1 Daten und Vorverarbeitung

1.1 Generierung der Dialogstrukturen

Als Grundlage für die Untersuchungen dient ein Korpus von 20.733.071 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 das komplette Korpus in eine MySQL Datenbank übertragen und in einem mehrstufigen Prozess für jeden Tweet eine Menge von Attributen erstellt. Dabei wurden einige Attribute direkt aus den JSON-Files der Twitter Streaming API ausgelesen, andere durch neuerliche Iterationen erzeugt.

Twitter Streaming API

- `id_str` – ID des Tweets
- `user_id_str` – ID der Users
- `friend_count` – Anzahl der Personen, dennen Tweet-Verfassers folgt
- `followers_count` – Anzahl der Follower des Tweet-Verfassers
- `in_reply_to_status_id_str` – ID des Tweet auf den aktueller Tweet Antwort ist
- `in_reply_to_user_id_str` – ID des Users des Tweet auf den aktueller Tweet Antwort ist
- `user_mentions_count` – Anzahl der User, die im Tweet erwähnt werden
- `user_mentions` – Liste der IDs der User, die im Tweet erwähnt werden
- `created_at` – Zeitpunkt des Erstellens des Tweet
- `source` – von welcher Platform wurde Tweet gesendet (Web, Smartphone)
- `text` – Textkörper des Tweet

nachbearbeitete Attribute

- `direct_replies_count` – Anzahl der Tweets, die direkt auf den aktuellen Tweet antworten
- `direct_replies_list` – Liste der Tweets, die direkt auf den aktuellen Tweet antworten
- `indirect_replies_count` – Anzahl der Tweets, die im Dialog-Baum unterhalb des aktuellen Tweet hängen
- `indirect_replies_list` – Eine Liste von allen Antworten, die sich auf diesen Tweet bezogen
- `is_base_tweet` – Root-Tweet eines Dialoges
- `is_question` – Tweet ist eine/ keine Frage
- `question_mark` – Anzahl der Fragezeichen eines Tweets
- `is_wh_question` – Tweet beginnt mit einem w-Fragewort
- `valid` – Tweet ist kein Spam.

Auf Grundlage dieser Informationen war es uns möglich, alle Dialoge in jeweils einem *Dialogbaum* abzubilden, um so Fragen und Antworten zu ermitteln. Dieser Dialogbaum wurde aufgebaut, indem alle Tweets auf ihr Attribut `in_reply_to_status_id_str` hin geprüft wurden und so der Tweet, zu dem der aktuelle Tweet eine Antwort darstellt, mit dem aktuellen verbunden werden konnte.

1.2 Basistweets

Ebenso war es uns möglich, 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

(1) @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:

(2) *Beginn*: @_danjl Dein Bild ist richtig scheiße (ID: 318482949844660224)

Antwort: @chrisgoescross Welches soll ich denn sonst nehmen? (ID: 318483248978219008)

1.3 Direkte und indirekte Antworten

Eine direkte Antwort ist ein Tweet, der sich direkt auf den Tweet bezieht, auf welchen er mit seinem Attribut `in_reply_to_status_id_str` zeigt, 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.

1.4 Filterung automatisch generierter Tweets

Um die statistische Auswertung nicht zu verzerren, haben wir insgesamt 25.736 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'. In diesem Fall wurde das Attribut `valid` auf 0 (false) gesetzt.