Proseminar Linguistische Annotation

Ines Rehbein und Josef Ruppenhofer

SS 2010

Verlässlichkeit der Annotation

- Letzte Sitzung:
  - G-Theory – Welche Faktoren sind verantwortlich für niedriges Agreement?
    (Bayerl & Paul. 2007. Identifying Sources of Disagreement)
  - Einfluss der Annotationsqualität auf Maschinelles Lernen – hohes Agreement (> 0.8) nicht immer ausreichend
    (Reidsma & Carletta. 2008. Reliability Measurement without Limits)
- Heute:
  - Welche Faktoren sind verantwortlich für niedriges Agreement?
  - Wie kann man systematische Muster in den annotierten Daten finden?
    (Passonneau et al. 2009. Making Sense of Word Sense Variation Passonneau et al. 2010. Word Sense Annotation of Polysemous Words by Multiple Annotators)
- Hypothese:
  - verschiedene Worte führen zu höherem/niedrigerem IAA
  - Gründe dafür sollten in NLP-Tasks modelliert werden, um automatische Systeme robuster zu machen

Verlässlichkeit der Annotation

- Letzte Sitzung:
  - G-Theory – Welche Faktoren sind verantwortlich für niedriges Agreement?
    (Bayerl & Paul. 2007. Identifying Sources of Disagreement)
  - Einfluss der Annotationsqualität auf Maschinelles Lernen – hohes Agreement (> 0.8) nicht immer ausreichend
    (Reidsma & Carletta. 2008. Reliability Measurement without Limits)
- Heute:
  - Welche Faktoren sind verantwortlich für niedriges Agreement?
  - Wie kann man systematische Muster in den annotierten Daten finden?
    (Passonneau et al. 2009. Making Sense of Word Sense Variation Passonneau et al. 2010. Word Sense Annotation of Polysemous Words by Multiple Annotators)
- Hypothese:
  - verschiedene Worte führen zu höherem/niedrigerem IAA
  - Gründe dafür sollten in NLP-Tasks modelliert werden, um automatische Systeme robuster zu machen

Verlässlichkeit der Annotation

- Letzte Sitzung:
  - G-Theory – Welche Faktoren sind verantwortlich für niedriges Agreement?
    (Bayerl & Paul. 2007. Identifying Sources of Disagreement)
  - Einfluss der Annotationsqualität auf Maschinelles Lernen – hohes Agreement (> 0.8) nicht immer ausreichend
    (Reidsma & Carletta. 2008. Reliability Measurement without Limits)
- Heute:
  - Welche Faktoren sind verantwortlich für niedriges Agreement?
  - Wie kann man systematische Muster in den annotierten Daten finden?
    (Passonneau et al. 2009. Making Sense of Word Sense Variation Passonneau et al. 2010. Word Sense Annotation of Polysemous Words by Multiple Annotators)
- Hypothese:
  - verschiedene Worte führen zu höherem/niedrigerem IAA
  - Gründe dafür sollten in NLP-Tasks modelliert werden, um automatische Systeme robuster zu machen

Übersicht

- Daten + Task
- 3 Faktoren des lexikalen Gebrauchs, die IAA beeinflussen
- Data Mining mit Assoziationsregeln (Association rules)

Übersicht

- Daten + Task
- 3 Faktoren des lexikalen Gebrauchs, die IAA beeinflussen
- Data Mining mit Assoziationsregeln (Association rules)

Daten + Task

- Kontext: Wortbedeutungs-Disambiguierung (WSD)
- MASC Project (Manually Annotated Sub-Corpus Project)
  - Erstellung eines kleinen, repräsentativen Korpus (American English)
  - geschriebene und gesprochene Sprache, Subkorpus des OANC (Open American National Corpus; `http://www.anc.org`)
  - Ziel: Erleichterung der Alignierung von WordNet und FrameNet
- 10 frequente, polyseme Worte (nicht enthalten in FN) als auch Worte, die in beiden Ressourcen vorkommen
- 1000 Vorkommen für jedes Wort (alle Vorkommen in MASC sowohl zufällig ausgewählte Vorkommen in OANC)
- annotiert von mindestens 1 studentischen Annotator
- 50 Vorkommen für jedes Wort annotiert von allen 6 Annotatorinnen, um IAA zu messen

Daten

Word	POS	No. Senses	Frequenz
fair	Adj	10	463
long	Adj	9	2706
quite	Adj	6	244
land	Noun	11	1288
time	Noun	10	21790
work	Noun	7	5780
know	Verb	11	10334
say	Verb	11	20372
show	Verb	12	11877
tell	Verb	8	4799

Inter-Annotator Agreement

- Keine Korrelation zwischen IAA und POS oder IAA und Annotator/in (Krippendorff's α, Cohen's κ)
- Statt dessen: systematische Pattern über alle Annotator/innen (sekundärer Effekt: POS)

POS	**Word**	α	κ	**No. Senses**	**Used**	**Freq.**
long	Adj	0.666	0.666	9	8	2706
fair	Adj	0.355	0.359	10	5	463
work	Noun	0.536	0.536	7	7	5780
land	Noun	0.263	0.267	11	8	1288
tell	Verb	0.415	0.416	8	8	4799
show	Verb	0.264	0.270	12	11	11877

Inter-Annotator Agreement II

- Annotatoren zeigen höhere Übereinstimmung auf manchen Worten (long, work, tell), niedrigere Übereinstimmung auf anderen Worten (fair, land, show)
- **Hypothese**: Ursache sind nicht Unterschiede zwischen Annotatorinnen, sondern **Unterschiede im Wortgebrauch**
  - größere Spezifizität des Kontexts → höheres Agreement
  - konkretere Wortbedeutungen → höheres Agreement
  - eng verwandte Wortbedeutungen → niedrigeres Agreement
- Keine signifikante Korrelation mit IAA:
  - Expertise der Annotatorinnen
  - Wortart
  - Anzahl der Wortbedeutungen in WordNet oder im Korpus
  - Verteilung der Wortbedeutungen

Übersicht

- Daten + Task
- 3 Faktoren des lexikalen Gebrauchs, die IAA beeinflussen
- Data Mining mit Assoziationsregeln (Association rules)

Spezifität des Kontexts

- **long - hohe Übereinstimmung**
- spezifischer Kontext:

(1) For 18 **long** months Michael could not find a job.

WN S1. temporal extent [N=6 of 6]

- unspezifisch:

(2) After I had submitted the manuscript my editor at Simon Schuster had suggested a number of cuts to streamline what was already a **long** and involved chapter on Brians ideas.

WN S2. spatial extent [N=3 of 6]
WN S1. temporal extent [N=2 of 6]
WN S9. more than normal or necessary [N=1 of 6]

Konkretheit der Wortbedeutungen

- **fair - geringe Übereinstimmung**
- abstrakte Wortbedeutung

(3) By insisting that everything Microsoft has done is **fair** competition they risk the possiblility that the kpublic if it accepts the judges finding to the cont5rary will conclude that Microsoft doesn't know the difference.

WN S1. free of favoritism/bias [N=6 of 6]

- wird oft im Kombination mit S2 annotiert

(4) I think that's true I can remember times my parents would say well what do you think would be a **fair** punishment.

WN S1. free of favoritism/bias [N=3 of 6]
WN S2. not excessive or extreme [N=3 of 6]

Ähnlichkeit der Wortbedeutungen

- **land - geringe Übereinstimmung**

(5) India is exhilarrating exhausting and infuriating a **land** where you'll find the practicalities of daily life overlay the mysteries that popular myth attaches to India.

WN S5. territory occupied by a nation [N=6 of 6]

- viele sehr ähnliche Wortbedeutung

(6) uh the Seattle area we lived outside outside of the city in the country and uh we have five acres of **land** up against a hillside where I grew up and so we did have a garden about a one a half acre garden

WN S4. solid part of the earth's surface [N=1 of 6]
WN S1. location of real estate [N=2 of 6]
WN S7. extensive landed property [N=3 of 6]
→ Wortbedeutungen zusammenlegen?

3 Faktoren, die IAA beeinflussen

1. Spezifität des Kontexts
2. Konkretheit der Wortbedeutungen
3. Ähnlichkeit der Wortbedeutungen

Aber - wie missst man die Ähnlichkeit von Wortbedeutungen?

- **Inter-Sense Similarity Measure (ISM)** *(Ide, 2006)*
- basiert auf dem **Lesk Ähnlichkeitsmaß**

Lesk (1985)

Relatedness of two words is proportional to to the extent of overlaps of their dictionary definitions.
Banerjee and Pedersen (2002) extended this notion to use WordNet as the dictionary for the word definitions. This notion was further extended to use the rich network of relationships between concepts present is WordNet.

3 Faktoren, die IAA beeinflussen

1. Spezifität des Kontexts
2. Konkretheit der Wortbedeutungen
3. Ähnlichkeit der Wortbedeutungen

   Aber - wie missst man die Ähnlichkeit von Wortbedeutungen?

- **Inter-Sense Similarity Measure (ISM)** *(Ide, 2006)*
- basiert auf dem **Lesk Ähnlichkeitsmaß**

Lesk (1985)

Relatedness of two words is proportional to to the extent of overlaps of their dictionary definitions.
Banerjee and Pedersen (2002) extended this notion to use WordNet as the dictionary for the word definitions. This notion was further extended to use the rich network of relationships between concepts present is WordNet.

3 Faktoren, die IAA beeinflussen

1. Spezifität des Kontexts
2. Konkretheit der Wortbedeutungen
3. Ähnlichkeit der Wortbedeutungen

   Aber - wie missst man die Ähnlichkeit von Wortbedeutungen?

- **Inter-Sense Similarity Measure (ISM)** *(Ide, 2006)*
- basiert auf dem **Lesk Ähnlichkeitsmaß**

Lesk (1985)

Relatedness of two words is proportional to to the extent of overlaps of their dictionary definitions.
Banerjee and Pedersen (2002) extended this notion to use WordNet as the dictionary for the word definitions. This notion was further extended to use the rich network of relationships between concepts present is WordNet.

Inter-Sense Similarity

Hypothese

Worte mit sehr ähnlichen Bedeutungen zeigen eine niedrigere Übereinstimmung in der Annotation

- Berechnung des ISM für alle Paare von Wortbedeutungen eines bestimmten Wortes w
- Berechnung eines Konfusionsgrenzwerts *(confusion threshold CT)*

$$CT_w = \mu ISM_w + \sigma ISM_w \quad (1)$$

wobei μISM_w = Erwartungswert für ISM_w
und σISM_w = Standardabweichung
ISM_w ist die Intersense Similarity für ein bestimmtes Paar von Wortbedeutungen für w

Inter-Sense Similarity

Hypothese

Worte mit sehr ähnlichen Bedeutungen zeigen eine niedrigere Übereinstimmung in der Annotation

- Berechnung des ISM für alle Paare von Wortbedeutungen eines bestimmten Wortes w
- Berechnung eines Konfusionsgrenzwerts *(confusion threshold CT)*

$$CT_w = \mu ISM_w + \sigma ISM_w \quad (1)$$

wobei μISM_w = Erwartungswert für ISM_w
und σISM_w = Standardabweichung
ISM_w ist die Intersense Similarity für ein bestimmtes Paar von Wortbedeutungen für w

Inter-Sense Similarity II

POS	Pairs	Max	Mean	Std. Dev	% $> CT$
long	36	0.71	0.28	0.18	0.17
fair	45	1.25	0.28	0.34	0.18
work	21	0.63	0.22	0.16	0.14
land	54	1.44	0.17	0.29	0.07
tell	28	1.22	0.15	0.25	0.07
show	66	1.38	0.18	0.27	0.12

- hohe Korrelation von IAA mit % $> CT$ (Anteil an Wortbedeutungen, die größer sind als der Konfusionsgrenzwert für dieses Wort) für Nomen, nicht aber für Verben und Adjektive
  $\rightarrow$ Hypothese nicht belegt
- Future work: Replikation der Studie mit mehr Daten

Übersicht

- Daten + Task
- 3 Faktoren des lexikalen Gebrauchs, die IAA beeinflussen
- Data Mining mit Assoziationsregeln (Association rules)

Data Mining mit Assoziationsregeln

- **Assoziationsanalyse**: Auffinden von Instanzen, die das Auftreten anderer Instanzen innerhalb eines Ereignisses wahrscheinlich machen
- **Anwendungsbeispiel**: Warenkorbanalyse
  (Kunden, die A gekauft haben, haben auch B gekauft)
- **Assoziationsregeln**: beschreiben Korrelationen zwischen gemeinsam auftretenden Instanzen

  wenn Instanz A, dann Instanz B oder $A \rightarrow B$
- Kenngrößen von Assoziationsregeln:
  - *Support:* relative Häufigkeit der Beispiele, in denen die Regel anwendbar ist
  - *Konfidenz:* relative Häufigkeit der Beispiele, in denen die Regel korrekt ist

Ziel

Auffinden von Mustern (systematischen Unterschieden) in der Wahl der Wortbedeutungen zwischen den verschiedenen Annotator/innen

Data Mining mit Assoziationsregeln

- **Assoziationsanalyse**: Auffinden von Instanzen, die das Auftreten anderer Instanzen innerhalb eines Ereignisses wahrscheinlich machen
- **Anwendungsbeispiel**: Warenkorbanalyse
  (Kunden, die A gekauft haben, haben auch B gekauft)
- **Assoziationsregeln**: beschreiben Korrelationen zwischen gemeinsam auftretenden Instanzen

  wenn Instanz A, dann Instanz B oder $A \rightarrow B$
- Kenngrößen von Assoziationsregeln:
  - *Support:* relative Häufigkeit der Beispiele, in denen die Regel anwendbar ist
  - *Konfidenz:* relative Häufigkeit der Beispiele, in denen die Regel korrekt ist

Ziel

Auffinden von Mustern (systematischen Unterschieden) in der Wahl der Wortbedeutungen zwischen den verschiedenen Annotator/innen

Data Mining mit Assoziationsregeln II

- Assoziationsregeln beschreiben Beziehungen zwischen Instanzen, basierend auf deren Attributen
- Mögliche Attribute:
  - Instanzen (Worte)
  - Annotator/innen ($A_1 - A_6$)
  - von den Annotator/innen zugewiesene Wortbedeutungen ($Sense1, Sense2, ..., SenseN$)
- Hauptinteresse auf bi-direktionalen Beziehungen zwischen Annotatoren:

$$A_i\_fair : Sense1 \rightarrow A_j\_fair : Sense2$$
$$A_j\_fair : Sense2 \rightarrow A_i\_fair : Sense1$$

Data Mining mit Assoziationsregeln III

- Evaluation der Assoziationsregeln mit Hilfe von Support und Konfidenz
- C, C_1, C_2: Bedingungen für Attribute
  - $\text{Supp}(C)$: Anteil an Instanzen, die Bedingung C erfüllen
  - $\text{Supp}(C_1 \Rightarrow C_2) = \text{Supp}(C_1)$
  - $\text{Conf}(C_1 \Rightarrow C_2) = \text{Supp}(C_1 \wedge C_2)/\text{Supp}(C_1)$

Gegeben seien die Grenzwerte MinSupp (minimum support) und MinConf (minimum confidence):
eine Regel ist überzeugend, wenn ihr Support größer ist als MinSupp und ihre Konfidenz größer als MinConf.

Data Mining mit Assoziationsregeln IV

- **Ziel**: finde Regeln, die Beziehungen zwischen Annotatorinnen beschreiben
  1. Vorverarbeitung der Daten: Erstellen einer 2-dimensionalen Tabelle jede Zeile beinhaltet ein Annotator_Sense-Paar
  2. für jedes Paar von Annotatoren: identifiziere gleiche Wortbedeutungen / systematische Unterschiede in der Zuweisung von Wortbedeutungen

Annotator_Sense-Paare	**Instanzen**				
	S_1	S_2	S_3	...	S_{50}
A_1 *Artefakt*	1	0	1	...	0
A_1 *Geschehen*	0	1	0	...	1
A_1 *Ort*	0	0	0	...	0
A_1 *Artefakt*	0	0	1	...	0
A_1 *Geschehen*	1	1	0	...	0
A_1 *Ort*	0	0	0	...	1
A_3 *Artefakt*	0	1	1	...	0
...	...	...	...	...	...
A_6 *Ort*	0	0	1	...	0

Data Mining mit Assoziationsregeln V

$Ann_i.S_j \Rightarrow Ann_m.S_n$		Supp(%)	Conf(%)
fair			
Sense 1 Agreements			
107.S1	101.S1	56.0	82.1
101.S1	107.S1	55.0	83.6
107.S1	105.S1	56.0	91.1
105.S1	107.S1	53.0	96.2
Disagreements			
107.S2	102.S1	56.0	28.6
102.S1	107.S2	31.0	51.6
105.S2	102.S1	53.0	24.5
102.S1	105.S2	31.0	41.9

- Ergebnisse der Assoziationsanalyse sind konsistent mit IAA:
  - IAA: *long* $>$ *fair* $>$ *quiet*(α : $0.67 > 0.54 > 0.49$)
  - *long*: größte Anzahl an Agreement-Assoziationsregeln mit Support $> 50\%$ (*long*: 22, *quiet*: 4, *fair*: 13)

Data Mining mit Assoziationsregeln V

$Ann_i.S_j \Rightarrow Ann_m.S_n$		Supp(%)	Conf(%)
fair			
Sense 1 Agreements			
107.S1	101.S1	56.0	82.1
101.S1	107.S1	55.0	83.6
107.S1	105.S1	56.0	91.1
105.S1	107.S1	53.0	96.2
Disagreements			
107.S2	102.S1	56.0	28.6
102.S1	107.S2	31.0	51.6
105.S2	102.S1	53.0	24.5
102.S1	105.S2	31.0	41.9

- Ergebnisse der Assoziationsanalyse sind konsistent mit IAA:
  - IAA: *long* $>$ *fair* $>$ *quiet*($\alpha : 0.67 > 0.54 > 0.49$)
  - *long*: größte Anzahl an Agreement-Assoziationsregeln mit Support $> 50\%$ (*long*: 22, *quiet*: 4, *fair*: 13)

Data Mining mit Assoziationsregeln VI

- Verhältnis zwischen Konfusionsmatrix und Assoziationsregeln?
  - beschreiben die gleichen Informationen
  - Konfusionsmatrix: nur möglich für 2 Annotatorinnen
  - Assoziationsregeln: bessere Interpretation durch *Support* und *Konfidenz*

Passonneau et al. (2010):

The main utility of the association rules is that they provide a more fine-grained analysis of the patterns of agreement and disagreement.

Data Mining mit Assoziationsregeln VI

- Verhältnis zwischen Konfusionsmatrix und Assoziationsregeln?
  - beschreiben die gleichen Informationen
  - Konfusionsmatrix: nur möglich für 2 Annotatorinnen
  - Assoziationsregeln: bessere Interpretation durch *Support* und *Konfidenz*

Passonneau et al. (2010):

The main utility of the association rules is that they provide a more fine-grained analysis of the patterns of agreement and disagreement.

Data Mining mit Assoziationsregeln VI

- Verhältnis zwischen Konfusionsmatrix und Assoziationsregeln?
  - beschreiben die gleichen Informationen
  - Konfusionsmatrix: nur möglich für 2 Annotatorinnen
  - Assoziationsregeln: bessere Interpretation durch *Support* und *Konfidenz*

Passonneau et al. (2010):

The main utility of the association rules is that they provide a more fine-grained analysis of the patterns of agreement and disagreement.

Zusammenfassung

- WSD-Experiment mit moderat polysemen Wörtern und mehr als 2 Annotator/innen
- Identifizierung von Faktoren, die IAA beeinflussen (Spezifität des Kontexts, Konkretheit der Wortbedeutungen, Ähnlichkeit der Wortbedeutungen)
- Assoziationsregeln zum Auffinden von systematischem Disagreement zwischen Annotator/innen

Referenzen

- Passonneau et al. 2009. Making Sense of Word Sense Variation
- Passonneau et al. 2010. Word Sense Annotation of Polysemous Words by Multiple Annotators
- Perl-Modul zur Berechnung von Lesk (implementiert von Ted Pedersen) `http://kobesearch.cpan.org/htdocs/WordNet-Similarity/WordNet/Similarity/lesk.pm.html`