3. Vorverarbeitung von Twitterdaten und Technisches

Seminar

Computerlinguistische Analyse von Twitterdaten

Tatjana Scheffler, Universität Potsdam

tatjana.scheffler@uni-potsdam.de

24.4.2013

Heute

- Finalisierung des Zeitplans
- Vorverarbeitung von deutschen Twitterdaten
- Technisches beim Arbeiten mit Twitter

Hausaufgabe

Weitere Seminarplanung

Sie sind gefragt!

Zeitplanung

8.5.	Isa	Normalisierung (HC&B 2012)
15.5.	Matthias	Topikerkennung (Karandikar 2010)
22.5.	Johannes	Trenderkennung (B&K 2012)
29.5.	Katarina	Sent.Ana. I (BFBLHH 2011)
5.6.	Anna	Sent.Ana II (P&P 2010)
12.6.	Ulf	Sent.Ana III (DT&R 2010)
19.6.	Frank	Conv. Retr. (MM&R 2012)
26.6.	Steve	Lokationsabhängigkeit (Ara. 2011)
3.7.	Kurzvorstellung der Projekte (alle)	
10.7.	Abschlussdiskussion, Zusammenfassung	

Vorträge

Vorverarbeitung

Bereitgestellt von Wladimir Sidorenko, Uni Potsdam

Warum Vorverarbeitung?

```
uuund der akku hält und hält....super :) #iphone4s

Der Tagesspiegel: Busemann: Keine Weisung an
Staatsanwaelte in Wulffff-Affaere - http://t.co/
Xef3vrUj #Pressemitteilung
```

- Fehler
- Normalisierung
- Performanz der nachfolgenden Module: Tagger, Parser, Maschinelles Lernen

Vorarbeiten zur Textnormalisierung

- Alegria (2008)
- Aw (2006)
- Beaufort (2010)
- Brody Diakopoulos (2011)
- Choudhury (2007)
- Clark and Araki (2011)
- Cook and Stevenson (2009)
- Han and Baldwin (2011)
- Kaufmann and Kalita (2010)
- Kobus (2008)
- Krawczyk (2009)
- Kukich (1992)
- Liu (2011)
- Melero (2012)
- Oliva (2012)
- Sproat (2001)
- Toutanova and Moore (2002)
- Wei (2011)
- Yvon (2010)

Klassifikation der Ansätze zur Textnormalisierung

Operationsebene:

- Grapheme / Phoneme
- Wörter
- Phrasen

Methodologie:

- Regelbasiert
- Statistisch
- Hybrid

Textnormalisierung – Schritte

- Noise Cleaner
- Umlaut Restorer
- Slang Normalizer
- Character Squeezer
- Sentence Splitter
- Tokenizer

(SocMedia-Projekt, bearbeitet von Wladimir Sidorenko)

Noise Cleaner

- Entfernung oder Ersetzung Twitter-typischer Elemente: Emoticons, @Handles, #Tags

```
@_0816_ Das ist lieb von Dir :-)
Ich suche weiter...in der Kueche*hihi* ;-)

Das ist lieb von Dir %PosSmiley
Ich suche weiter...in der Kueche%PosSmiley %PosSmiley
REPLACED     0      0                          @_0816_
REPLACED     21     10       %PosSmiley        :-)
REPLACED     62     10       %PosSmiley        *hihi*
REPLACED     73     10       %PosSmiley        ;-)
```

Umlautwiederherstellung

Der israelisch-palaestinensische Konflikt ist ein Konflikt um Land, die Sicherheit von Grenzen und um die Staatlichkeit zweier Nationen.

Der israelisch-palästinensische Konflikt ist ein Konflikt um Land, die Sicherheit von Grenzen und um die Staatlichkeit zweier Nationen.

Slangnormalisierung

- Im Englischen wichtiger?
  do tngrs luv 2 txt msg?

```
@rmmarchy Und da isser wieder, der VIP-Vertrag
für's Bobby Car. cdu wulff
```

```
@rmmarchy Und da ist er wieder, der VIP-
Vertrag für das Bobby Car. cdu wulff
```

Character Squeezer

```
Ichhh haaasssseee diiieeseeen Tiiisch

Ich hasse diesen Tisch
```

- “Hase” oder “hasse”?
- dictionary lookup für alle Zeichenketten mit drei oder weniger gleichen Buchstaben in Folge: Ichhh, Ichh, Ich
- Alle gefundenen Varianten bleiben erhalten

Satzgrenzenerkennung

```
Dieter Golombek, Jurysprecher Dt. Lokalj.-
Preis: "Wir brauchen erklärenden Journalismus
mehr denn je." drehscheibe.org/weblog/?p=3926
#video #bpbwahl
```

```
Dieter Golombek, Jurysprecher Dt. Lokalj.-
Preis: "Wir brauchen erklärenden Journalismus
mehr denn je."<sentence/>
drehscheibe.org/weblog/?p=3926<sentence/>
#video #bpbwahl<sentence/>
```

Tokenisierung

```
@_0816_ Das ist lieb von Dir :-) Ich suche
weiter...in der Kueche*hihi* ;-)
```

TreeTagger Tokenizer:

```
@_0816_ Das ist lieb von Dir :- ) Ich suche
weiter ... in der Kueche*hihi* ; - )
```

Custom Tokenizer:

```
@_0816_ Das ist lieb von Dir :-) Ich suche
weiter ... in der Kueche *hihi* ;-) <sentence/>
```

Technisches & Praktisches

Probleme bei der Benutzung von Twitterdaten, Python, etc.

Projektideen?

- Dialoge auf Twitter
- Spamerkennung
- Generierung von Hashtags
- Analyse eines linguistischen Phänomens (z.B. weil-V2, Genreanalysen)
- Identifikation von ortsrelevanten Wörtern
- Reguläre Hausarbeit – Vergleich von Ansätzen, Evaluierung (z.B. spezieller Fokus auf deutsche Daten)

Signifikantes Vokabular

- Chi-squared-Test zum Vergleich von Subkorpora
- Tag (8am to 6pm) vs. Nacht (6pm to 8am)
  - Tag: verschiedene Genres, Themen
  - Nacht: schlafen, TV, Chats
- Tag1 vs. Tag2
- Trends verfolgen?
- Echtzeit?

Fragen

- tatjana.scheffler@uni-potsdam.de
- Sprechzeiten:
  Dienstags, 10-12 Uhr und nach Vereinbarung
  Haus 14, Raum 2.33
  Bitte per Email voranmelden!
- Aktuelle Informationen, Literatur, etc. auf der Webseite:

  http://www.ling.uni-potsdam.de/~scheffler/teaching/2013twitter.html