

"Isn't it obvious?" - Rhetorical Question Identification

Abu-Khader, Meier, Moslemi, Pisetta

Institut für Computerlinguistik
Ruprecht-Karls-Universität Heidelberg
Dozent: Herr Dr. Ruppenhofer und Frau Karimova
WS 2017/2018

November 30, 2017

Gliederung

1. Einführung
2. Fachlicher Bezug
3. Daten
 - 3.1 Switchboard Dialogue Act Corpus
 - 3.2 Sarcasm Corpus
 - 3.3 Inter-annotator Agreement
4. Projektablauf
 - 4.1 Feature Extraktion
 - 4.2 Klassifizierung
5. Evaluierung

Ziel

■ Erkennung von rhetorischen Fragen



"I didn't feel answers were necessary.
All the questions seemed rhetorical."

Automatic Identification of Rhetorical Questions

- Bhattasali et. al
- Relevante Aufgabe für Informationsextraktion und Textzusammenfassung
- N-Gram basiertes Modell um rhetorische Fragen im Switchboard Dialogue Act Corpus zu identifizieren

Feature set	Acc	Pre	Rec	F1	Error 95%
Question	92.41	35.00	60.16	44.25	7.59 \pm 1.02
Precedent	85.64	12.30	30.47	17.53	14.36 \pm 1.36
Subsequent	78.98	13.68	60.16	22.29	21.02 \pm 1.58
Question + Precedent	93.82	41.94	60.94	49.68	6.18 \pm 0.93
Question + Subsequent	93.27	39.52	64.84	49.11	6.73 \pm 0.97
Precedent + Subsequent	84.93	19.62	64.84	30.14	15.07 \pm 1.38
Question + Precedent + Subsequent	94.87	49.03	59.38	53.71	5.13 \pm 0.86

Switchboard Dialogue Act Corpus

Name

doc
sw00utt
sw01utt
sw02utt
sw03utt
sw04utt
sw05utt
sw06utt
sw07utt
sw08utt
sw09utt
sw10utt
sw11utt
sw12utt
sw13utt
README

sw_0001_4325	sw_0021_4168	sw_0041_4048
sw_0002_4330	sw_0022_4320	sw_0042_4060
sw_0003_4103	sw_0023_4341	sw_0043_4148
sw_0004_4327	sw_0024_4688	sw_0044_4177

- enthält Telefongespräche, die anhand von Transkriptionsregeln niedergeschrieben wurden
- Verzeichnis mit mehreren Ordnern
- von Hand annotierter Korpus
- frei verfügbare Daten
- Datengröße insgesamt 16,1MB

Switchboard Dialogue Act Corpus

- Korpus enthält ca. 500 rhetorische Fragen

Tag	Sprecher	Aussage	Text
qy	B.48	utt3:	Can you force somebody to be a good productive citizen? /
b	B.50	utt1:	Yeah. /
sv	B.50	utt1:	I don't think you can. /

Table: Ausschnitt aus dem Switchboard Dialogue Corpus

Relevante Tags

Tag	Bedeutung	Beispiel
qh	Rhetorische Frage	Ist das nicht offensichtlich?
qy	Ja-Nein-Frage	Hast du gerade Zeit?
qy^d	Deklarative Ja-Nein-Frage	So you can afford to get a house
qw	W-Fragen	Wo ist die nächste Haltestelle?
qw^h	Deklarative W-Fragen	You are what kind of buff?
bh	Nachfrage	Ist das richtig?
qo	Offene Fragen	Was ist mit dir?
^g	Tag-Frage	Richtig?

Table: Tags für verschiedene Fragetypen

- Später: One vs. Rest Klassifizierung

Switchboard Auszug

aa B.108 utt1: I know. /
qh B.108 utt2: Who has time? /
sd B.108 utt3: I don't have time to sit, I know. /

sd A.11 utt2: {C and } that is all we have been doing all weekend. /
qh B.12 utt1: You know what? /
sv B.12 utt2: It seems like we are doing it here forever. /

b A.3 utt1: Yeah. /
qh A.3 utt2: Who has spare time <laughter>? /
x B.4 utt1: <Laughter>.

b B.60 utt1: Yeah. /
qh A.61 utt1: {D You know. } /
b B.62 utt1: Yeah. /

Switchboard Extraktion

I know.

Who has time?

I don't have time to sit, I know.

C and that is all we have been doing all weekend.

You know what?

It seems like we are doing it here forever.

Yeah.

Who has spare time <laughter>?

<Laughter>.

Yeah.

D You know.

Yeah.

Sarcasm V2 Corpus

- Subset des Internet Argument Corpus(IAC)
- Text mit zugehöriger Antwort (quote-response)
- Enthält drei Arten von Sarkasmus:
 - Genereller Sarkasmus (ca. 3260 Einheiten, 1630 sarkastisch)
 - Hyperbeln (ca. 582 Einheiten, 291 sarkastisch)
 - Rhetorische Fragen (c.850 Einheiten, 425 sarkastisch)

Auszug Sarcasm V2 Corpus

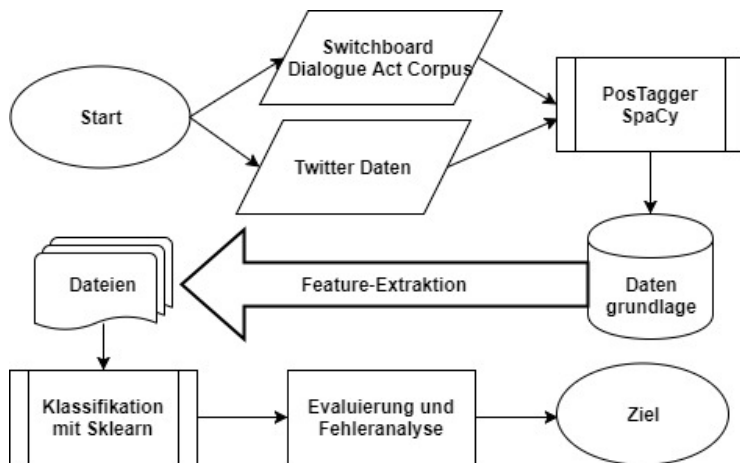
RQ, sarc, RQ_sarc0000, if the founders and early government leaders believed that the modern liberal interpretation of separation of church and state is correct then why is the image of moses carrying the ten commandments carved into the supreme court building? ten commandments stunner: feds lying at the supreme court, "ummmmmm, because it was built in 1932 – 1935, and has nothing whatsoever to do with the founders or early government leaders, nor with what they believed was correct? you really are the master of the art of self-pwn, archie."

RQ, sarc, RQ_sarc0001, "So Stalin didn't actually practice Communism (not that anyone has), but really just used bits and pieces to further his rule and such ideology us not communism but Stalinism. So on essence, Stalin wasn't a commie.", You seem to require a slick answer to a complicated question - so how about this: Question: Was Stalin a communist ? Answer: That's what he said he was. Hope this is glib and vacuous enough to meet with your exacting standards.

Inter Annotator Agreement

- 25 rhetorische und 25 nicht-rhetorische Fragen aus Switchboard
- zufällig ausgewählt, Tags entfernt, annotiert nach: rhetorisch - nicht-rhetorisch
- Kappa Score von: 0.23 nur Frage allein; 0.45 mit Aussage davor und danach

Projekttablauf



Eigenschaften rhetorischer Fragen

- Frage, auf die keine Antwort erwartet wird (Meinungsgleichheit)
- Dient somit nicht Informationsgewinn
- Verstärkung einer Aussage
- Suggestivfrage
- Du willst doch bestimmt auch wie alle anderen studieren gehen?
- Sehe ich in diesem Kleid nicht wirklich super aus?

Arten der rhetorischen Fragen

1. Implizit negativ formulierte Fragen, die das Gegenteil meinen: *"Für was soll das denn gut sein?"*
2. Fragen, die quasi schon eine Antwort auf eine andere Frage beantworten: *"Was ist dein Name?" "Seh ich so aus, als ob ich antworten will?"*
3. Fragen, für die es implizite Antworten gibt: *"Who is your daddy?"*
4. Blamierende oder kritisierende Fragen: *"Was denkst du eigentlich, was das ist?"*
5. Der Sprecher drückt seinen Unglauben in der Frage aus: *"Was macht die Fliege in der Suppe?"*

Features

- Unigram
- Bigram
- Pos-N-Grams
- Strong-negative-polarity-items (NPIs), in years, lift a finger
- Modal auxiliary, "could", "would"
- Bestimmte Ausdrücke wie "yet", "after all"

Pos-Tagging

SpaCy-Ausgabe

you PRON PRP
need VERB VBP
to PART TO
do VERB VB
it PRON PRP

Tri-Grams

['DT', 'NN', 'VBP'], ['NN', 'VBP', 'PRP'], ['VBP', 'PRP', 'IN'],
['PRP', 'IN', '.'], ['IN', '.', 'IN'], ['.', 'IN', 'NN'], ['IN', 'NN', 'DT']...

Aufteilung in Featuresets

- Fragen ohne Kontext
- Fragen mit davorstehender Äußerung
- Frage mit dahinterstehender Äußerung
- Frage mit beiden Kontexten

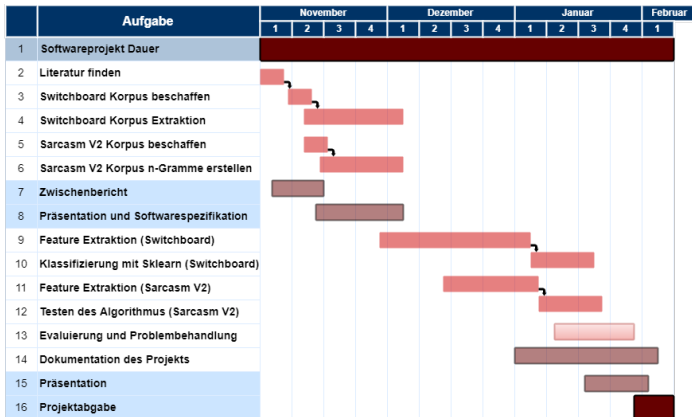
Klassifikation

- Klassifikation mit Naive Bayes und SVM in SKLearn
- Eingabe der Features als Counts
- Training mithilfe von Cross Validation

Evaluierung

- klassische Maße Accuracy, Precision, Recall & F1-Measure
- Baseline: Set, das aus der Frage allein besteht
- Beachtung von false positives bei unterschiedlichen Datensets

Zeitplan



Optionale Ziele

NPI als eigenes Feature

Lexikon mit zwei Dateien:

1. Minus Plus NPIs
2. Plus Minus NPIS

weitere Möglichkeiten

Anwendung des Algorithmus auf Twitter-Daten

Literatur

BHATTASALI, Shohini; CYTRYN, Jeremy; FELDMAN, Elana & PARK, Joonsuk, "Automatic Identification of Rhetorical Questions", Cornell University, 743-749, 2015.

ORABY, Shereen; HARRISON, Vrindavan; REED, Lena & HERNANDEZ, Ernesto; RILOFF, Ellen; WALKER, Marilyn; "Creating and Characterizing a Diverse Corpus of Sarcasm in Dialogue", University of California, Santa Cruz, 31-41, 2013.

RANGANATH, Suhas; HU, Xia; TANG, Juliang; WANG, Suhang; LIU, Huan; "Identifying Rhetorical Questions in Social Media", Arizona State University, 667-670, 2017.

BAMMAN, David; SMITH, Noah A.; "Contextualized Sarcasm detection on Twitter", Carnegie Mellon University, 574-577, 2015.

LU, Yujie; SAKAMOTO, Kotaro; SHIBUKI, Hideyuki; MORI, Tatsunori; "Construction of a Multilingual Annotated Corpus for Deeper Sentiment Understanding in Social Media", Carnegie Mellon University, 205-265, 2017