und rechts für Ermittlung von Verb-Nomen-Paaren).
- Was wird ermittelt? Statistische Assoziation zwischen *Grundformen* oder *Vollformen*
    - Für die Ermittlung von Assoziationspaaren von Vollformen ist keine Lemmatisierung notwendig und es können folglich keine Fehler in der Grundformenreduzierung unterlaufen; allerdings ist die benötigte Datenmenge höher.
    - Die Grundformenreduzierung ist vor allem für Lexeme mit zahlreichen verschiedenen Wortformen (v.a. Verben, Adjektive) sinnvoll, da damit verschiedene Formen der selben Konstruktion auf eine Grundform abgebildet werden und die Chance erhöhen, dass auch für seltenere Assoziationspaare genug Daten gefunden werden. Zur Lemmatisierung ist ein Vollformenlexikon oder ein Modul zur morphologischen Analyse erforderlich.
- Korpusgröße (in Wörtern), bzw. Zahl aller extrahieren Bigramme
- Häufigkeit jeder Wortform im Text bzw. in der Wortpaarliste (bzw. bei Ermittlung von Grundformen: Häufigkeit jeder Grundform).
- Häufigkeit jedes Wortpaares im Text bzw. in der Wortpaarliste (Wortpaare können sein: Wortpaare von Wortformen oder Grundformen).

Die meisten statistischen Tests können mit diesen Parametern durchgeführt werden; für einige komplexere Tests kann noch die Varianz und die Verteilung der Wortformen/Lemmata im Korpus eine Rolle spielen.

Hier die Tabelle mit einigen gängigen statistischen Werten. Dabei ist:

- f_i die Häufigkeit der Wortform oder des Lemmas w_i in der Wortpaarliste
- $f_{<i;j>}$ ist die Häufigkeit des Wortpaars $<w_i,w_j>$
- N die Zahl aller Wortpaare

Statistischer Wert	**Formel**	**Anmerkung**
relative Wortpaarfrequenz	$f_{(i;j)}/N$	Je häufiger die Einzelwörter, desto eher ist auch dieser Wert hoch. Ergibt zu hohe Werte für Wortpaare aus häufigen Wörter, auch wenn sie überhaupt nicht assoziiert sind
Mutual Information Allgemein Im Fall von Wortpaaren	$I(A_1;A_2) = \log \frac{p(A_1 \cup A_2)}{p(A_1) * p(A_2)}$ $I(w_1;w_2) = \log \frac{f(w_1,w_2) * N}{f(w_1) * f(w_2)}$	Problematisch für seltene Wortpaare (hier ist der MI-Wert zu hoch); Wortpaare mit Frequenz<3 sollten daher herausgefiltert werden
t-test Allgemein Im Fall von Wortpaaren	$t = \frac{\bar{x}_{<i;j>} - m_{<i;j>}}{\sqrt{\frac{s^2_{<i;j>}}{N}}} -$ $t = \frac{f_{<i;j>} - \frac{f_i * f_j}{N}}{\sqrt{f_{<i;j>}}} -$	Herleitung s..u.
Chi-Quadrat-Test	$\chi^2 = \frac{N * (O_{11} * O_{22} - O_{11}O_{21})^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})}$	
Log-Likelihood		

Tabelle 2 Assoziationsmaße

5.3. Der T-test

Der T-Wert (T-Score) ist ein Maß für die Signifikanz einer Abweichung eines Werts vom Erwartungswert. Im Falle von Assoziationspaaren in einem Korpus misst er die Signifikanz der Abweichung der tatsächlichen Häufigkeit des Wortpaars von der Häufigkeit, die zu erwarten wäre, wenn die beiden Wörter zufällig über das Korpus verteilt wären.

Die folgende Formel ist die allgemeine Form für den T-Test.

- X ist der Mittelwert der Stichprobe
- m ist der Mittelwert der Referenzstichprobe
- s2 ist die Varianz - um den t-Test anwenden zu können, sollte die Varianz in der Stichprobe und der Referenzstichprobe in etwa gleich sein.
- N ist die Größe der Stichprobe.

$$t = \frac{\bar{x}_{<i;j>} - m_{<i;j>}}{\sqrt{\frac{s^2_{<i;j>}}{N}}} -$$

Der t-Wert muss dann anhand einer Tabelle interpretiert werden. Bei einer hohen Zahl N kann folgende Tabelle benutzt werden.

Wahrscheinlichkeit, dass die Abweichung zufällig zustande kommt	0.5	0.1	0.05	0.02	0.01	0.001
T-Test Wert	0.674	1.645	1.960	2.326	2.576	3.291

Tabelle 3 : T-Test-Tabelle

5.3.1 Beispiel - Wortlänge

Nehmen wir an, die durchschnittliche Wortlänge in einem großen Korpus (10 Mio. Wörter) mit Texten aller Art sind 5 Buchstaben. Wir haben nun einen kleinen linguistischen Fachtext aus dem Korpus vor uns. Dieser Text enthält 100 Wörter und die durchschnittliche Wortlänge ist 7 Buchstaben, die Varianz der Wortlänge in diesem Text ist 4,0.

- Unsere Hypothese ist nun folgende: diese Abweichung ist nicht zufällig - linguistische Texte (oder dieser Text) hat eine größere durchschnittliche Wortlänge und die gefundene Abweichung ist tatsächlich signifikant.
- Die Gegenhypothese: der Mittelwert weicht hier nur zufällig ab, und die Abweichung ist im Rahmen des Normalen bei solch einer Stichprobe.

Versuchen wir nun diese Formel für unsere Zwecke zu verwenden:

- X ist der Mittelwert im linguistischen Text, also 7
- m ist der angenommene Mittelwert für die Gegenhypothese, also 5
- s2 ist 4,0
- N ist die Größe der Stichprobe, also 100

Dann ist der T-Wert (7-5)/Wurzel(4/100) = 2/ (1/5) = 10.
Das heißt nun, die Wahrscheinlichkeit, dass diese Abweichung zufällig ist, ist wesentlich kleiner als 0,001 (weniger 1 Promille).

5.3.2 Übertragung auf Assoziationspaare

Wie wenden wir nun den T-Test auf Assoziationspaare an?
Nehmen wir an, wir haben aus einem Korpus 1 Million Wortpaare der Form <w1 w2> extrahiert.
Wir interessieren uns nun für das Wortpaar <wi wj>

- Die Häufigkeit f(wi) ist 100
- Die Häufigkeit f(wj) ist 100
- Die Häufigkeit f(<wi;wj>) ist 10.
- N = 1 Million

Unsere Hypothese: das Wortpaar <wi,wj> hat eine Häufigkeit, die signifikant höher ist, als der Erwartungswert, falls alle Wörter zufällig über das Korpus verteilt wären.

Statistische Vorannahmen

Die durchgeführte Zählung ist ein Bernoulli-Experiment (ein Experiment mit den Ausgängen 0 oder 1).
1 steht für: ein Wortpaar ist das untersuchte Wortpaar <wi,wj>
0 steht für: ein Wortpaar ist nicht das gesuchte Wortpaar.

Was ist nun der Mittelwert für die tatsächliche Verteilung, also x?
Dieser ist $f(<w_i,w_j>)$ (Häufigkeit des Wortpaars) geteilt durch die Menge der Wortpaare N.
$f(<w_i,w_j>) / N$ = 10/1 Million = 1/100 000
(Dieser Wert ist also einfach die relative Häufigkeit h(<wi,wj>); bzw. die Wahrscheinlichkeit, dass ein zufällig herausgegriffenes Wortpaar genau das untersuchte Wortpaar ist)

Was wäre nun der Mittelwert bei einer zufälligen Verteilung? Ganz einfach. Wenn Wort w_i 100 mal im Korpus vorkommt, und w_j ebenfalls, dann ist die Wahrscheinlichkeit, bzw. h(w) jeweils 1/10 000. Die Wahrscheinlichkeit, dass ein Wortpaar aus w_1, w_2 besteht ist also 1/10 000 * 1/10 000, das ist 1/100 Millionen. Die erwartete Häufigkeit in allen Bigrammen (1 Million) wäre also 1/100 - der erwartete Mittelwert somit 1/ 100 Millionen.

Was ist nun die Varianz der Stichprobe?

Die Varianz ist bekanntlich die Summe der Quadrate aller Abweichungen vom Mittelwert / Stichprobengröße, in unserem Fall:
$f(<w_i,w_j>)*(1-h(<w_i,w_j>))^2) + (N-f(<w_i,w_j>))*(h(<w_i,w_j>))^2) / N$

Dies lässt sich umformen zu $h(<w_i,w_j>)*(1- h(<w_i,w_j>))$. Da $(1- h(<w_i,w_j>)$ in etwa etwa 1 ist, ist dies ungefähr $h(<w_i,w_j>)$ - diesen Wert setzen wir also als Varianz ein.
Damit kommen wir auf die folgende Formel

$$t = \frac{h_{<i;j>} - h_i * h_j}{\sqrt{\frac{h_{<i;j>}}{N}}} -$$

Im Ergebnis gibt dies für unsere Beispielzahlen oben ca 3,16 - wir können also (s. Tabelle 3) recht sicher sein, hier ein signifikantes Paar, sprich ein Assoziationspaar, vor uns zu haben.

5.4. Mutual Information

Mutual Information ist ein Maß für die Assoziation zweier Zufallsvariablen aus der Informationstheorie. Die hier vorgestellte Formel berechnet allerdings nur die punktweise Mutual Information (pointwise mutual information) für einen bestimmten Wert einer Zufallsvariablen, d.h. für zwei Elementarereignisse. Die Formel für die MI von zwei Ereignissen A1 und A2 ist die folgende:

$$I(A_1; A_2) = \log \frac{p(A_1 \cup A_2)}{p(A) * p(A_2)}$$

D.h. Die Mutual Information errechnet sich aus zwei Wahrscheinlichkeiten: a) der tatsächlichen Wahrscheinlichkeit dass beide Ereignisse gemeinsam auftreten. b) aus der Wahrscheinlichkeit, die ein gemeinsames Auftreten hätte, gegeben, die beiden Ereignisse sind unabhängig. Der Logarithmus dient dazu, die Werte, die zwischen 0 und unendlich liegen, mit dem neutralen Punkt bei 1, auf eine Skala abzubilden, die ihren neutralen Punkt bei 0 hat, und auf der positive Werte Assoziiertheit bedeuten.
Die MI ist das einfachste und auf Anhieb einsichtigste Maß zur Errechnung von Assoziationspaaren.
Zur Errechnung von Assoziationspaaren nehmen wir folgendes an:

- Es handelt sich um ein Bernouilli-Experiment (s.o.)
- p(w1) = h(w1); d.h. die Wahrscheinlichkeit eines Wortes ist gleich der relativen Häufigkeit
- p(<w1,w2>) = h(<w1,w2>).

Die entsprechenden Werte müssen dann nur noch in die obige Formel eingesetzt werden.

Die MI bringt als Assoziationsmaß die Schwierigkeit mit sich, dass sie die Signifikanz nicht berücksichtigt. Dies ist v.a. für Bigramme mit niedriger Häufigkeit (insbesondere Häufigkeit 1) problematisch.

5.5. Chi-Quadrat-Test

Der Chi-Quadrat-Test ist ein Hypothesentest, der sich besonders gut dazu eignet, erwartete Häufigkeiten mit tatsächlich beobachteten Häufigkeiten zu vergleichen. Er berechnet die Signifikanz der Abweichung einer Menge von Zufallsvariablen von einem hypothetisch angenommenen Wert.

In unserem Fall sind – wie bei den bisher genannten Tests - die tatsächlichen Häufigkeiten die Wortpaarhäufigkeiten im Korpus, während die hypothetischen Häufigkeiten die Bigrammfrequenzen sind, die anzunehmen wären unter der Bedingung, dass die Verteilung der Wörter im Korpus zufällig ist. Die Werte zur Durchführung des Chi-Quadrat-Tests sind die Häufigkeitsauflistungen in einer Kreuztabelle. Im Falle von Wortpaarhäufigkeiten ergibt sich folgende Tabelle mit zwei Zeilen und zwei Spalten der Frequenzen aller möglicher Kombinationen aus zwei Wörtern w1 und w2:

	w1	~w1
w2	$f_{<w1w2>}$	$f_{<\sim w1w2>}$
~w2	$f_{<w1\sim w2>}$	$f_{<\sim w1\sim w2>}$

Tabelle 4 Kreuztabelle der tatsächlichen Frequenzen für eine gegebene Wortkombination

Es ist deutlich, dass die hier verwendeten Werte sich aus den bisher verwendeten Frequenzen von Einzelwörtern und Wortpaaren im t-Test und der Mutual Information ableiten lassen, wenn man berücksichtigt, dass die Summe der ersten Spalte die Frequenz von w1, die Summe der ersten Zeile die Frequenz von w2 ist. Damit lassen sich aus fw1, fw2 und f<w1,w2> sowie N (Zahl der Wortpaare) alle Werte der Tabelle berechnen.
Wesentlich bei der Beurteilung der Ergebnisse des X-Quadrat-Tests sind die so genannten Freiheitsgrade, d.h. die Zahl der Parameter die im angenommenen Bezugsrahmen variieren können. Unter der Annahme, dass N, fw1 und fw2 nicht variieren, ist es in der gegebenen Tabelle nur möglich, einen Parameter zu ändern - die Frequenz $f_{<w1w2>}$ – alle anderen Frequenzen ergeben sich dann in Abhängigkeit aus diesem Wert und den Rahmenbedingungen, d.h. der Beurteilung des Chi-Quadrat-Test-Werts liegt der Freiheitsgrad 1 zu Grunde.

Die Formel für die Berechnung des Chi-Quadrat-Wertes ist nun sehr einfach (vgl. Manning/Schütze 1999: 169):

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Hier sind i und j die Indizes für die Felder in der Tabelle – d.h. es wird über alle Felder in der Kreuztabelle aufsummiert - in unserem Falle für die Werte $f_{<w1w2>}$, $f_{<\sim w1w2>}$, $f_{<w1\sim w2>}$, $f_{<\sim w1\sim w2>}$. Dabei sind die Oij-Werte jeweils die tatsächlich observierten Werte, die Eij-Werte die Erwartungswerte unter der Prämisse, dass das Vorkommen von w1 und w2 unabhängig ist. Diese Erwartungswerte lassen sich nun sehr einfach berechnen: Fest gegeben sind N, fw1 und fw2. Die Wahrscheinlichkeit für das Auftreten von <w1,w2> unter der Prämisse der Unabhängigkeit wurde bereits für die anderen statistischen Werte berechnet und ist $p_{E<w1w2>}$ = fw1/N * fw2/N. Daraus ergibt sich für unser Korpus als angenommene durchschnittliche Häufigkeit für <w1,w2> in einem Korpus der Größe N unter der Prämisse der Unabhängigkeit der Wert f_E<w1,w2> = $p_{E<w1w2>}$* N. Daraus lassen sich nun all anderen Werte der Kreuztabelle für die Hypothese der Unabhängkeit ableiten:

	w1	~w1
w2	$f_{w1} * f_{w2}/N$	$f_{w2}-(f_{w1} * f_{w2)})/N$
~w2	$f_{w1}-(f_{w1} * f_{w2}/N)$	N-fw1-fw2+ $(f_{w1} * f_{w2)}/N$

Tabelle 5 Kreuztabelle der angenommenen Frequenzen für eine gegebene Wortkombination unter der Prämisse der Unabhängigkeit

Da sich die Werte für die angenommene Unabhängkeit der Ergebnisse aus den tatsächlich vorhandenen Werten berechnen lassen – aus dem obigen folgt dass - f_E<w1,w2> = fw1/N * fw2/N.* N – lässt sich nun auch der chi-Quadrat-Wert aus den gegebenen Werten in der Tabelle berechnen. Durch eine Umformung der oben genannten Formel für den Chi-Quadrat-Test erhält man:

$$\chi^2 = \frac{N * (O_{11} * O_{22} - O_{11}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

(vgl. Manning Schütze 1999: 170f).

Mit dem hieraus errechneten Wert muss nun wie beim T-Test aufgrund einer Tabelle die Signifikanz der Abweichung der tatsächlichen Werte von den Erwartungswerten ermittelt werden.

Um Deutlichkeit zu erlangen, wie sich der errechnete Wert verhält ist es sinnvoll sich einige Extrembeispiele zu betrachten, für die eine linguistische Intuition besteht. Bei der Mutual Information

hatten wir bereits gesehen, dass der Wert – entgegegen den Erwartungen, die wir an einen Assoziationswert zweier Wörter stellen – bei gleich bleibender Korpuslänge für ein Wortpaar aus zwei Wörtern, die nur innerhalb des Wortpaars – d.h. in keinem anderen Wortpaar – auftritt mit steigender Frequenz des Wortpaars tatsächlich sinkt. In der Kreuztabelle sind in diesem Falle zwei Werte gleich 0 Wie verhält es sich nun der Chi-Quadrat-Wert in diesem Fall?

	w1	~w1
w2	f<w1,w2>	0
~w2	0	N-f<w1,w2>

Tabelle 6 Kreuztabelle der angenommenen Frequenzen für eine gegebene Wortkombination für ein Wortpaar aus zwei Wörtern, die stets gemeinsam auftreten

Für die Formel ergibt sich damit durch die Einsetzung des Wertes 0 für die Werte folgende Vereinfachung, wenn alle Produkte, die Null ergeben, weggelassen werden:

$$\chi^2 = \frac{N*(O_{11}*O_{22})^2}{(O_{11})(O_{11})(O_{22})(O_{22})}$$

Durch Kürzung ergibt sich der Wert N – d.h. das Ergebnis des Chi-Quadrat-Test ist in diesem Fall völlig unabhängig von der Häufigkeit der Wörter. Dies ist für die Extraktion von Assoziationspaaren ebenfalls nicht optimal – an sich wäre es wünschenswert, dass häufigere Paare in diesem Fall etwas höher gewichtet würden.

5.6. Log-Likelihood

Aufgrund der Unzureichendheiten bei der Berechnung von Assoziiertheit zwischen Wörtern über die Mutual Information, den T-Test, den Chi-Quadrat-Test, und anderen bis dahin vorgeschlagene Assoziationsmaße schlägt Dunning (1993) als Assoziationsmaß die log-likelihood (Log-Wahrscheinlichkeit) vor. Der Ausgangspunkt seiner Argumentation für dieses Assoziationsmaß ist die Tatsache, dass die meisten der anhand von Korpora untersuchten Phänomene sich mit seltenen Ereignissen beschäftigen, die meisten vorgeschlagenen statistischen Maße sich aber nicht dazu eignen, die Signifikanz seltener Ereignisse korrekt zu erfassen, und insbesondere nicht zum Vergleich von seltenen Ereignissen mit häufigeren, was zu einer Überbewertung von Wortpaaren mit geringer Frequenz führt.

Dies ist leicht nachvollziehbar, wenn man die Erörterung der Extremfälle für die Mutual Information und den Chi-Quadrat-Test betrachtet. Hier konnte ja in den vorangehenden Abschnitten bereits gezeigt werden, dass seltene Bigramme tatsächlich einen zu hohen Assoziationswert erhalten.

Der Log-Likehood-Wert vergleicht die Wahrscheinlichkeiten zweier Hypothesen.
Als Datengrundlage des Log-Likelihood-Tests dienen die folgenden Werte – die sich wiederum aus der bereits für den Chi-Quadrat-Test herangezogenen Kreuztabelle ergeben:

- f<w1,w2>: Die Zahl der Wortpaare aus Wort 1 und Wort 2
- f<w1,~w2>: Die Zahl der Wortpaare mit Wort 1 ohne Wort 2
    - (hieraus): fw1 = f<w1,~w2>+ f<w1,w2>
- N: die Zahl aller Wortpaare

In unserem Fall beginnen wir mit der folgenden Annahme: die Zahlen in der Kreuztabelle (s.o. beim Chi-Square-Test) lassen sich wesentlich besser erklären unter der Annahme, dass p(w1|~w2) und p(w1|w2) ungleich sind – das also die Distribution des Wortes w1 nicht von der Tatsache abhängt, ob es vor w2 auftritt oder nicht. Unter der Gegenannahme wäre p(w1|~w2) = p(w1|w2) – d.h. die beiden Wahrscheinlichkeiten wären gleich. Formalisieren wir die beiden Hypothesen:

- **Hypothese 1**: Die Frequenzen der Wortpaare der Art <w1,w2> und <w1,~w2> im Korpus ergeben sich aus der bedingten Wahrscheinlichkeit des Wortes p(w1|w2) und der bedingten Wahrscheinlichkeit p(w1|~w2), wobei beide Wahrscheinlichkeiten unterschiedlich sind.
- **Hypothese 2**: Die Frequenz der Wortpaare ergibt sich aus der allgemeinen Wahrscheinlichkeit des Auftretens p(w1), d.h. p(w1|w2) = p(w1|~w2) und damit =p(w1).

Die Wahrscheinlichkeiten für die tatsächlich beobachteten Frequenzen f<w1,w2>: und f<~w1,w2> lasen sich nun auf der Basis der folgenden Formel für die Wahrscheinlichkeit von Binomialverteilungen[1] für Bernoulli-Experimente errechnen:

$$P(k;n,p) = \binom{n}{k} \bullet p^k \bullet (1-p)^{n-k}$$

Hier ist:

- k :Zahl der Treffer, in unserem Fall, Zahl der Auftreten von w1 (in einem Fall mit bzw. ohne w2)
- n : Zahl der Versuche, in unserem Fall, Zahl der Wortpaare mit bzw. ohne w2
- p : Wahrscheinlichkeit für das Auftreten eines Treffers, in unserem Falle das Auftreten von w1.

[1] Zur Erinnerung: Die Binomialwahrscheinlichkeit ist die Wahrscheinlichkeit, dass ein bestimmtes Ereignis in einer Kette von Bernoulli-Versuchen – also von Zufallsexperiment mit zwei möglichen Ergebnissen – bei n Versuchen k-mal auftritt. In unserem Fall wird die Wahrscheinlichkeit berechnet, dass ein Wort in n Bigrammen k-mal zu finden ist, und dies unter verschiedenen Parameterannahmen für p (der Wahrscheinlichkeit für ein Einzelereignis).

Die Wahrscheinlichkeit, dass w1 genau f(<w1,w2>) mal vor dem Wort w2 auftritt ist dann, unter Verwendung der bedingten Wahrscheinlichkeit für w1, zu errechnen als $P(f_{<w1,w2>},f_{w2},p_{w1|w2})$. Die Wahrscheinlichkeit, dass das w1 genau $f_{<w1,\sim w2>}$ mal in anderen Kontexten auftritt ist: $P(f_{<w1,\sim w2>},f_{\sim w2},pw1|\sim w2)$. Die Wahrscheinlichkeit, dass beide Werte gemeinsam auftreten ist das Produkt beider Wahrscheinlichkeiten.
Der analoge Wert wird nun berechnet unter der Annahme, dass $p_{w1|w2}$ und $p_{w1|\sim w2} = p_{w1}$ sind, d.h. wir berechnen das Produkt, wobei wir als Wahrscheinlichkeit die Wahrscheinlichkeit des Wortes p_{w1} setzen: $P(f_{<w1,w2>},f_{w2},p_{w1}).* P(f_{<w1,\sim w2>},f_{\sim w2},p_{w1})$.

- unter **Hypothese 1** ist die Wahrscheinlichkeit für die beobachteten Frequenzen für w1 mit bzw. ohne w2: P(f(<w1,w2>,fw2,pw1).* P(f(<w1,~w2>,f(~w2),pw1)
- unter **Hypothese 2** ist die Wahrscheinlichkeit für die beobachteten Frequenzen für w1 gegeben w2 bzw. nicht gegeben w2: P(f(<w1,w2>,fw2,pw1|w2) * P(f(<w1,~w2>,f(~w2),pw1|w2).

Der maximale Wert beider Wahrscheinlichkeiten soll nun verglichen werden. Dazu werden die beiden Werte in einem Bruch zueinander in Relation gesetzt – dies ist der Likelihood-Bruch:

$$\lambda = \frac{\max_{p_{w1|w2},p_{w1|\sim w2}} \binom{f_{w2}}{f_{<w1,w2>}} \bullet p_{w1|w2}^{\,f_{<w1,w2>}} \bullet (1-p_{w1|w2})^{f_{w2}-f_{<w1,w2>}} * \binom{f_{\sim w2}}{f_{<w1,\sim w2>}} \bullet p_{w1|\sim w2}^{\,f_{<w1,\sim w2>}} \bullet (1-p_{w1|\sim w2})^{f_{\sim w2}-f_{<w1,\sim w2>}}}{\max p_{w1} \binom{f_{w2}}{f_{<w1,w2>}} \bullet p_{w1}^{\,f_{<w1,w2>}} \bullet (1-p_{w1})^{f_{w2}-f_{<w1,w2>}} * \binom{f_{\sim w2}}{f_{<w1,\sim w2>}} \bullet p_{w1}^{\,f_{<w1,\sim w2>}} \bullet (1-p_{w1})^{f_{\sim w2}-f_{<w1,\sim w2>}}}$$

Nun lässt sich zeigen, dass die maximalen Wahrscheinlichkeiten sich setzen lassen als die tatsächlich beobachteten Wahrscheinlichkeiten (d.h. relativen Häufigkeiten) im gegebenen Korpus. Das heißt, die höchste Wahrscheinlichkeit für die gegebene Verteilung wird dann erzielt, wenn man die tatsächliche Wahrscheinlichkeit bzw. die relative Häufigkeit im Korpus anschaut. Das heißt:

- Die maximal bedingte Wahrscheinlichkeit w1|w2 unter den Daten im gegebenen Korpus - genauer: die Wahrscheinlichkeit des Auftretens von w1 vor w2 errechnet sich folgendermaßen: p(w1|w2) = f<w1,w2> / f(w1).
- Die maximale bedingte Wahrscheinlichkeit w1|~w2, genauer: die bedingte Wahrscheinlichkeit des Auftretens von w1 vor allen anderen Wörtern als w2. Sie ist p(w1|~w2) = f<w1,~w2>/N-f<w1,w2>
- Die maximale bedingte Wahrscheinlichkeit von pw1 ist die Wahrscheinlichkeit im Korpus an sich, dh. hw1

Diese beiden Wahrscheinlichkeiten werden nun über den Likelihood-Bruch in Relation zueinander gesetzt – wir lassen einfach das "max" weg, und nehmen für die bisher hypothetischen Wahrscheinlichkeiten die tatsächlichen Wahrscheinlichkeiten an.
Da die N über X Ausdrücke oben wie unten gleich sind, lässt sich dieser Bruch erfreulicherweise folgendermaßen kürzen:

$$\lambda = \frac{p_{w1|}{}^{f_{<w1,w2>}} \bullet (1-p_{w1})^{f_{w2}-f_{<w1,w2>}} * p_{w1}{}^{f_{<w1,\sim w2>}} \bullet (1-p_{w1})^{f_{\sim w2}-f_{<w1,\sim w2>}}}{p_{w1|w2}{}^{f_{<w1,w2>}} \bullet (1-p_{w1|w2})^{f_{w2}-f_{<w1,w2>}} * p_{w1|\sim w2}{}^{f_{<w1,\sim w2>}} \bullet (1-p_{w1|\sim w2})^{f_{\sim w2}-f_{<w1,\sim w2>}}}$$

Nun haben wir den Wert des Wahrscheinlichkeitsbruches λ berechnet, der wiedergibt, wie viel wahrscheinlicher die NULL-Hypothese als unsere Annahme ist. Tatsächlich ist allerdings der Wert -2logλ interessant, da dieser sich in der Chi-Quadrat-Tabelle nachschlagen lässt, um die Wahrscheinlichkeit der Zufälligkeit der Abweichung nachzuprüfen. Will man diesen Wert berechnen, wird der Bruch durch die Logarithmierung zur Differenz, wir erhalten dann:

$$-2\log\lambda = \log(p_{w1}{}^{f_{<w1,w2>}} * (1-p_{w1})^{f_{w2}-f_{<w1,w2>}} + \log(p_{w1}{}^{f_{<w1,\sim w2>}} * (1-p_{w1})^{f_{\sim w2}-f_{<w1,\sim w2>}})$$
$$-(\log p_{w1|w2}{}^{f_{<w1,w2>}} * (1-p_{w1|w2})^{f_{w2}-f_{<w1,w2>}}) - \log(p_{w1|\sim w2}{}^{f_{<w1,\sim w2>}} * (1-p_{w1|\sim w2})^{f_{\sim w2}-f_{<w1,\sim w2>}})$$

Anzumerken zu diesem Maß:

1. die Werte für seltene Wortpaare sind weniger stark gewichtet. Häufige Wortpaare werden damit stärker gewichtet.
2. Nur für Wortpaare mit einer Häufigkeit < 5 führt dies zu einer entscheidenden Herunterstufung des Werts gegenüber anderen Assoziationsmaßen
3. der Wert wird ebenfalls hoch für statistische disassoziierte Wortpaare – d.h. Wortpaare aus Wörtern, die mit hoher Signifikanz nicht gemeinsam auftreten. Diese müssen mit Hilfe eines Wertes, der Abweichung nach oben von Abweichung nach untern unterscheidet, aussortiert werden
4. denselben Effekt wie durch die Berechnung dieses statistisch etwas aufwendigen Wertes kann man durch eine ad-hoc Reduzierung des Assoziationswerts seltener Bigramme für einfache Assoziationsmaße erreichen

5.7. Weitere Assoziationsmaße

Es gibt noch einige weitere Assoziationsmaße - für eine Übersicht s. Manning-Schütze, Kapitel 5.

6. Evaluation

6.1. Einleitung

Angenommen, wir wollen die im letzten Kapitel diskutierten Assoziationsmaße in Bezug auf ihren Extraktionserfolg verteilen – wie würden wir vorgehen? Wir brauchen Maße, die uns sagen, wie gut die einzelnen Assoziationsmaße in der Lage sind, relevante Wortpaare im Korpus zu entdecken, und irrelevante Wortpaare auszusortieren. Maße, die wir hier verwenden könnten sind Recall und Precision, die im nachfolgenden Abschnitt vorgestellt werden.

6.2. Recall / Precision (Ausbeute / Präzision)

Recall und Precision sind die wichtigsten Maße zur Evaluierung von Text-Retrieval-Systemen, und sind auch für die Evaluation zahlreicher anderer Systeme relevant (etwa automatische Textklassifikation). Exemplarisch wird hier die Verwendung der beiden Werte im Falle von Text-Retrieval-Systemen beschrieben.
Gegeben ist eine Dokumentensammlung D mit einer Zahl von Dokument N_D. Für eine gegebene Anfrage A gibt es eine Untermenge A von Dokumenten, die N_A relevante Dokumenten enthält (diese sind z.B. durch eine manuelle Klassifizierung der Dokumente ermittelt worden).
Das System (das natürlich nicht perfekt funktioniert) findet nun für die Anfrage A eine Dokumentenmenge F, die N_F Dokumente enthält. Unter diesen Dokumenten sind einige relevante Dokumente und einige nichtrelevante.

Recall und Precision sind nun folgende Werte:
Der **Recall** misst, welcher Anteil der tatsächlich relevanten Dokumente tatsächlich gefunden wurde: Er errechnet sich aus der Zahl der tatsächlich relevanten und gefundenen Dokumente, also der Größe der Schnittmenge aus den relevanten und den gefundenen Dokumenten, geteilt durch die Zahl der relevanten Dokumente. Der Recall beantwortet also die Frage: Wie viele von den Dokumenten, die das System hätte finden sollen, hat es denn wirklich gefunden?

Recall $R = N_{A \cap F} / N_A$ (A: relevante für Anfrage; F: gefundene Dokumente)

Die **Precision** ist der Anteil der tatsächlich gefundenen Dokumente die relevant waren (also wieder die Größe der Schnittmenge aus gefundenen und relevanten Dokumenten) an der Zahl aller gefundenen

Dokumente - d.h. sie beantwortet die Frage: wie viele von den Dokumenten die ich gefunden habe, sind den wirklich brauchbar?

Precision: $P = N_{A \cap F} / N_F$ (A: relevante für Anfrage; F: gefundene Dokumente)

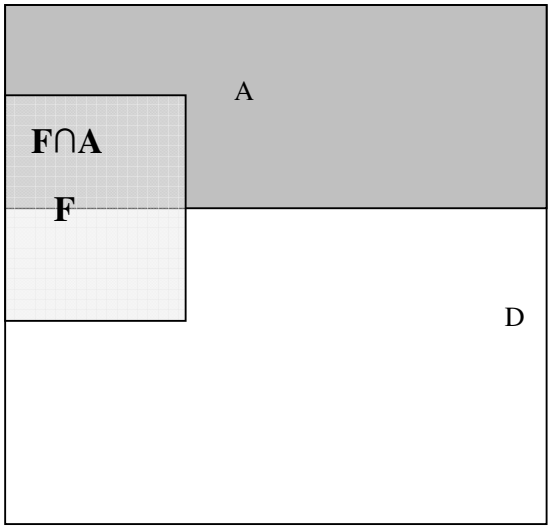

Abbildung 1 Schema zum Verständnis von Recall und Precision. D: Menge aller Objekte. A: Menge aller in Bezug auf eine Aufgabe/Anfrage interessanten Objekte (Objekte, die gefunden werden sollen). F: Menge aller gefundenen Objekte, A∩F: Menge aller gefundenen und interessanten Objekte.

Weder Recall noch Precision allein erlauben es, die Qualität eines Systems zu bestimmen: der Recall kann ganz leicht auf 1 gebracht werden, indem man alle Dokumente findet (und dann in der Flut der Dokumente untergeht), die Precision lässt sich ab besten erhöhen, indem das System extrem restriktiv arbeitet (und dann die meisten interessanten Dokumente nicht findet).
Ein Maß dass Recall und Precision kombiniert, ist das F-Maß

$$F = \frac{1}{\alpha * \frac{1}{P} + (1-\alpha) * \frac{1}{R}}$$

Hier ist P die Precision, R der Recall, und α ein Wert, der zwischen 1 und 0 liegt, und die Gewichtung der beiden Maße angibt – oft wird für α der Wert 0,5 gewählt.

Verrechnung von Recall und Precisionwerten: Mikrobewertung / Makrobewertung

Üblicherweise werden wir für unterschiedliche Anfragen auch unterschiedliche Recall- und Precisionwerte erhalten. Wie geht man nun vor, wenn man den Recall bzw. die Precision für das Gesamtsystem ermitteln will. Hier gibt es prinzipielle zwei verschiedene Möglichkeiten: ermittle die durchschnittlichen Wert über alle Anfragen hinweg (Makrobewertung) bzw. berechne den Wert über die für alle Anfragen gefundenen Dokumente (Mikrobewertung).

	Dokumentenanzahl	Relevant	Gefunden	Gefunden∩Relevant	Precision	Recall
1	D1	A1	F1	A1∩F1	A1∩F1/F1	A1∩F1/A1
2	D2	A2	F2	A2∩F2	A2∩F2/F2	A2∩F2/A2
3	D3	A3	F3	A3∩F3	A3∩F3/F3	A3∩F3/A3
4	…	…	…	…	…	…
n	Dn	An	Fn	An∩Fn	An∩Fn/Fn	An∩Fn/An
S		∑Ai	∑Fi	∑(Ai∩Fi)	∑(Ai∩Fi/Fi)	∑(Ai∩Fi/Ai)

Tabelle 7 Tabelle zur Veranschaulichung der Mikrobewertung und Makrobewertung von Recall und Precision

Mikrobewertung: Rechne alle Dokumentenzahlen zusammen, ermittle dann Precision bzw. Recall. In der Tabelle 7 werden die Summen ∑Ai (Recall) ∑Fi (Precision) und ∑(Ai∩Fi) (beide) aus der Reihe (S) verwendet, und aus diesen Summen der Recall- bzw. Precisionwert berechnet.

Makrobewertung: Berechne den Mittelwert aller Recall-/ bzw. Precisionwerte – d.h. verwende die Summen ∑(Ai∩Fi/Fi) (Precision) bzw. ∑(Ai∩Fi/Ai) (Recall), teile sie jeweils durch i (Zahl aller Wert in der Spalte).

Mikrobewertung und Makrobewertung liefern unterschiedliche Werte, wenn die Zahl der zu findenden Dokumente für die unterschiedlichen Anfragen nicht gleich sind.

Beispiel: Nehmen wir an, wir haben 10 Anfragen. Für neun der Anfragen gibt es jeweils ein relevantes Dokument, diese wird auch gefunden. Recall & Precision sind jeweils 1. Für eine Anfrage gibt es 100 relevante Dokument; es werden nur 10 Dokumente gefunden, davon ist nur eines relevant. Recall ist 1/100, Precision 1/10. Nach der Makrobewertung ergibt sich der durchschnittliche Recallwert von 0,901, und die Precision 0,91. Nach der Mikrobewertung gibt sich ein Recall von 1/11, eine Precision von 10/19.

6.3. Übung

Eine Textsammlung hat Texte der Kategorien *Kochrezepte - Wetterberichte- Andere*.

Eine Stichprobe von 200 Texten wird von Hand klassifiziert, mit folgendem Ergebnis:

- 50 Kochrezepte
- 50 Wetterberichte
- 100 Andere

Ein Textklassifikationssystem klassifiziert nun folgendermaßen:

- 70 Kochrezepte (davon 40 wirkliche, 10 Wetterberichte, 20 Andere)
- 10 Wetterberichte (davon alle 10 korrekt)
- 120 Andere (davon 80 Andere, 30 Wetterberichte, 10 Kochrezept)

Was ist der Recall und die Präzision für jede Kategorie (Kochrezepte, Wetterberichte, Andere).

Was ist der Gesamtwert für den Recall und die Precision für alle Kategorien (Makro- und Mikrobewertung).

7. N-Gramm-Modelle

(Dieses Kapitel basiert v.a. auf Manning/Schütze, Kap 6)

7.1. Ziele und Anwendungen

Ziel von N-Gramm-Modellen über Wortformen ist die Berechnung der Wahrscheinlichkeit einer Wortform aufgrund des Kontextes – üblicherweise N Wörtern des linken Kontextes:

Ich trinke ein	*Bier*
	Weißbier
	Wasser
	Cola
	…

Anwendungen für N-Gramm-Modelle:

- Spracherkennungssysteme (automatische Umsetzung gesprochener in geschriebene Sprache) – hier werden N-Gramm-Modell eingesetzt um unter verschiedenen Kandidatenwörtern, die durch die phonetische Komponente erkannt wurden, eine Auswahl zu treffen;
- OCR (optical character recognition – Umsetzung gedruckter, eingescannter Texte in elektronischen Text);
- Kommunikationshilfen - Wortvorschlagssysteme (z.B. bei Sprachbehinderungen) – Systeme, die aufgrund der bisher eingegebenen Teiläußerung mögliche Fortsetzungen vorschlagen, um die Eingabe zu beschleunigen;

7.2. Algorithmen

Die Grundidee bei der Anwendung von N-Gramm-Modellen ist die Vorhersage der nächsten Wortform aufgrund ihrer Wahrscheinlichkeit, die auf Basis einer Anzahl von vorausgehenden Wörtern berechnet wird. Wenn man (inklusive des zu bestimmenden Wortes) N Wörter zur Berechnung heranzieht, spricht man von einem N-Gramm-Modell: Wird nur ein Wort des linken Kontextes herangezogen, spricht man also von einem Bigrammmodell, sind es zwei, hat man ein Trigrammmodell. Die einfachsten Algorithmen verwenden ausschließlich ein N-Gramm-Modell, komplexere Algorithmen verrechnen z.T. verschiedene Werte.

Ein N-Gramm-Modell muss zunächst trainiert werden, d.h. vor dem Einsatz müssen die Werte ermittelt werden, die später zur Wortvorhersage verwendet werden sollen. Bei Wortvorschlagssystemen in Kommunikationshilfen ist es üblich, dass das System bei der Benutzung das Sprachmodell aktualisiert, d.h. dazulernt.

7.2.1 Einfache N-Grammmodelle

Für N-Gramm-Modelle wird die bedingte Wahrscheinlichkeit eines N-Gramms, gegeben die Wahrscheinlichkeit der n-1 Wörter zuvor berechnet. Beide Wahrscheinlichkeiten entsprechen wieder den relativen Häufigkeiten. Da die Wahrscheinlichkeiten sich aus der Gesamtzahl aller N-Gramme M im Nenner errechnen (d.h. p(N-Gramm) = $c_{N\text{-}Gramm}/M$) kann man die Wahrscheinlichkeit aus der Zahl der N und N-1-Gramme berechnen (c ist hier die Anzahl der N-Gramme):

$$p(w_n \mid w_1..w_{n-1}) = \frac{p(w_1..w_n)}{p(w_1..w_{n-1})} = \frac{c(w_1..w_n)}{c(w_1..w_{n-1})}$$

Will man nur das wahrscheinlichste Wort berechnen (arg_{max}) - d.h. $c(w_1 \ldots w_{n-1})$ ist konstant – reicht es aus den Wert $c(w_1 \ldots w_n)$ heranzuziehen (d.h. das häufigsten N-Gramm gibt die beste Vorhersage für das letzte Wort den N-Gramms, wenn die Wörter zuvor bereits bekannt sind).
Probleme: Wenn wir die Wahrscheinlichkeit eines Wortes im Bigramm,- Trigramm- oder 4.Gramm-Modell berechnen, stoßen wir auf das Problem, dass viele Trigramme und noch mehr 4-Gramme noch nicht beobachtet wurden, d.h. sehr viele Wörter erhalten in die Wahrscheinlichkeit 0 zugewiesen. Dies ist aber nicht unbedingt wünschenswert, da auch noch nicht beobachtete N-Gramme eine gewisse Wahrscheinlichkeit zugewiesen bekommen sollten. Prinzipiell gibt es zur Lösung dieses Problems zwei Möglichkeiten:

1. Wahrscheinlichkeiten werden auch für nicht beobachtete N-Gramme vergeben
2. N-1, N-2 etc-Gramme miteinbeziehen
    a. Verrechnung der Wahrscheinlichkeiten
    b. Verwendung des Modell mit dem größten N

Wahrscheinlichkeiten an nicht beobachtete N-Gramme

P_{Lap} (Laplace-Wahrscheinlichkeit)– Zur Häufigkeit jedes N-Gramms wird 1 hinzugezählt; d.h. nicht beobachtete N-Gramme erhalten die Häufigkeit 1. eins. Die Wahrscheinlichkeit eines N-Gramms wird folgendermaßen berechnet:

$$p_{lap}(w_1..w_n) = \frac{c(w_1..w_n)+1}{M+B}$$

Hier ist M die Gesamtzahl aller Bigramme (Summe aller c), B die Zahl aller unterschiedlichen Bigramme. Das Problem ist hier, die zu hohe Wahrscheinlichkeiten für noch nicht beobachtete N-Gramme

P_{Lid} (Lidstone) Dies ist dasselbe Verfahren, nur wird nicht 1, sondern ein Faktor λ hinzugezählt, der z.B. ½ betragen kann:

$$p_{lid}(w_1..w_n) = \frac{c(w_1..w_n) + \lambda}{M + \lambda B}$$

Kombinationen der Wahrscheinlichkeiten für unterschiedliche N

Für die Kombination von einfachen Wahrscheinlichkeiten, Bigrammen, Trigrammen etc. gibt es prinzipielle zwei verschiedene Möglichkeiten:
verwende alle Modelle gleichzeitig und verrechne die Wahrscheinlichkeiten;
verwende das Modell mit dem höchsten N, dass noch vernünftige Ergebnisse liefern könnte (d.h. wo die Frequenz hoch genug ist).
Zunächst eine Formel zur Verrechnung verschiedener Modelle:

$$p_{comb}(w_n \mid w_1..w_{n-1}) = \lambda_1 p(w_n) + ... + \lambda_{n-1} p(w_n \mid w_{2..}w_{n-1}) + \lambda_n p(w_n \mid w_{1..}w_{n-1})$$

dabei gilt $\sum\lambda_i = 1$

Die Werte für λ geben das Gewicht der Modelle für verschiedene Ns an.

Backing off

(Manning-Schütze 219f).
Die Grundidee hier ist, dass ein N-Gramm-Modell mit höherem N nur verwendet wird, wenn genügend Daten (d.h. genügend N-Gramme) zur Verfügung stehen. Sonst wird das Modell mit N-1 gewählt. Die größte Schwierigkeit hier ist es, die Wahrscheinlichkeiten so zu verteilen, dass bei der Gesamtwahrscheinlichkeit für alle Fortsetzungen wieder 1 herauskommt.

7.3. Evaluierung von N-Gramm-Modellen

Trainingsdaten & Testdaten

N-Gramm Modelle werden auf einem Textkorpus trainiert. Um die Vorhersageleistung beurteilen zu können, brauchen wir zusätzlich ein Testkorpus. Testkorpus und Trainingskorpus dürfen nicht identisch sein – es ist augenscheinlich, dass beim Testen auf dem Trainingskorpus mit völlig unzureichenden Algorithmen gute Resultate erzielt werden können.

Zur Verfügung stehende Daten müssen also in Trainingsdaten und Testdaten unterteilt werden, wobei die Testdaten in der Regel den wesentlich geringeren Anteil ausmachen (ca. 5-10%).
Für einige Algorithmen werden die Trainingsdaten nochmals in zwei Untermengen unterteilt: die Daten für das anfängliche Training und die Daten für die Feinjustierung des Modells.

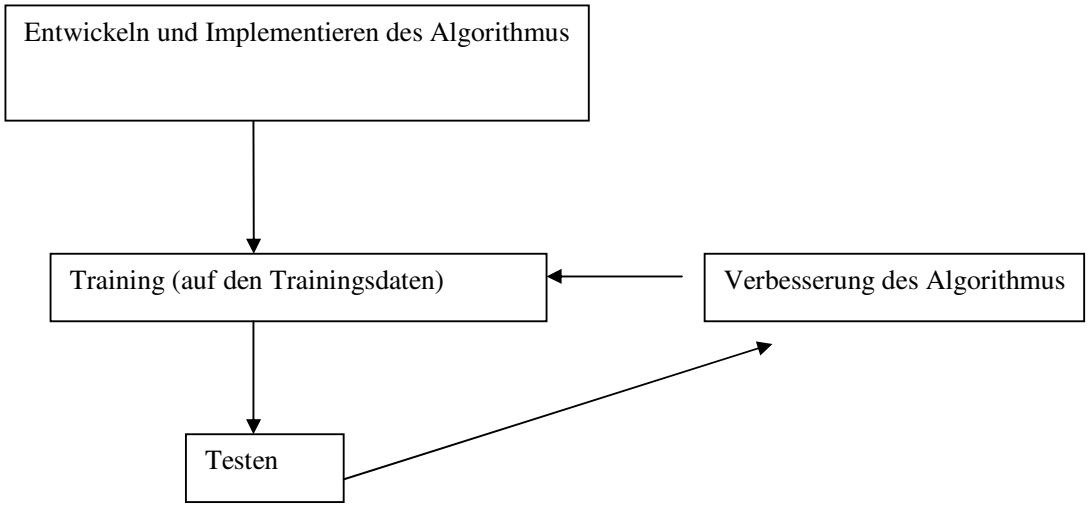

Abbildung 2 Vorgehen für die Verwendung eines N-Gramm-Sprachmodells

8. Wortbedeutungsdisambiguierung

(Manning/Schütze S. 229ff)

8.1. Einleitung

In diesem Kapitel geht es um statistische Methoden zur Feststellung der konkret vorliegenden Bedeutung eines mehrdeutigen (ambigen) Wortes.

8.1.1 Ambiguität und Disambiguierung

Ambiguität bedeutet Mehrdeutigkeit. Ambige Wörter sind solche, für die in einem Wörterbuch mehrere unterschiedliche Bedeutungsbeschreibungen eingetragen werden müssen.
In zahlreichen theoretischen Arbeiten zur Wortbedeutung unterschieden zwischen Homonymie (unterschiedliche Wörter haben dieselbe Form; die Bedeutungen liegen soweit auseinander, dass man von unterschiedlichen Wörtern spricht) und Polysemie (ein Wort hat unterschiedliche Bedeutungen; dies ist der Fall, falls man die Bedeutungen noch aufeinander beziehen kann).
Bei Homonymen ist noch zu unterscheiden zwischen partiellen Homonymen (nicht alle Formen eines Wortes sind gleich) und vollständigen Homonymen (alle Formen sind gleich).

Beispiele

- *Bank* (GELDINSTITUT, SITZGELEGENHEIT) – partielle Homonyme (*Bank-Banken*… - *Bank – Bänke*).
- *Er hat zuhause eine schwarze **Maus** (*polysem – COMPUTERMAUS / TIER)
- *Peter war schon wieder **schlecht**.* (syntaktische+semantische Ambiguität - Peter hat schlechte Leistung hervorgebracht. Peter spürte Übelkeit.)
- *Essen macht Spaß.* (syntaktische Ambiguität: *essen* – Verb, *Essen* – Nomen)

Unter **Disambiguierung** versteht man die Entscheidung, welche Bedeutung das ambige Wort in der konkreten Äußerung (also in einem konkreten gegebenen Kontext) hat.
Wir nehmen an, dass Wörter eine endliche Anzahl von diskreten Bedeutungen haben (die wir im günstigsten Fall in einem Lexikon, Thesaurus oder einer anderen Quelle finden können).
In den meisten Fällen ist es aber unklar, wo die Grenze zwischen jeweiligen Bedeutungen liegt und wie fein man sie definieren kann / soll.

Beispiele

Titel

ehrenvoller, durch eine Prüfung erworbener od. für Verdienste verliehener Zusatz zum Namen (Doktortitel);

Standesbezeichnung (Grafentitel, Herzogstitel);

Überschrift eines literar. od. musikal. Werkes (Buchtitel, Filmtitel, Operntitel);

Name od. Ziffer des Abschnitts eines Gesetzes, des Haushaltsplans usw.;

Abbildung 3 Auszug aus Wahrig Deutsches Wörterbuch

Titel

1 *Bezeichnung des Ranges einer Person;* einen Titel führen; jmdn. mit seinem Titel anreden

1.1 *ehrenvoller, durch eine Prüfung erworbener od. für Verdienste verliehener Zusatz zum Namen;* DoktorTitel; akademischer Titel

1.2 *durch Geburt erworbene Bezeichnung des Ranges als Zusatz zum Namen;* GrafenTitel; HerzogsTitel
1.3 *Amtsbezeichnung,* z.B. Regierender Bürgermeister

1.4 *in sportl. Wettkämpfen errungene Bezeichnung des Ranges;* Titel des Weltmeisters im Boxen

2 *kennzeichnender Name eines Buches od. Kunstwerkes;* BuchTitel; FilmTitel; OpernTitel

3 *Name od. Ziffer des Abschnitts eines Gesetzes od. einer Verordnung;* diese Mittel sind unter Titel 5 des Haushaltsplanes ausgewiesen

Abbildung 4 Auszug aus Wahrig Fremdwörterlexikon

Aus diesen Beispielen sehen wir, dass die Unterscheidung zwischen einzelnen Bedeutungen eines Wortes sogar innerhalb eines Lexikon-Verlages sehr stark variiert.

8.1.2 Anwendungen:

Maschinelle Übersetzung:

dt. Bank -> engl. bank

dt. Bank -> engl. bench

8.2. Methodologie - Begriffe

Überwachtes Lernen – supervised learning

Die Trainingsdaten sind klassifiziert / markiert / bearbeitet, bzw. wir verfügen über andere Informationsquellen (Lexika, Thesauri, allignierte Korpora)

Unüberwachtes Lernen – unsupervised learning

Die Trainingsdaten sind nicht vorklassifiziert

Pseudowörter

Wenn wir zu Evaluation eines Systems nicht über ausreichend manuell disambiguierte Testkorpora verfügen, können solche Testkorpora automatisch generiert werden, indem mehrere (meist zwei) nicht-ambige Wörter als eine sogenanntes Pseudowort zusammengefasst werden, dass dann natürlich mehrdeutig ist – etwa Erdbeere+Fahrrad zum mehrdeutigen Pseudowort Erdbeere-Fahrrad. Da uns ja bekannt ist, welches der beiden künstlich zusammengefassten Wörter ursprünglich im Text stand, können wir das Wortbedeutungssystem dadurch evaluieren, dass es das ursprüngliche Wort (d.h. die zugrunde liegende Bedeutung) herausfinden muss.

Upper bound – obere Grenze

Maximale Leistung, optimale Lösung für ein Problem (meistens von einem Menschen) – diese Grenze gibt an, was eine Wortbedeutungssystem maximal leisten kann.

Lower bound (baseline) – untere Grenze

Minimale Leistung (meisten von dem einfachsten Algoritmus) – dies ist die Leistung, die ein Algorithmus bringen würde, der Wortbedeutungen nach dem Zufallprinzip zuweist

8.2.1 Supervised Disambiguation – überwachte Disambiguierung

Bayesian classification – Klassifizierung nach Bayes

Setzt voraus, dass wir ein Trainingskorpus haben, wo alle Vorkommen des ambigen Wortes korrekt disambiguiert sind

Das Wort wird gemeinsam mit seinem Kontext betrachtet. D .h. es werden zur Disambiguierung Wörter herangezogen, die in der Umgebung des ambigen Wortes innerhalb eines bestimmten Fensters auftreten. Ein solches Fenster kann ein Teilsatz, ein Satz oder ein Abschnitt sein.

Bayes decision rule – Entscheidungsregel von Bayes

- entscheide s' wenn $P(s'|c) > P(sk|c)$ für $sk \neq s'$

Das bedeutet, dass für das Wort w die Bedeutung s gewählt wird, die mit der größten Wahrscheinlichkeit im Kontext c auftritt

Naive Annahme von Bayes

- $P(c|sk) = P(\{vj \mid vj \text{ in } c\} \mid sk) = \prod_{vj \text{ in } c} P(vj \mid sk)$

Wir nehmen an, dass

1. die Reihenfolge und die Struktur der Wörter, die den Kontext bilden, ohne Bedeutung ist (bag of words model) (es ist keine Menge, weil wir das Mehrfachauftreten zulassen)
2. das Auftreten von jedem Wort im Kontext völlig unabhängig von allen anderen Wörtern, die den Kontext bilden, ist

Wir bekommen dann folgende Entscheidungsregel:

entscheide dich für die Bedeutung s' wenn $s' = \arg\max [\log P(sk) + \sum_{vj \text{ in } c} \log P(vj \mid sk)]$

Dabei ist

$P(vj \mid sk) = C(vj, sk) / C(sk)$

$P(sk) = C(sk) / C(w)$

Abkürzungen:

w	ein ambigues Wort
s1,…,sk,…sK	Bedeutungen des Wortes w
c1,…,ci,…,cI	Kontext von w im Korpus
v1,…,vj,…,vJ	Wörter, die den Kontext bilden
C(vj, sk)	Anzahl von vj im Kontext von sk im Trainingskorpus
C(sk)	Anzahl von sk im Trainingskorpus
C(w)	Anzahl von w

9. Clustern und Klassifizieren

9.1. Einleitung

Unter Klassifikation wird die Zuordnung vordefinierter Klassifikatoren (etwa syntaktische Kategorien aus einer Grammatik) zu Objekten verstanden - in unserem Fall zu sprachlichen Einheiten. D.h. die Kategorien, zu denen einzelne Objekt zugeordnet werden sollen, sind vordefiniert. Im Gegensatz dazu werden beim Clustering keine Kategorien vorgegeben, sondern die Kategorien entstehen erst durch den Gruppierungsprozeß. Es ist allerdings möglich (und oft notwendig) mögliche Menge der Cluster und evt. Größen der Cluster zu bestimmen. Beim automatischen statistischen Clustering geht es um die Einteilung von sprachlichen Objekten in Gruppen von ähnlichen Objekten.
Objekte können sein:

- Wörter
- Sätze
- Texte

Anwendung können wir uns etwa die automatische semantische Einteilung von Wörtern durch die Hinzuziehung ihres Kontextes vorstellen. Im Folgenden geht exemplarisch um das Clustering von Wörtern - d.h. die Zusammenfassung von Wörtern zu Gruppen von ähnlichen Wörtern. Als Merkmale - d.h. Einteilungskriterium - für das Clustering kommt in erster Linie die Distribution der Wörter (bzw. Wortformen) im Kontext in Frage - d.h. meist andere Wörter die in der Textumgebung der zu klassifizierenden Wörter auftauchen. Ebenfalls in Frage kommt die Distribution der Wörter in einer Sammlung von Texten.

Zum Clustern von Objekten müssen wir zwei Voraussetzungen erfüllen:

1. wir müssen ein Ähnlichkeitsmaß zwischen Objekten definieren, d.h. ein Maß, wie sehr sich zwei Objekte gleichen (etwa Wörter in einem Korpus, oder auch Dokument im Information Retrieval).
2. Wir müssen Algorithmen definieren, die Aufgrund der Ähnlichkeit zwischen Objekten, eine größere Menge von Objekten in Gruppen von ähnlichen Objekten (Cluster) einteilt.

9.2. Ähnlichkeitsmaße

9.2.1 Vektoren und Vektorähnlichkeit

(Manning-Schütze 296ff, Kap. 8.5.1 f)

Vektoren können herangezogen werden, um den Kontext eines Wortes oder den Inhalt eines Dokuments zu beschreiben.
Ein Vektor ist eine geordnete Menge von numerischen Werten mit N-Dimensionen. Um zwei Vektoren vergleichen zu können, sollten sie dieselbe Anzahl von Dimensionen besitzen.
Folgendes etwa ist ein 3-dimensionaler Vektor

$$\vec{x} = \begin{pmatrix} 7 \\ 3 \\ 0 \end{pmatrix}$$

Wie man sieht, wird ein Vektor durch einen Buchstaben mit einem Pfeil darüber notiert.

Spezialfall eines Vekotrs sind die sogenannten binären Vektoren – sie sind eine geordnet Folge von Nullen und Einsen (d.h. die Wert können nur 0 und 1 sein):

$$\vec{x} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

Als Vektor lässt sich nun etwa die Umgebung eines Wortes darstellen. Nehmen wir etwa folgende Matrix, die die Wörter angibt, die im Kontext von *Herschel, Jupiter, Saturn* und *Uranus* auftauchen:

Wort/Kontext	*umkreisen*	*entdecken*	*Durchmesser*	*leuchten*	*lachen*
Herschel	0	5	0	0	15
Jupiter	5	0	4	11	0
Saturn	3	0	5	12	0
Uranus	4	2	4	10	0

Abbildung 5 Kontextmatrix

Den Kontext von *Herschel* könnte man nun etwa durch folgenden Vektor darstellen:

$$\vec{x} = \begin{pmatrix} 0 \\ 5 \\ 0 \\ 0 \\ 15 \end{pmatrix}$$

Anstatt der Häufigkeiten der Kontextwörter könnte man auch ihre bedingten Wahrscheinlichkeiten eintragen. In einen binären Vektor könnte man diesen Vektor umformen, indem alle Werte >0 gleich 1 setzt (also nur kodiert, ob eine Wort im Kontext auftritt oder nicht.
Folgende Begriff werden noch benötigt:

Länge eines Vektors

$$|\vec{x}| = \sqrt{\sum_{i=1}^{n} x_i^2}$$

(notiert mit zwei senkrechten Strichen): Die Wurzel aus den summierten Quadraten der Werte im Vektor.

Normalisierter Vektor

Ein Vektor mit der Länge 1.

Skalarprodukt (engl. dot product)

$$\vec{x} \cdot \vec{y} = \sum_{i=1}^{n} x_i y_i$$

Für zwei Vektoren lassen sich nun unterschiedliche Ähnlichkeitsmaße definieren, die in folgender Tabelle aufgelistet sind.

Maß	Definition	Einschränkungen	Anmerkungen
Einfache Übereinstimmung	$\vec{x} \cap \vec{y}$	nur binäre Vektoren	Summe aller Einträge in beiden Vektoren, die 1 sind.
Dice-Koeffizient	$\frac{2\lvert\vec{x} \cap \vec{y}\rvert}{\lvert\vec{x}\rvert + \lvert\vec{y}\rvert}$	nur binäre Vektoren	Länge der Schnittmenge beider Vektoren durch die Summe der Länge beider Vektoren
Jaccard-Koeffiziernt	$\frac{\lvert\vec{x} \cap \vec{y}\rvert}{\lvert\vec{x} \cup \vec{y}\rvert}$	nur binäre Vektoren	
Euklidische Distanz	$\lvert\vec{x} - \vec{y}\rvert = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$		Distanzmaß!

Manhattan-Metrik, L1-Metrik	$\sum_{i=1}^{n} \lvert x_i - y_i \rvert$	nach Manning-Schütze v.a. für Vektoren, die bedingte Wahrscheinlichkeiten enthalten	Distanzmaß! Senkrechte Striche sind hier Betragsstriche
Cosinus	$\frac{\vec{x} \cdot \vec{y}}{\lvert\vec{x}\rVert\vec{y}\rvert}$		Distanzmaß! Sind die Vektoren normalisiert (Länge 1) ist dieser Wert einfach das Skalarprodukt

9.2.2 Andere Distanz/Ähnlichkeitsmaße

Neben Vektorähnlichkeitsmaßen können wir für die Ähnlichkeiten zweier Reihen in einer Kookkurenzmatrix wie oben auch probabilistische Maße verwenden.
Dazu wird zunächst für jedes Kontextwort die bedingte Wahrscheinlichkeit gegeben das zu klassifizierende Wort errechnet (diese ist gleich der Häufigkeit des Kontextwortes geteilt durch die aufsummierten Häufigkeiten aller Kontextwörter).
Als Ähnlichkeitsmaße werden in Manning-Schütze nun noch genannt (303f.):

Distanzmaß	Definition	Einschränkungen	Anmerkungen
Kullback-Leibler (KL)-Distanz	$D(p \parallel q) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right)$	nicht symetrisch	
Informationsradius	$D(p \parallel \frac{p+q}{2}) + D(q \parallel \frac{p+q}{2})$		KL-Distanz vom Mittelwert beider Wahrscheinlichkeiten

9.3. Clustering

9.3.1 Typen von Clustering

Hierarchisches Clustering

Es wird eine Hierarchie von Clustern erzeugt. Je nach Vorgehensweise wird ein erstes Cluster, das alle Objekte enthält über mehrere Schritte hinweg in kleinere Cluster zerlegt (top-down) oder zahlreiche kleine Cluster, die anfangs nur 1 Objekt enthalten, werden zu immer größeren Clustern zusammengesetzt.

Nicht-hierarchisches Clustering

Die Objekte werden direkt durch geeignet Algorithmen einer Zahl n von Clustern zugeteilt.

Hartes vs. weiches Clustering

Beim harten Clustering gehört jedes Objekt entweder in ein Cluster oder nicht. Es gibt keine graduierbare Gruppenzugehörigkeit. Beim weichen Clustering kann jedes Objekt einem Cluster mehr oder weniger angehören.

9.3.2 Der K-means Clustering-Algorithmus

Voraussetzung: Anzahl der zu bildenden Cluster; anfängliche Cluster-Zentren (beliebig gewählt); Ähnlichkeitsmaße zwischen Objekten (meist Euklidische Distanz oder Manhattan-Metrik)
Algorithmus:

1. Weise jedes Objekt dem ähnlichsten Clusterzentrum zu;
2. Berechne den neuen Mittelwert für jedes Cluster (Mittelwert für jede Position in den Vektoren)
3. Wiederhole die Prozedur, bis sich nichts mehr ändert.

9.3.3 Der EM-Algorithmus zum "weichen" Clustern

Exkurs Normalverteilung (a. Gaußsche Verteilung)

Zum Verständnis der Anwenung des EM-Algorithmus im Clustering ist es nötig, die sog. Gaußsche Verteilung zu kennen, die auch Normalverteilung genannt wird.. Sie ist symmetrisch, unimodal (hat nur ein Maximum) und glockenförmig. Sie ist deswegen bedeutsam, weil viele Variablen in Zufallsprozessen

zur Normalverteilung neigen - d.h. es gibt eine Häufung beim Mittelwert, und ober- bzw. unterhalb des Mittelwerts fallen die Werte so ab, dass die Kurve zunächst steiler, dann wieder flacher wird.
Die Formel für die Normalverteilung einer Dichtekurve mit Normalverteilung ist:

$$f(x \mid \sigma, \mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2)$$

In dieser Formel ist: x der Wert, für den die Wahrscheinlichkeit berechnet werden soll; sigma (σ) die Varianz, μ (mü) der Mittelwert.

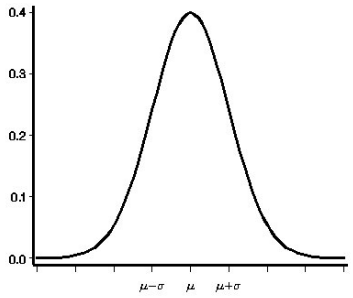

Abbildung 6 Die Kurve einer Normalverteilung

Aufgrund der Form wird diese Kurve auch als Glockenkurve bezeichnet.

Beispiel für Normalverteilungen: Ein Beispiel ist die Zahl der Wappen/Köpfe bei N Münzwürfen. Für größere N nähert sich die Kurve der Wahrscheinlichkeiten für die einzelnen Häufigkeiten der Normalverteilung an. Nehmen wir an, wir werfen eine Münze 100-mal. Wenn man nun die Wahrscheinlichkeiten aufgrund der Binomialverteilung berechnet, stellen wir fest, dass das Balkendiagramm seinen Scheitelpunkt bei 50-mal Kopf hat. Dies sei nun unser Mittelwert für die Berechnung der Normalverteilungskurve. Die Standardabweichung lässt sich ebenfalls aus den Werten der Binomialverteilung ableiten und beträgt in unserem Fall 5. Betrachtet man die beiden Graphiken unten stellt man fest, dass die Kurve der Normalverteilung der Biniomialverteilung sehr stark ähnelt.

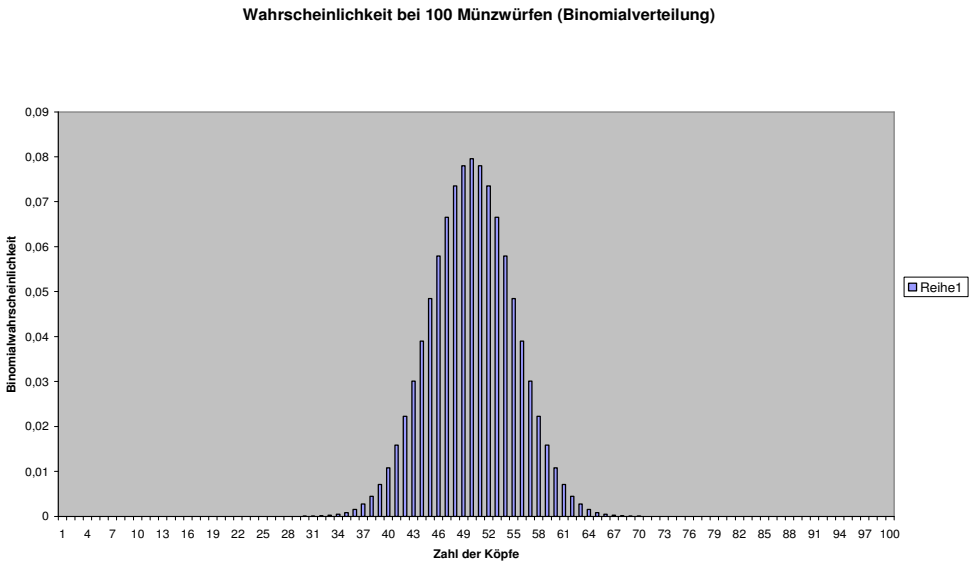

Abbildung 7 Binomialverteilung für die Zahl der Köpfe bei 100 Münzwürfen (faire Münze) – deutlich ist die Annährung der Kurve an die Kurve einer Normalverteilung (s.u.)

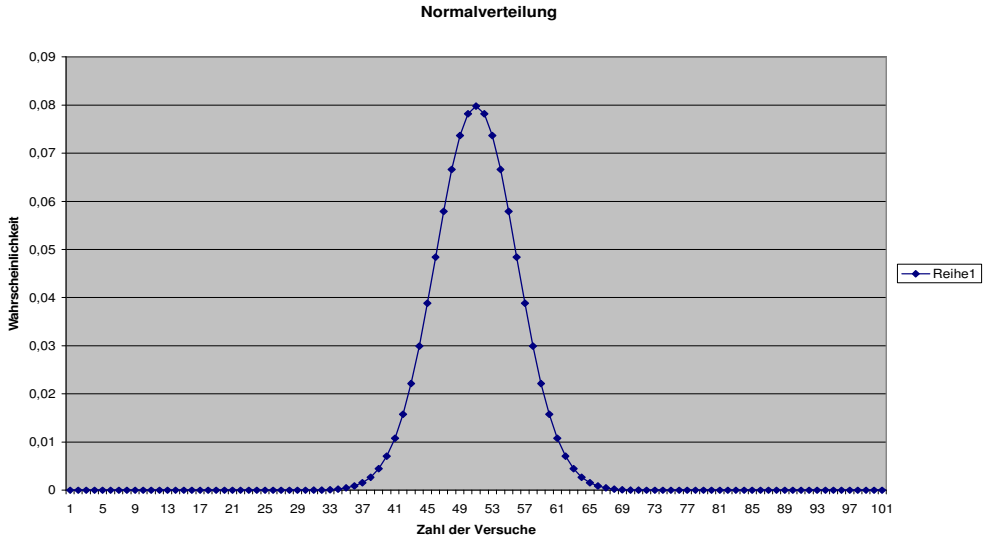

Abbildung 8 Normalverteilung für den Mittelwert 50, und die Standardabweichung 5 (entspricht obigem Münzwurfexperiment).

Fortsetzung EM-Algorithmus

Doch nun zurück zum EM-Algorithmus

Voraussetzung für Ihn ist wieder eine Matrix von Kookkurenzen, wobei diese normalisiert sein sollten. Annahme ist, dass diese Matrix – d.i. eine Menge von Vektoren X={ $\vec{x}_1$... $\vec{x}_n$} von M zugrunde liegenden unbekannten probablistischen Prozessen P erzeugt werden, die normalverteilte Daten erzeugen. Die Erzeugungswahrscheinlichkeit eines Vektors $\vec{x}$ durch einen Prozess P_j ist die Wahrscheinlichkeit der Clusterzugehörigkeit des angenommenen Clusters C_j. – d.h. jeder Prozess entspricht, salopp gesprochen, einem Clustererzeugungsmechanismus; die Zahl der Prozesse (M) ist gleich die Zahl der angenommenen Cluster.

Ziel des EM-Algorithmus ist es, die Parameter der angenommenen Prozesse zu bestimmen, d.h. den Mittelwert μ und die Varianz σ.

Dazu muss zunächst ausgerechnet werden können, mit welcher Wahrscheinlichkeit die gegebenen Daten (d.h. die Vektoren) durch diese Prozesse generiert werden. Hierzu verwenden wir die oben diskutierte Wahrscheinlichkeit eines Wertes aufgrund der Normalverteilung, wobei wir in diesem Fall nicht auf einzelnen Werten x_i eines Vektors operieren, sondern über den gesamten Vektor arbeiten. Die Formel für die Wahrscheinlichkeit von Vektoren aufgrund der Normalverteilung ist wie folgt:

$$n(\vec{x},\vec{\mu},\Sigma) = \frac{1}{\sqrt{2\pi^m|\Sigma|}}\exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})\right]$$

Hier ist:

$\vec{x}$ der Vektor, für den die Realisierungswahrscheinlichkeit aufgrund der anderen Parameter berechnet werden soll;

$\vec{\mu}$ der Mittelwert-Vektor; der Mittelwert von Vektoren wird berechnet, indem für jede Dimension der Vektoren das arithmetische Mittel berechnet wird;

Σ die Kovarianzmatrix, diese kodiert die Abweichungen aller Werte aller Vektoren vom Mittelwert-Vektor;

?^{T} ist die Transposition einer xy Matrix; d.h. die Vertauschung der Dimensionen x und y.

π ist die Konstante PI (3,141…);

Berechnung der Wahrscheinlichkeiten unserer Daten aufgrund des gegebenen Modells

Mithilfe dieser Formeln können wir jeden Prozesse P_j und alle Vektoren x_i die Wahrscheinlichkeit errechnen, dass dieser Vektor von diesem Prozess generiert wird.

Wenn wir nun fürjeden Prozess Pj eine Gewichtung g_j festlegen können wir die Gesamtwahrscheinlichkeit aller Daten aufgrund der gegebenen Zufallsprozesse (d.h. letztlich Clusterrepräsentationen) berechnen:

$$L = \prod_{i=1}^{n}\sum_{j=1}^{k} g_j\, n(\vec{x}_i,\vec{\mu}_j,\Sigma_j)$$

Die Summe ist für einen gegebenen Vektor i die mit g_j gewichtete Summe der Wahrscheinlichkeiten, dass er von den Zufallsprozessen Pj generiert wird - $n(\vec{x}_i, \vec{\mu}_j, \Sigma_j)$ ist dabei die Wahrscheinlichkeit, dass der Vektor $\vec{x}_i$ von einem (normalverteilten) Zufallsprozess P_j mit den Parametern $\vec{\mu}_j \Sigma_j$ erzeugt wird.

Das Produkt ist die Gesamtwahrscheinlichkeit aller Daten, d.h., die Produkt der Wahrscheinlichkeit für alle Vektoren.

Nun wissen wir also, wie wahrscheinlich unsere Daten aufgrund eines Modells sind.

Wenn wir nun die Wahrscheinlichkeit der Daten kennen ist die nächste Aufgabe, einen Algorithmus zu finden, der diese Wahrscheinlichkeit durch Veränderung der Parameter $\vec{\mu}_j \Sigma_j$ möglichst hoch macht.

Dies ist der EM-Algorithmus.

Algorithmusablauf

Initialisierung

Als Anfangswerte setzen wir fest:

- die Zahl der Zufallsprozesse – d.h. die Zahl der angenommenen Cluster;
- den Mittelwert dieser Zufallsprozess – diese können zufällige Vektoren sein, oder die Clusterzentren, die wir aus einem anderen Clusteringalgorithmus (z.B. K-Means) gewonnen haben;
- eine beliebige Initialisierung der Kovarianzmatrix;

Anfangsberechnung

Wir berechnen nun die die Wahrscheinlichkeit unserer Daten aufgrund des gegebenen Modells (s.o.) und Merken sie uns. Nennen wir sie L_0

1. Schritt - E-Schritt (Expection / Erwartung)

Nun berechnen wir die bedingte Wahrscheinlichkeit hij für jeden Vektor i und jeden Clustererzeugenden Prozess P_j, dass der Vektor von dem erzeugenden Prozess generiert wird, im Verhältnis dazu dass er von anderen Prozessen generiert wird:

$$h_{ij} = \frac{n(\vec{x}_i, \vec{\mu}_j, \Sigma_j)}{\sum_{l=1}^{k} n(\vec{x}_i, \vec{\mu}_l, \Sigma_l)}$$

2. Schritt - M-Schritt - Maximierung

Dieser Schritt ist die Anpassung des Modells an die Ergebnisse des vorherigen Schritts:

Mithilfe diese Maßes berechnen wir nun einen neuen gewichteten Mittelwert für jeden Prozess, der sich aus allen Vektoren, dem Grad ihrer Zugehörigkeit zum Cluster (=dem Prozess) ergibt, und eine neue gewichtete Varianz. Die Gewichtung ist in beiden Fällen eben die Wahrscheinlichkeit h_{ij}.für die Clusterzughörigkeit des Vektors

$$\mu_j = \frac{\sum_{i=1}^{n} h_{ij} \vec{x}_i}{\sum_{i=1}^{n} h_{ij}}$$

$$\Sigma_j = \frac{\sum_{i=1}^{n} h_{ij} (\vec{x}_i - \mu_j)(\vec{x}_i - \mu_j)^T}{\sum_{i=1}^{n} h_{ij}}$$

Schließlich muss noch das Gewicht g jedes Prozesses Pj neu berechnet werden (n ist hier die Zahl aller Vektoren):

$$g_j = \frac{\sum_{i=1}^{n} h_{ij}}{n}$$

Damit haben wir nun alle Parameter des Modells neu berechnet.

Wiederberechnung der Wahrscheinlichkeit der Daten

Wir berechnen nun wiederum die die Wahrscheinlichkeit unserer Daten aufgrund des neu berechneten Modells Nennen wir sie L_1. Ist sie größer als L_0, wiederholen wird Schritt 1 und Schritt 2; und zwar so lange bis Li nicht mehr signifikant größer ist als L_{i-1}.

Auf diese Weise bekommen wir ein Modell, das unsere Daten – ausgehend von den Anfangparametern, möglichst wahrscheinlich macht.
Allerdings kann es sein, dass dieses Modell nicht das wirklich optimalste Modell ist – sind die Anfangsparameter unglücklich gewählt, kann das Ergebnis aus nicht optimalen Clustern bestehen.

10. Hidden Markov-Modelle

Hidden-Markov-Modelle sind stochastische Modelle, die im Bereich der Computerlinguistik u.a. zur Modellierung von Spracherkennung und beim Tagging eingesetzt werden.
Hidden-Markov-Modelle ähneln in mancher Hinsicht endlichen Automaten. Zur Erinnerung: Folgendes ist ein endlicher Automat:

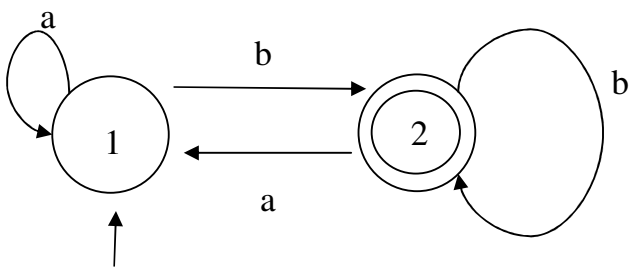

Abbildung 9 Endlicher Automat

Wie man hier sieht, ist ein endlicher Automat ein Quintupel:

Spezifikation	Beschreibung	Im Beispiel
$S = \{s_1 \dots s_n\}$	Menge der Zustände	{1,2}
$K = \{k_1 \dots k_m\}$	Menge der Ausgabesymbole / der erkannten Symbole	{a,b}
$D = \{d_{ijx} \dots d_{kly}\}$ i,j,k.l $\in$ S; x,y $\in$ K	Menge der Übergänge zwischen Zuständen mit den emittierten Symbolen	d_{12b} (beim Wechsel von Zustand 1 zu 2 wird das Symbol b erkannt bzw. emittiert)
a $\in$ S	der Startzustand	1 (bezeichnet durch den eingehenden Pfeil)
e1,e2… $\in$ S	die Menge der Endzustände	2 (bezeichnet durch den zusätzlichen Kreis)

Ein solcher Automat eignet sich zu zweierlei:
- erkenne eine Zeichenkette (Folge von Ausgabesymbolen) die der Automatenspezifikation entspricht (hier z.B. (b oder ab oder aabbab);

- generiere eine Zeichenkette, die der Automatenbeschreibung entspricht;

Markovkette

Eine Markovkette ist nun ein stochastischer Vorgang, der sich mithilfe eines ähnlichen Modells wie ein endlicher Automat darstellen lässt. Die wichtigste Modifikation ist, dass Übergängen zwischen Zuständen Wahrscheinlichkeiten zugewiesen werden:

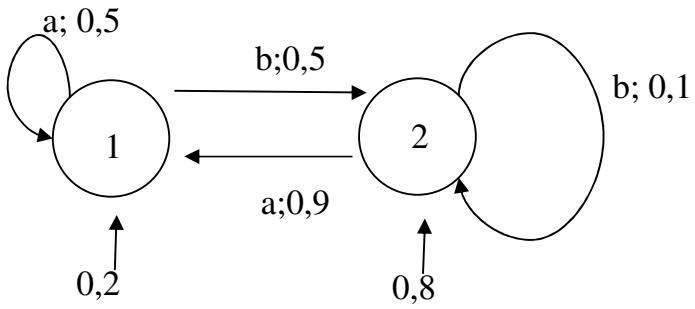

Abbildung 10 Markovkette

Folgende Modifikationen (in Fett) ergeben sich in der formalen Spezifikation zu den endlichen Automaten:

Spezifikation	Beschreibung	Im Beispiel
$S = \{s_1 \dots s_n\}$	Menge der Zustände	{1,2}
$K = \{k_1 \dots k_m\}$	Menge der Ausgabesymbole	{a,b}
$A = \{a_{ijx} \dots a_{kly}\}$ i,j,k.l ∈ S; x,y ∈ K	**Wahrscheinlichkeiten** der Übergänge zwischen Zuständen mit den emittierten Symbolen	a_{12b} = 0,5 (beim Wechsel von Zustand 1 zu 2 wird das Symbol b erkannt bzw. emittiert; die Wahrscheinlichkeit des Übergangs ist 0,5)
P = {pi… } mit .i ∈ S	**der Startzustand entfällt, an seine Stelle treten die Wahrscheinlichkeiten, in einem Zustand s zu beginnen**	1 → 0,2
~~**e1,e2…∈ S**~~	**die Menge der Endzustände entfällt**	-

Eine solche Markovkette eignet sich dazu, alle Vorgänge zu modellieren, bei der nur eine eingeschränkte Zahl von Daten aus dem linken Kontext Beachtung finden soll – dies ist der sogenannte beschränkte Horizont dieser Modelle. So kann man beispielsweise ein –Bigrammmodell über Wörter (zur Vorhersage des nächsten Wortes) sehr einfach mit einer Markovkette implementieren. Die bedingten Wahrscheinlichkeiten, dass ein Wort einem anderen folgt, werden durch Übergangswahrscheinlichkeiten modelliert; jeder Übergang emittiert ein Wort und führt in einen Zustand, der vom emittierten Wort determiniert ist. Auch alle anderen N-Gramm-Modelle mit endlichem N kann man als Markovketten implementieren, indem man in den Zuständen die N-1 vorangehende Wörter kodiert.

Die Wahrscheinlichkeit einer Ausgabe bzw. einer erkannten Sequenz ergibt sich aus dem Produkt der Wahrscheinlichkeiten der durchlaufenen Übergänge – d.i. die Wahrscheinlichkeit des durchlaufenen Pfades. Gibt es mehrere Pfade durch die Markovkette, die dieselbe Ausgabe generieren, müssen die Pfadwahrscheinlichkeiten addiert werden.

Etwas komplexer noch ist ein Hidden-Markov-Modell. Der Name rührt daher, dass man hier in der Regel die Zustände und die Übergangswahrscheinlichkeiten nicht direkt beobachtet, sondern nur die Ausgabesequenz betrachtet.
Die hauptsächliche Modifikation zur Modellierung der Markovketten, ist, dass ein Übergang nicht mehr ein bestimmtes Symbol emittiert, sondern jeder Übergang ein oder mehrere Symbole k mit einer gewissen Wahrscheinlichkeit emittiert.

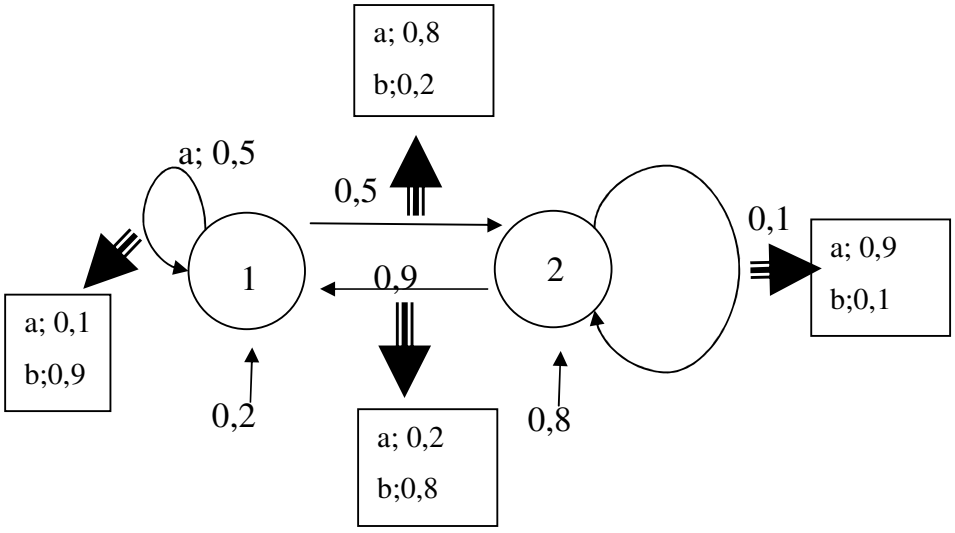

Abbildung 11 Hidden-Markov-Modell

Folgende Modifikationen (in Fett) ergeben sich zur Markovkette:

Spezifikation	Beschreibung	Im Beispiel
$S = \{s_1 \dots s_n\}$	Menge der Zustände	{1,2}
$K = \{k_1 \dots k_m\}$	Menge der Ausgabesymbole	{a,b}
$A = \{a_{ij} \dots a_{kl}\}$ i,j,k.l ∈ S; **x,y ∈ K**	Wahrscheinlichkeiten **der Übergänge zwischen Zuständen** (ohne emittierte Symbole)	a_{12b} = 0,5 (der Wechsel von Zustand 1 zu 2 hat die Wahrscheinlichkeit 0,2)
$B = \{b_{ijx} \dots b_{kly}\}$ i,j,k.l ∈ S; **x,y ∈ K**	**Wahrscheinlichkeiten, dass bei den Übergängen bestimmte Symbole emittiert werden**	b_{12b} = 0,2 (die Emission des Zeichens b beim Wechsel von Zustand 1 zu 2 hat die Wahrscheinlichkeit 0,5)
$P = \{p_i \dots\}$ mit .i ∈ S	der Startzustand ist auch hier durch Anfangswahrscheinlich-keiten ersetzt	1 → 0,2

Achtung: Häufiger als dieser Typ von Hidden-Markov-Modellen sind sogenannte zustandsemmitierende Modelle (die hier ist ein übergangsemittierendes Modell), bei denen die Symbole beim Eintritt in einen Zustand ausgegeben werden – wieder mit einer gewissen Wahrscheinlichkeit. Die beiden Modelle lassen sich ineinander überführen.

10.1. Drei wichtige Berechnungen für Markov-Modelle:

Drei Berechnungen sind in Bezug auf Markov-Modelle relevant.

- Wie wahrscheinlich ist eine Beobachtung, gegeben ein Markov-Modell (P(O|m)? Diese Frage etwa wichtig um, gegeben ein Markovmodell entscheiden zu können welche von zwei möglichen Ausgaben die wahrscheinlichere ist. Sie wird z.B. bei der Spracherkennung verwendet um zu entscheiden, welches Phonem erkannt wurde.
- Wie ermitteln wir die Zustandssequenz, die am besten unseren Beobachtungen entspricht? Diese Frage ist z.B. im syntaktischen Tagging relevant. Hier werden in den Zuständen die Syntaxkategorien kodiert, in den Ausgabesymbolen die Wörter. Hier soll für eine gegeben Wortfolge die wahrscheinlichste Tagfolge gefunden werden, um jedem Wort die wahrscheinlichste syntaktische Kategorie zuzuweisen;
- Wie ermitteln wir das Modell, das am ehesten unseren Beobachtungen entspricht. Dies ist Frage nach dem Training eines HMM. Wie ermittle ich aufgrund der Trainingsdaten das Modell, mit dem ich später arbeiten will.

Im Skript wird nur Punkt 1) behandelt; für 2) und 3) s. Manning/Schütze, Kapitel 10.2

Wie wahrscheinlich ist eine Beobachtung, gegeben ein HMM

Naive Berechnung: Berechne die Wahrscheinlichkeit für alle Pfade, die die Beobachtung erzeugen; addiere die Wahrscheinlichkeiten. Die Wahrscheinlichkeit einer Ausgabesequenz in Bezug auf eine bestimmte Abfolge von Zuständen errechnet sich hier aus dem Produkt der Übergangswahrscheinlichkeiten zwischen den Zuständen, und den Emissionswahrscheinlichkeiten für die Symbole in der Ausgabesequenz; hinzu kommt noch die Wahrscheinlichkeit des Anfangszustandes. Die Wahrscheinlichkeit der Ausgabesequenz "ab" in der Zustandsfolge 1 2 1 in obigem Modell m wäre:

$$p(\text{"ab"}|\text{"1 2 1"}, m) = p_i * (a_{12} * b_{12a}) * (a_{21} * b_{21b}) = 0{,}2 * (0{,}5*0{,}8) * (0{,}9 * 0{,}8)$$

Will man nun die Gesamtwahrscheinlichkeit für eine beobachtete Sequenz für alle möglichen Pfade durch das Modell berechnen, muss man die Wahrscheinlichkeiten für alle Pfade aufsummieren.
Dass heisst wir berechnen:

$$\sum_{\forall X} p(O \mid X, m)$$

(hier ist X ein Pfad durch das Modell, mit der Länge der Beobachtungssequenz+1, m ist unser Modell; O ist die Ausgabesequenz)
Das Problem ist nun, dass die Zahl der möglichen Pfade kann für längere Beobachtungen sehr hoch wird, dass die Wahrscheinlichkeit nicht mehr zu berechnen ist. So ist die Zahl der Zustandsfolgen für ein Modell mit N Zuständen und für eine Eingabesequenz der Länge L exponenentiell, d.h. N^L.

Lösung für dieses Komplexitätsproblem ist die: Darstellung der Wahrscheinlichkeit, einen Zustand für die Eingabesequenz O nach t Schritten zu erreichen, als Verband. Für jeden Zustand und jeden Zeitpunkt t existiert damit nur noch eine Wahrscheinlichkeit.
Hier die Tabelle, für die Berechnung der Wahrscheinlichkeit für die Sequenz "aba" in obigem Modell. Die Wahrscheinlichkeiten zu einem Zeitpunkt ti für die gegebenen Eingabesequenz in einem Zustand j zu sein wird kodiert als $P_{ti\text{-}j}$. Alle anderen Werte stammen aus dem Modell oben. Wie man sieht, ergibt sich die Wahrscheinlichkeit, zu einem bestimmten Zeitpunkt für eine bestimmte Eingabe in einem bestimmten Zustand zu sein, aus den Wahrscheinlichkeiten zum vorausgehenden Zeitpunkt. Für jedes Zeichen der Eiingabesequenz muss nur für jeden Zustand eine Neuberechnung der Wahrscheinlichkeit durchgeführt werden. Die Wahrscheinlichkeit zum Zeitpunkt t0 ergibt sich aus den Anfangswahrscheinlichkeiten für jeden Zustand.

Zustandswahrscheinlichkeit für s1	p_1	$P_{t1\text{-}1\text{-}a} = p_1 * a_{11} * b_{11a} + p_2 * a_{21} * b_{21a}$	$P_{t2\text{-}1\text{-}a} = P_{t1\text{-}1\text{-}a} * a_{11} * b_{11b} + P_{t2\text{-}2\text{-}a} * a_{21} * b_{21b}$	…
Zustandswahrscheinlichkeit für s2	p_2	$P_{t1\text{-}2\text{-}a} = p_1 * a_{12} * b_{12a} + p_2 * a_{22} * b_{23a}$	$P_{t2\text{-}1\text{-}a} = P_{t1\text{-}1\text{-}a} * a_{12} * b_{12b} + P_{t1\text{-}2\text{-}a} * a_{22} * b_{22b}$	…
Eingabesymbole	-	a	b	…
Zeitpunkt	t0	t1	t2	…

11. Grundlagen der Informationstheorie

11.1. Entropie (Entropy)

(Basiert auf Manning/Schütze: S. 60ff)

Entropie misst den Informationsgehalt einer Zufallsvariablen. Sie ist:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

Hierbei ist p(x) die Wahrscheinlichkeit für die verschiedenen Werte x der diskreten Zufallsvariablen X ist (probability mass function). Eine diskrete Zufallsvariable ist eine Variable mit abzählbaren Werten; z.B. Ergebnis eine Münzwurfs (Werte Kopf oder Zahl) oder eine die Länge in Buchstaben einen zufällig aus einem Korpus ausgewählten Wortes.

Liegt für X eine Laplace'sche Wahrscheinlichkeitsverteilung – d.h. alle p(x) sind gleich - lässt sich die Formel vereinfachen zu:

$$H(X) = \log_2 n$$

(Dies ist leicht herzuleiten)

Beispiel 1 : wird eine faire Münze geworfen ist p(X) jeweils 1/2 für p(Kopf) und p(Wappen). Die Entropie ist dann -1/2*log 1/2 + 1/2*log 1/2 =- log 1/2 = 1.

Beispiel 2: Ein Würfel ist nicht fair. Mit p(6) = 1/2, p(1) ... p(5) = 1/10. Dann ist die Entropie: 1/2*1+1/2*(log 10) = 0,5+1,7 = 2,2

Die Entropie mit dem Logarithmus auf der Basis 2 misst die durchschnittliche Länge einer Informationssequenz in Bit, um einen Wert der Zufallsvariablen zu übermitteln, d.h. um n Werte der Zufallsvariablen X zu übermitteln sind n*H(X) Bit nötig.

Beispiel 1: Eine faire Münze wird 20 mal geworfen. Die Entropie ist, wie oben berechnet, ein Bit. Dann sind 20 Bit nötig, um das Ergebnis zu übermitteln.

Bedingte Entropie

$$H(X \mid Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x \mid y)$$

Die bedingte Entropie H(X|Y) misst die zusätzliche Information, die notwendig ist, um den Wert einer zweiten Zufallsvariablen X zu spezifizieren wenn der Wert von Y bereits bekannt ist. Sind die Variable unabhängig voneinander so ist H(X|Y) = H(X).

11.2. Die Mutual Information (Transinformation)

Die Mutual Information I(X;Y) zweier Zufallsvariablen X und Y ergibt sich aus dem Differenz aus Entropie einer Variable H(X) und der bedingten Entropie H(X|Y). Also:

$$I(X;Y) = H(X)-H(X|Y).$$

Sind beide Variablen unabhängig, so ist sie 0; bei maximaler Abhängigkeit ist sie =H(X).

Durch Umformung der obigen Formel ergibt sich:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Als Mutual Information zweier Werte einer Zufallsvariablen (pointwise mutual information) wird folgender Wert bezeichnet:

$$I(x;y) = \log \frac{p(x,y)}{p(x)p(y)}$$

Dieser Wert ist in der Computerlinguistik sehr beliebt, um die Korrelation eines Assoziationspaares von Wörtern zu messen (s.o.)