Analyzing Disagreements

Beata Beigman Klebanov, Eyal Beigman and Daniel Diermeier

July 1, 2010

Analyzing Disagreements (Beigman Klebanov, Beigman, Diermeier)

- Hauptfrage: kann man auf der Instanzebene unterscheiden, welche Instanzen mit Disagreements das Resultat von Unachtsamkeit etc sind und welche die Folge von abweichendem Verständnis (der Richtlinien, des Kontexts, etc.)?

Analyzing Disagreements (Beigman Klebanov, Beigman, Diermeier)

- Hauptfrage: kann man auf der Instanzebene unterscheiden, welche Instanzen mit Disagreements das Resultat von Unachtsamkeit etc sind und welche die Folge von abweichendem Verständnis (der Richtlinien, des Kontexts, etc.)?
- Das Papier befasst sich mit einer Entdeckungs-Aufgabe (ähnlich: Sentiment!) und nicht mit einer Klassifizierungsaufgabe, bei der die zu annotierenden Instanzen vorgegeben sind.

Analyzing Disagreements (Beigman Klebanov, Beigman, Diermeier)

- Hauptfrage: kann man auf der Instanzebene unterscheiden, welche Instanzen mit Disagreements das Resultat von Unachtsamkeit etc sind und welche die Folge von abweichendem Verständnis (der Richtlinien, des Kontexts, etc.)?
- Das Papier befasst sich mit einer Entdeckungs-Aufgabe (ähnlich: Sentiment!) und nicht mit einer Klassifizierungsaufgabe, bei der die zu annotierenden Instanzen vorgegeben sind.
- Vorgehen:
  - aus allen Annotationen mit disagreements sollen die "reliably deliberate annotations" identifiziert werden, also die, von denen man annehmen kann, dass sich die Annotatoren dabei was gedacht haben.
  - die reliably deliberate annotations sind aber natürlich immer noch solche mit disagreements!

2 Arten von Annotationssaufgaben

- Klassifizierung vorgegebener Einheiten (z.B. WSD, POS)
- Entdeckung von Instanzen eines Phänomens
- Hier: Entdeckung von 4 Arten von Metaphern

Frage: wie soll man Disagreements verstehen?

- Wenn einige Personen etwas als Metapher markieren und andere nicht, ist das ein Zeichen von
  - Unterschiedlichem Verständnis
  - Unachtsamkeit ?

Daten

- 151 Artikel aus der britischen Presse
- 4 Metaphern (Love, Vehicle, Authority, Building)
- Annotationsebene: Paragraph (Enthält der Paragraph ein Vorkommen der Metapher?)

Experiment

- Production task: 9 Annotatoren
- Validation task: 7 Annotatoren
- 6 Wochen pro Aufgabe; 25 Texte pro Woche
- Insgesamt 2364 Paragraphen
- Metaphern sind von einander unabhängig: jede Metapher wird als eigene Annotationsaufgabe angesehen

Übereinstimmung der Annotatoren

- Ergebnisse pro Metapher

Typ	K	marked
VEHICLE	0.66	4.0%
LOVE	0.66	2.5%
AUTHORITY	0.39	2.7%
BUILD	0.43	1.7%

- Fazit: Annotationen insgesamt nicht zuverlässig

Reliably deliberate annotations

- Ziel: die annotierten Instanzen identifizieren, die mit hoher Wahrscheinlichkeit mit Überlegung annotiert wurden.
- Bezug auf Methode aus früherem Artikel, in dem Kohärenz untersucht wurde
  - Annahme: alle Annotatoren annotieren zufällig (mit individuellen Wahrscheinlichkeiten)
  - Wenn eine Instanz von sehr vielen Annotatoren markiert wird, ist das ein sehr unwahrscheinliches Ereignis $\Rightarrow$ solche Instanzen werden als nicht-zufällig angesehen!
  - Man kann berechnen, wieviele Annotatoren eine Instanz annotieren müssen, um auszuschließen, dass die Annotation zufällig erfolgt ist

Berechnung der erforderlichen Annotatorenanzahl für valide Instanzen

- jeder Annotator i, für i aus 1,...,20 hat eine Annotationswahrscheinlichkeit p_i
- S ist die Anzahl der Annotatoren, die eine bestimmte Instanz annotiert haben (1 bis 20)
- Erwartungswert für S ist: $E(S)=\sum_{i=1}^{20} p_i$
- Varianz von S: $V(S)=\sum_{i=1}^{20} p_i(1-p_i)$ [S hat eine binomiale Verteilung!]

Berechnung der erforderlichen Annotatorenanzahl für valide Instanzen II

- p(Val): ein Datenpunkt aus einer Normalverteilung mit gegebenem μ =E(S) und Standardabweichung $\sigma = \sqrt{V(S)}$ fällt in das Intervall (-inf(inity), Val]
- Wenn wir Val=0.5 setzen, berechnen wir die Wahrscheinlichkeit, dass kein Annotator die Instanz annotiert hat
- Wenn wir Val=12.5 setzen, berechnen wir die Wahrscheinlichkeit, dass weniger als 13 Annotatoren die Instanz annotiert haben
  - 1-p(12.5) ist die Wahrscheinlichkeit, daß mindestens 13 Annotatoren die Instanz annotiert haben
  - p(0.5) + (1-p(12.5)) ist die Wahrscheinlichkeit, daß kein Annotator die Instanz annotiert hat oder dass mindestens 13 Annotatoren die Instanz annotiert haben
  - das Ergebnis von p(0.5) + (1-p(12.5)) ist 0.0039, d.h. es ist viel kleiner als 0.01, die angestrebte Signifikanzschwelle

Unachtsamkeit versus unterschiedliches Verständnis

- In den Fällen, in denen die Annotationen absichtlich erzeugt wurden, wie soll man die negativen Annotationen auffassen?

Unachtsamkeit versus unterschiedliches Verständnis

- In den Fällen, in denen die Annotationen absichtlich erzeugt wurden, wie soll man die negativen Annotationen auffassen?
- Adaption eines Validierungsexperiments aus früherer Arbeit
  - Annotatoren bekommen alle Instanzen als annotiert zu sehen, die von überhaupt irgendjemand annotiert wurden

Unachtsamkeit versus unterschiedliches Verständnis

- In den Fällen, in denen die Annotationen absichtlich erzeugt wurden, wie soll man die negativen Annotationen auffassen?
- Adaption eines Validierungsexperiments aus früherer Arbeit
  - Annotatoren bekommen alle Instanzen als annotiert zu sehen, die von überhaupt irgendjemand annotiert wurden
  - "potentially agreeable items": Wenn Nicht-Annotationen nur auf Unachtsamkeit zurückzuführen sind, dann würden sie beim zweiten Blick im Validierungsexperiment akzeptiert

Unachtsamkeit versus unterschiedliches Verständnis

- In den Fällen, in denen die Annotationen absichtlich erzeugt wurden, wie soll man die negativen Annotationen auffassen?
- Adaption eines Validierungsexperiments aus früherer Arbeit
  - Annotatoren bekommen alle Instanzen als annotiert zu sehen, die von überhaupt irgendjemand annotiert wurden
  - "potentially agreeable items": Wenn Nicht-Annotationen nur auf Unachtsamkeit zurückzuführen sind, dann würden sie beim zweiten Blick im Validierungsexperiment akzeptiert
  - Annotatoren würden es aber ablehnen, Instanzen beim zweiten Sehen durchgehen zu lassen, bei denen sie sich beim ersten Annotieren mit Überlegung gegen die Annotation als Metapher entschieden haben.

Unachtsamkeit versus unterschiedliches Verständnis

- In den Fällen, in denen die Annotationen absichtlich erzeugt wurden, wie soll man die negativen Annotationen auffassen?
- Adaption eines Validierungsexperiments aus früherer Arbeit
  - Annotatoren bekommen alle Instanzen als annotiert zu sehen, die von überhaupt irgendjemand annotiert wurden
  - "potentially agreeable items": Wenn Nicht-Annotationen nur auf Unachtsamkeit zurückzuführen sind, dann würden sie beim zweiten Blick im Validierungsexperiment akzeptiert
  - Annotatoren würden es aber ablehnen, Instanzen beim zweiten Sehen durchgehen zu lassen, bei denen sie sich beim ersten Annotieren mit Überlegung gegen die Annotation als Metapher entschieden haben.
  - Zufällig ausgewählte Annotationen wurden mit ins Validierungsexperiment aufgenommen, um zu kontrollieren, ob die Annotatoren in der Validierung bereit sind, einfach alles zu akzeptieren, wie es vorgegeben wird

Unachtsamkeit versus unterschiedliches Verständnis

- In den Fällen, in denen die Annotationen absichtlich erzeugt wurden, wie soll man die negativen Annotationen auffassen?
- Adaption eines Validierungsexperiments aus früherer Arbeit
  - Annotatoren bekommen alle Instanzen als annotiert zu sehen, die von überhaupt irgendjemand annotiert wurden
  - "potentially agreeable items": Wenn Nicht-Annotationen nur auf Unachtsamkeit zurückzuführen sind, dann würden sie beim zweiten Blick im Validierungsexperiment akzeptiert
  - Annotatoren würden es aber ablehnen, Instanzen beim zweiten Sehen durchgehen zu lassen, bei denen sie sich beim ersten Annotieren mit Überlegung gegen die Annotation als Metapher entschieden haben.
  - Zufällig ausgewählte Annotationen wurden mit ins Validierungsexperiment aufgenommen, um zu kontrollieren, ob die Annotatoren in der Validierung bereit sind, einfach alles zu akzeptieren, wie es vorgegeben wird
- Im früheren Experiment wurden die obigen Annahmen/Vorhersagen bestätigt

Anwendung auf Metaphernannotation I

Subset	#	Acc	Subset	#	Acc
$\text{Rand}_{Vehicle}$	94	5%	Hum_V	194	73%
Rand_{Love}	56	6%	Hum_L	137	64%
$\text{Rand}_{Authority}$	62	12%	Hum_A	258	51%
$\text{Rand}_{Building}$	40	1%	Hum_B	126	68%
Rand	252	6%	Hum_V	715	62%

Table: % angenommener Instanzen in der Validierung

Anwendung auf Metaphernannotation II

Subset	#	Acc	Subset	#	Acc
$\text{Unrel}_{Vehicle}$	92	49%	Rel_V	102	94%
Unrel_{Love}	81	43%	Rel_L	56	95%
$\text{Unrel}_{Authority}$	218	42%	Rel_A	40	96%
$\text{Unrel}_{Building}$	86	55%	Rel_B	40	96%
Unrel	477	46%	Rel_V	238	95%

Table: % akzeptierte Instanzen in der Validierung [der in der Produktionsaufgabe annotierten Labels: 477+238=715]

Selbstakzeptanz

- Potentieller Einwand: viele Fälle von Akzeptanz stellen Beibehaltung des früheren Urteils dar
  $\Rightarrow$ Selbstakzeptanzfälle aussen vor lassen

Selbstakzeptanz

- Potentieller Einwand: viele Fälle von Akzeptanz stellen Beibehaltung des früheren Urteils dar
  $\Rightarrow$ Selbstakzeptanzfälle aussen vor lassen
- Konzentration auf accept-Urteile in Fällen, in denen die Annotatorin in der production task zunächst keine Metapher annotiert hatte

Selbstakzeptanz

- Potentieller Einwand: viele Fälle von Akzeptanz stellen Beibehaltung des früheren Urteils dar
  $\Rightarrow$ Selbstakzeptanzfälle aussen vor lassen
- Konzentration auf accept-Urteile in Fällen, in denen die Annotatorin in der production task zunächst keine Metapher annotiert hatte
- wenn von 7 Annotatoren X die Metapher in der production task annotiert hatten, dann sehen wir uns in der Validierung die anderen 7-X Anntoatorinnen an
- Datenset nun kleiner–184 aus 238 zuverlässigen Instanzen–da es Instanzen gibt, die von allen 7 Annotatoren gelabelt worden waren

Selbstakzeptanz

- Potentieller Einwand: viele Fälle von Akzeptanz stellen Beibehaltung des früheren Urteils dar
  $\Rightarrow$ Selbstakzeptanzfälle aussen vor lassen
- Konzentration auf accept-Urteile in Fällen, in denen die Annotatorin in der production task zunächst keine Metapher annotiert hatte
- wenn von 7 Annotatoren X die Metapher in der production task annotiert hatten, dann sehen wir uns in der Validierung die anderen 7-X Anntoatorinnen an
- Datenset nun kleiner–184 aus 238 zuverlässigen Instanzen–da es Instanzen gibt, die von allen 7 Annotatoren gelabelt worden waren
- Im Unreliable set liegt die Fremd-Accept-rate bei 41%, für das reliable set liegt sie bei ca. 91%

Selbstablehnung

- Wieviele Instanzen, die man in der production task als Metapher gelabelt hat, lehnt man bei Wiedervorlage ab?

Selbstablehnung

- Wieviele Instanzen, die man in der production task als Metapher gelabelt hat, lehnt man bei Wiedervorlage ab?
- Wenn es sich um eine Instanz handelt, die zum unreliable set gehört, dann liegt die Ablehnungsrate bei ca. 23%.
- Bei Instanzen, die im reliable set waren, liegt sie mit 4% viel niedriger. Umgekehrt: in 96% der Fälle bleiben Annotatoren bei den Daten im reliable set bei ihrem ersten Urteil

Selbstablehnung

- Wieviele Instanzen, die man in der production task als Metapher gelabelt hat, lehnt man bei Wiedervorlage ab?
- Wenn es sich um eine Instanz handelt, die zum unreliable set gehört, dann liegt die Ablehnungsrate bei ca. 23%.
- Bei Instanzen, die im reliable set waren, liegt sie mit 4% viel niedriger. Umgekehrt: in 96% der Fälle bleiben Annotatoren bei den Daten im reliable set bei ihrem ersten Urteil
- Daten bei denen Selbstablehnung vorliegt spiegeln entweder wieder, dass
  - die Annotation ein Fehler war
  - die Entscheidung über die Annotierbarkeit schwierig war

Fazit

- Anwendung der Methode zur Identifikation von zufälligem Noise bei der Identifikation und Annotation von Metaphern
- Festlegung eines *agreement threshold*, der zwischen Flüchtigkeitsfehlern und bewußter Nicht-Übereinstimmung unterscheidet (4 von 9 Annotator_innen für eine Reliabilität von 99.9%)
- Analyse der Daten hat gezeigt, dass
  - bewusste Annotionen konsistent sind (niedrige Selbstablehnungsrate)
  - Flüchtigkeitsfehler jedoch mit höherer Wahrscheinlichkeit während der Validierung abgelehnt werden
- Methode erlaubt darüber hinaus die Identifikation von "harten Fälllen" für die Annotation
  $\Rightarrow$ Änderung des Annotationsschemas nötig?