Einführung in die Computerlinguistik

Statistische Verfahren zur Wortbedeutungs-Disambiguierung

WS 2013/2014

Manfred Pinkal

Lexikalische Mehrdeutigkeit

- Äußerungs- und Textverstehen impliziert die Erkennung der korrekten, im Kontext intendierten Äußerungsbedeutung.

- Wörter sind vielfach mehrdeutig:
  - *Bank: Geldinstitut / Sitzmöbel*
  - *Maschine: Flugzeug / Motorrad/ Technisches Gerät*
  - *Absatz: Schuh/ Treppe/ Text/ Verkauf*
  - *aufgeben: einen Plan / einen Koffer aufgeben*

- Die Disambiguierung der Wortbedeutung (engl. "Word-sense disambiguation": WSD) ist zentrale Aufgabe der Computerlinguistik.

Wissensbasierte Disambiguierung

The box was in the pen – The pen was in the box

- Disambiguierung in der Wissensbasierten Sprachverarbeitung: durch Inferenz über Alltagswissen.
- Das Inferenzproblem ist grundsätzlich lösbar (Logik, Deduktion, Theorembeweiser)
- Das harte Problem ist die Kodierung von Alltagswissen in ausreichendem Umfang und in der angemessenen Form.

WSD und Weltwissen

- Wissensbasierte Disambiguierung durch Inferenz mit Weltwissen:

  Ich gehe nachher einkaufen, deshalb muss ich jetzt dringend zur Bank.
- Aber: Unterschiedliche Verwendungen von „*Bank*“ benötigen unterschiedliches Wissen:
  - *Ich will ein Haus bauen, ...*
  - *Ich habe geerbt, ...*
  - *Ich brauche eine neue Kreditkarte, ...*
- Viele Wörter sind mehrdeutig, viele sind vielfach mehrdeutig, und jedes Wort erfordert spezifisches Wissen: *Bank – Absatz*
- Wissensbasierte Disambiguierung ist praktisch nicht machbar: Riesige Mengen an handkodiertem Weltwissen wären nötig: „Knowledge Bottleneck“
- Attraktive Alternative: Statistische Modellierung

Statistische Modellierung: Allgemeines Schema

- Manuelle Korpusannotation
- Merkmalsspezifikation
- Automatische Merkmalsextraktion
- Training eines statistischen Modells
- Evaluierung

WSD: Korpusannotation

- Spezifikation des Annotationsschemas: Übernahme von Wortbedeutungen aus einem Wörterbuch oder Thesaurus (Standard: WordNet-Synsets)

- Annotation aller Zielwort-Instanzen im Trainingskorpus mit einer Wortbedeutung

Trainings-Korpus

...

(A237) ... Für diejenigen, denen Komfort wichtig ist, haben wir eine Bank mit leicht schwingender Rückenlehne entwickelt. ...

(A295) ... Ich suche noch eine Bank für meinen Garten und sondiere deshalb gerade Angebote. ...

(A303) ... Habe im März 2000 einen höheren Betrag bei einer Bank angelegt. ...

(A452) ... Beim Test Anlageberatung der Banken löste kein Institut die einfache Frage nach einer sicheren Anlage wirklich gut. ...

...

Trainings-Korpus: Annotation mit WSD-Information

...

(A237) ... Für diejenigen, denen Komfort wichtig ist, haben wir eine Bank <bank1> mit leicht schwingender Rückenlehne entwickelt. ...

(A295) ... Ich suche noch eine Bank <bank1> für meinen Garten und sondiere deshalb gerade Angebote. ...

(A303) ... Habe im März 2000 einen höheren Betrag bei einer Bank<bank2> angelegt. ...

(A452) ... Beim Test Anlageberatung der Banken <bank2> löste kein Institut die einfache Frage nach einer sicheren Anlage wirklich gut. ...

...

WSD: Merkmalsspezifikation

- Wir verwenden Kontextwörter als Merkmale für die Disambiguierung. Präziser ausgedrückt:

- Wir nehmen für alle Zielwörter eine gemeinsame Merkmalsmenge an: Vorkommen/ Nichtvorkommen der n (z.B. n=1000) häufigsten Inhaltswörter (Substantive, Verben, Adjektive) im Kontext des Zielwortes.

- Den Kontext einer Instanz legen wir als den Satz fest, in dem die Instanz vorkommt (alternativ: das Fenster mit fester Länge von k Wörtern rechts und links von der Instanz (z.B. k=10)).

Zielwörter und Kontextwörter

...

(A237) ... Für diejenigen, denen Komfort wichtig ist, haben wir eine Bank <bank1> mit leicht schwingender Rückenlehne entwickelt. ...

(A295) ... Ich suche noch eine Bank <bank1> für meinen Garten und sondiere deshalb gerade Angebote. ...

(A303) ... Habe im März 2000 einen höheren Betrag bei einer Bank<bank2> angelegt. ...

(A452) ... Beim Test Anlageberatung der Banken <bank2> löste kein Institut die einfache Frage nach einer sicheren Anlage wirklich gut. ...

...

WSD: Merkmalsextraktion

- Wir lemmatisieren die Kontextwörter (Bestimmung des Wortstamms durch morphologische Analyse)

- Wir ermitteln für jedes Vorkommen eines Zielwortes ein spezifisches Merkmalsmuster v, indem wir für jedes i: $1 \leq i \leq 1000$ setzen:
  - $v_i = 1$, wenn das Wort w_i als Kontextwort im Satz auftritt
  - $v_i = 0$ sonst.

- Alle Merkmale sind Boole'sche Merkmale ($\in \{0,1\}$). Das Merkmalsmuster v kann als Vektor in einem hochdimensionalen Raum betrachtet werden.

Merkmalsextraktion, Beispiel

Instanz Nr	A237	A295	A303	A452	...
Annotation	bank1	bank1	bank2	bank2	...
Frage	0	0	0	1	...
Komfort	1	0	0	0	...
anlegen	0	0	1	0	...
Betrag	0	0	1	0	...
Garten	0	1	0	0	...
suchen	0	1	0	0	...
fahren	0	0	0	0	...
richtig	0	0	0	0	...
Test	0	0	0	1	...
...	...	...	...	...	...

Neue Instanz:

Keine Frage: In einen ordentlichen Garten gehören neben einer Bank auch die richtigen Möbel.

Neu
?
1
0
0
0
1
0
0
0
1
0
...

WSD: Statistisches Modell

- Wie bestimmen wir den Wortsinn einer neuen Instanz von *Bank* auf der Grundlage des Musters von Kontextwörtern v ?

Versuch: Analog zum POS-Modell der letzten Woche:

- Wir zählen zunächst für jedes Merkmalsmuster aus, wie oft im Trainingskorpus das Muster mit $bank1$ und $bank2$ vorkommt.
- Wir schätzen die bedingten Wahrscheinlichkeiten $P(bank1|v)$ und $P(bank2|v)$ auf der Grundlage dieser Frequenzen.

$$P(s \mid v) = \frac{P(s,v)}{P(v)} \approx \frac{Fr(s,v)}{Fr(v)}$$

- Wir weisen den wahrscheinlicheren Wortsinn zu.
- Sparse-Data-Problem:
  - 1000 Wörter, je 2 Werte: 2^{1000} alternative Kontextmuster.

Das Bayessche Theorem

- Das Bayessche Theorem oder die Bayes-Regel:

$$P(E \mid F) = \frac{P(F \mid E) \cdot P(E)}{P(F)}$$

- Die Bayes-Regel ist ein elementares Gesetz der Wahrscheinlichkeitstheorie. Sie ist überall da nützlich, wo der Schluss von einer Größe F auf eine andere Größe E bestimmt werden soll (typischerweise von einem Symptom auf eine relevante Eigenschaft/ die Ursache), die Abhängigkeit in der anderen Richtung (von der Ursache auf das Symptom) aber besser zugänglich ist.

Bayes-Theorem und WSD

- Merkmalsmuster v : Symptom
- Wortsinn s : Ursache
- Mit Bayes-Regel : $$P(s \mid v) = \frac{P(v \mid s) \cdot P(s)}{P(v)}$$
- Der wahrscheinlichste Wortsinn: $$\max_s P(s \mid v) = \max_s \frac{P(v \mid s) \cdot P(s)}{P(v)}$$ $$= \max_s P(v \mid s) \cdot P(s)$$
- $P(s)$ ist die globale, "a priori"-Wahrscheinlichkeit des Wortsinns s.
- $P(v)$, die Wahrscheinlichkeit des Merkmalsmusters, wird nicht mehr benötigt.
- Wie ermitteln wir $P(v \mid s)$? – Wiederum: Sparse-Data-Problem:
  - Die Auftretenshäufigkeit eines bestimmten Musters mit einer Lesart ist typischerweise klein (meistens sogar 0) und erlaubt deshalb keine verlässliche Abschätzung von $P(v|s)$.

Unabhängigkeitsannahme

- Unter der Voraussetzung, dass zwei Ereignisse E_1 und E_2 unabhängig sind, ist die gemeinsame Wahrscheinlichkeit von E_1 und E_2 das Produkt der Einzelwahrscheinlichkeiten:

$$P(E_1, E_2) = P(E_1) * P(E_2)$$

$$P(E_1, E_2 \mid F) = P(E_1 \mid F) * P(E_2 \mid F)$$

- Die Regel gilt für Produkte mit beliebig vielen Faktoren. Unter der ("naiven")Annahme, dass die Merkmale unabhängig voneinander auftreten, lässt sich die Wahrscheinlichkeit eines Merkmalsmusters approximieren als Produkt der Einzelwahrscheinlichkeiten für seine Komponenten:

$$P(v \mid s) \approx \prod_{v_i} P(v_i \mid s)$$

- Maschinelle Lernverfahren, die diese Unabhängigkeitsnnahme nutzen, um Wahrscheinlichkeiten trotz geringer Datenmengen zu approximieren, heißen "Naive Bayes Classifier".

Von binären Merkmalsmustern ...

Instanz Nr	...	A237	A295	A303	A452	...
Annotation	...	bank1	bank1	bank2	bank2	...
Frage	...	0	0	0	1	...
Komfort	...	1	0	0	0	...
anlegen	...	0	0	1	0	...
Betrag	...	0	0	1	0	...
Garten	...	0	1	0	0	...
suchen	...	0	1	0	0	...
fahren	...	0	0	0	0	...
richtig	...	0	0	0	0	...
Test	...	0	0	0	1	...
...	...	...	...	...	...	...

	bank1	**bank2**
Frage	11	15
Komfort	7	3
anlegen	3	84
Betrag	5	41
Garten	40	1
suchen	7	32
fahren	4	24
richtig	12	21
Test	2	5
...	...	...

... zu Wahrscheinlichkeitsschätzungen von Kontextmerkmalen

- Wir gehen von insgesamt 500 Instanzen von "Bank" im Trainingskorpus aus, davon 200 als bank1 und 300 als bank2 annotiert.

	bank1	**bank2**
Frage	11	15
Komfort	7	3
anlegen	3	84
Betrag	5	41
Garten	40	1
suchen	7	32
fahren	4	24
richtig	12	21
Test	2	5
...	...	...

	$P(1\|bank1)$	$P(0\|bank1)$
Frage	0,055	0,945
Komfort	0,035	0,965
anlegen	0,015	0,985
Betrag	0,025	0,975
Garten	0,200	0,800
suchen	0,035	0,965
fahren	0,020	0,980
richtig	0,060	0,940
Test	0,010	0,990
...	...	...

... zu Wahrscheinlichkeitsschätzungen von Kontextmerkmalen

- Wir gehen von insgesamt 500 Instanzen von "Bank" im Trainingskorpus aus, davon 200 als bank1 und 300 als bank2 annotiert.

	bank1	**bank2**
Frage	11	15
Komfort	7	3
anlegen	3	84
Betrag	5	41
Garten	40	1
suchen	7	32
fahren	4	24
richtig	12	21
Test	2	5
...	...	...

	$P(1\|bank2)$	$P(0\|bank2)$
Frage	0,050	0,950
Komfort	0,010	0,990
anlegen	0,280	0,720
Betrag	0,137	0,863
Garten	0,003	0,997
suchen	0,107	0,893
fahren	0,080	0,920
richtig	0,070	0,930
Test	0,017	0,983
...	...	...

Beispiel

Keine Frage: In einen ordentlichen Garten gehören neben einer Bank auch die richtigen Möbel.

$$\max_s P(s \mid v) = \max_s P(v \mid s) \cdot P(s)$$

$$s \in \{bank1, bank2\}$$

$$P(v \mid bank1) \approx \prod_{v_i} P(v_i \mid bank1)$$

$$P(v \mid Bank2) \approx \prod_{v_i} P(v_i \mid Bank2)$$

	$P(1\|bank1)$	$P(0\|bank1)$
Frage	0,055	0,945
Komfort	0,035	0,965
anlegen	0,015	0,985
Betrag	0,025	0,975
Garten	0,200	0,800
suchen	0,035	0,965
fahren	0,020	0,980
richtig	0,060	0,940
Test	0,010	0,990
...	...	...

v_i
1
0
0
0
1
0
0
1
0
...

	$P(1\|bank1)$	$P(0\|bank1)$
Frage	0,055	0,945
Komfort	0,035	0,965
anlegen	0,015	0,985
Betrag	0,025	0,975
Garten	0,200	0,800
suchen	0,035	0,965
fahren	0,020	0,980
richtig	0,060	0,940
Test	0,010	0,990
...	...	...

v_i	$P(v_i\|bank1)$
1	0,055
0	0,965
0	0,985
0	0,975
1	0,200
0	0,965
0	0,980
1	0,060
0	0,990
...	...

$$P(v \mid bank1) \approx \prod_{v_i} P(v_i \mid bank1) = 0{,}000572$$

	$P(1\|bank2)$	$P(0\|bank2)$
Frage	0,050	0,950
Komfort	0,010	0,990
anlegen	0,280	0,720
Betrag	0,137	0,863
Garten	0,003	0,997
suchen	0,107	0,893
fahren	0,080	0,920
richtig	0,070	0,930
Test	0,017	0,983
...	...	...

v_i	$P(v_i\|bank2)$
1	0,050
0	0,990
0	0,720
0	0,863
1	0,003
0	0,893
0	0,920
1	0,070
0	0,983
...	...

$$P(v \mid Bank2) \approx \prod_{v_i} P(v_i \mid Bank2) = 0{,}000006$$

Beispiel

Keine Frage: In einen ordentlichen Garten gehören neben einer Bank auch die richtigen Möbel.

$$\max_s P(s \mid v) = \max_s P(v \mid s) \cdot P(s)$$

$$s \in \{bank1, bank2\}$$

$$P(v \mid bank1) \approx \prod_{v_i} P(v_i \mid bank1) = 0{,}000572 \qquad P(bank1) = 0{,}4$$

$$P(v \mid bank2) \approx \prod_{v_i} P(v_i \mid bank2) = 0{,}000006 \qquad P(bank2) = 0{,}6$$

$$P(v \mid bank1) \cdot P(bank1) = 0{,}000228$$

$$P(v \mid bank2) \cdot P(bank2) = 0{,}000004$$

$$\max_s P(s \mid v) = bank1$$

Performanz von WSD-Verfahren

- Wenn man durchgängig die häufigste Wortbedeutung wählt, erreich man bereits ca. 80% Akkuratheit („most frequent wordsense baseline“).

- Die Übereinstimmung menschlicher Annotatoren („Inter-Annotator Agreement“) liegt bei 95% (Obergrenze, „Upper bound“ für automatische Annotation).

- Die besten WSD-Systeme erreichen heute 90-92% Akkuratheit.

WSD

- Eine der schwierigsten Aufgaben in der Computerlinguistik:
- Sehr viele Wörter sind auf sehr unterschiedliche Weise mehrdeutig. Man benötigt riesige Mengen von Trainingsmaterial.
- Alle bisher vorgestellten Lernverfahren sind „überwachte“ (supervised) Lernvefahren: Sie erfordern die manuelle Annotation eines Trainingskorpus.
- Attraktiver sind „halbüberwachte“ (semi-supervised) Verfahren, bei denen ein großes Trainingskorpus (teil-)automatisch auf der Grundlage einer kleinen Menge von handannotierten „Seed-Daten“ erzeugt wird.
- Noch attraktiver sind „unüberwachte“ statistische Verfahren, die Resultate ohne jedes Training erzielen. – Die Akkuratheit von unüberwachten Systemen ist deutlich niedriger.