Fragen auf Twitter?

Johannes Gontrum & Steve Wendler

24. März 2015

Your abstract.

Inhaltsverzeichnis

1 Einleitung

Texte als sprachliche Einheit unterliegen im Gegensatz zu Phrasen oder Einzelsätzen einer pragmatischen Anreicherung: sie folgen argumentativen Pfaden. Das bedeutet, dass Sätze in ihrer linearen Abfolge Bezug nehmen auf vorangegangene oder auch folgende Sätze. Diese Bezugnahmen bzw. Relationen können hinsichtlich unterschiedlicher Aspekte aufeinander verweisen: so kann ein Satz die Begründung für die Satzaussage eines anderen Satzes darstellen oder aber auch ein Gegenargument. Ebenso kann der Sprecher/Schreiber seine Einstellung zum

2 Daten und Vorverarbeitung

2.1 Generierung der Dialogstrukturen

Als Grundlage für die Untersuchungen dient ein Korpus von etwa 20,6 Millionen deutschen Tweets (exklusive Retweets), die im April 2013 gesammelt wurden. Dieses Korpus wurde mit Hilfe einer deutschen Stoppwortliste durch die Twitter Streaming API generiert und anschließend durch LangID gefiltert, sodass Einträge entfernt wurden, die nicht in deutscher Sprache verfasst wurden.

Um Anfragen an dieses Korpus zu vereinfachen, haben wir alle Tweets in eine MySQL Datenbank übertragen und für jeden Tweet eine Liste von direkten und indirekten Antworten erzeugt. Eine direkte Antwort ist dabei ein Tweet, der sich direkt auf einen anderen bezieht, während indirekte Antworten transitiv auch Tweets bezeichnen, die auf eine Nachricht antworten, die wiederum eine Antwort auf den Basistweet ist. Zur schnelleren Analyse haben wir ebenfalls die Anzahl der direkten und indirekten Antworten in die Datenbank mit aufgenommen.

Ebenso war es uns wichtig, Tweets zu markieren, die einen Dialog starten. Diese Basistweets beziehen sich auf keine anderen Tweets, es gibt jedoch Nachrichten, die auf diese Basistweets antworten. Die Identifizierung setzt jedoch voraus, dass NutzerInnen sich bei einer Antwort direkt auf einen Tweet beziehen und nicht manuell eine Nachricht mit einem @-Handle verfassen. So wurde der Tweet „@DeigningDiamond – wenigstens etwas, das ich machen konnte, ohne dass du es auch nur merken konntest.]“ (ID: 318484529570529281) von unserem System als Basistweet markiert, betrachtet man aber den Kontext, wird klar, dass er eigentlich Teil einer Diskussion ist. Diese fälschlich markierten Tweets können leider nicht vermieden werden, da es nicht ungewöhnlich ist, einen Dialog mit einem an eine Userin / einen User gerichteten Tweet beginnt:

Beginn: „@_danjl Dein Bild ist richtig scheiße“ (ID: 318482949844660224)
Antwort: „@chrisgoescross Welches soll ich denn sonst nehmen?“ (ID: 318483248978219008)

2.2 Filterung automatisch generierter Tweets

Um die statistische Auswertung nicht zu verzerren, haben wir insgesamt 25736 Tweets entfernt, die eindeutig automatisch generiert wurden. Dazu zählen z.B. Benachrichtigungen aus Videospielen, Musik-Updates oder Foursquare-Mitteilungen. Es wurden alle Tweets als ungültig markiert, die eines der Folgenden Tokens enthalten: '@YouTube', 'Gutschein', '#4sq', '#androidgames', '#nowplaying', '#np', 'Verkehrsmeldungen' und 'Wetterdaten'.

2.3 Identifikation von Fragen

Da es bei einer so großen Datenmenge aus Geschwindigkeitsgründen nicht möglich ist, Fragen akkurat und linguistisch korrekt zu identifizieren, mussten wir uns einer Näherungslösung bedienen.

Liste der W-Wörter, etwas mehr zur Theorie?

100 zufällig ausgewählte Tweets, die als wh$_q$uestion getagged sind:
Falsch 27, Richtig 73

100 zufällig ausgewählte Tweets, die als question$_m$ark getagged sind:
Falsch 0, Richtig 100

100 zufällig ausgewählte Tweets, die nicht als Frage getagged sind:
Falsch 0, Richtig 100

Schönere Statistik

=> question_mark ist zuverlässiger Tagger

2.4 Einteilung der UserInnen anhand der Anzahl ihrer Follower

Während des untersuchten Zeitraums zwischen dem 1. April und dem 30. April 2013 waren 1.577.083 unterschiedliche Accounts auf Twitter aktiv.

3 Beantwortung von Fragen