Computerlinguistische Analyse von Twitterdaten

Tatjana Scheffler
10.4.2013

Organisatorische Details:

BSc Computerlinguistik/Linguistik, Universität Potsdam
Sommersemester 2013
Seminar, AM4
Kurshomepage: http://www.ling.uni-potsdam.de/~scheffler/teaching/2013twitter.html

Termine:

Mittwochs, 12-14 Uhr - Haus 5/106

Kursbeschreibung:

Soziale Medien wie Twitter bieten neue Datenquellen für linguistische Analysen. Erste Arbeiten existieren zur Verarbeitung von Twitterdaten und deren sprachwissenschaftlicher Betrachtung, beziehen sich allerdings fast ausschließlich auf englische Daten: Diese sind zahlreicher, leichter zu erhalten, und oft auch leichter zu analysieren. Zahlreiche Tools, z.B. ein dezidierter Twitter-Tagger existieren schon für die Verarbeitung von englischsprachigen Social Media-Daten. In dem Blockseminar soll die computerlinguistische Analyse von deutschen Tweets anhand von vorhandenen und neu gecrawlten Daten gemeinsam erarbeitet werden. Vorverarbeitungsskripte sind vorhanden und können angepasst werden. Mögliche Themen sind Stimmungsanalyse, Themenklassifizierung, die Erstellung von Subkorpora, lexikalische Studien (Zeit- oder Ortsbezug von Wörtern) und vieles mehr.

Voraussetzungen:

möglichst Kenntnis einer Skriptsprache (Perl, Python)

Anforderungen/Leistungsnachweise:

Präsentation
Projekt und Ausarbeitung

Leseaufgabe für nächste Woche:

Wisdom/BuzzFeed (6.1.2013): “How Twitter gets in the way of research”:
http://www.buzzfeed.com/nostrich/how-twitter-gets-in-the-way-of-research

Themen:

Vorverarbeitung, Säuberung	Han, Cook, Baldwin, 2012: "Automatically Constructing a Normalisation Dictionary for Microblogs" Proc. of EMNLP www.cs.toronto.edu/~pcook/Hanetal2012.pdf Petrovic, Osborne, Lavrenko: "The Edinburgh Twitter Corpus" (deprecated) Tokenizer, Emoticons: https://github.com/brendano/tweetmotif
Topikerkennung	O'Connor, Krieger, Ahn, 2010: "TweetMotif: Exploratory Search and Topic Summarization for Twitter" https://github.com/brendano/tweetmotif http://anyall.org/oconnor_krieger_ahn.icwsm2010.tweetmotif.pdf http://brenocon.com/blog/2009/05/announcing-tweetmotif-for-summarizing-twitter-topics-with-a-dash-of-nlp/ Kireyev, Palen, Anderson, 2009: "Applications of Topics Models to Analysis of Disaster-Related Twitter Data" www.umiacs.umd.edu/~jbg/nips_tm_workshop/15.pdf Karandikar, 2010: "Clustering short status messages: A topic model based approach" http://ebiquity.umbc.edu/get/a/publication/518.pdf
Trenderkennung und -verfolgung	Mathioudakis, Koudas, 2010: "TwitterMonitor: Trend Detection over the Twitter Stream" http://www.inf.utfsm.cl/~mmendoza/descargas/p1155-mathioudakis.pdf Benhardus, Kalita, 2012: "Streaming Trend Detection in Twitter" http://www.cs.uccs.edu/~jkalita/papers/2012/BenhardusJamesIJWBC2012.pdf Becker, Naaman, Gravano, 2011: "Beyond Trending Topics: Real-World Event Identification on Twitter" http://academiccommons.columbia.edu/download/fedora_content/download/ac:135416/CONTENT/cucs-012-11.pdf
Tonalitätsanalyse	Meinungsbild der Zielgruppe (Sentiment Analysis) Pak, Paroubek, 2010: "Twitter as a Corpus for Sentiment Analysis and Opinion Mining" Proc. of LREC http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf Davidov, Tsur, Rappoport, 2010: "Enhanced Sentiment Learning Using Twitter Hashtags and Smileys" www.aclweb.org/anthology/C10-2028 Barbosa, Feng, 2010: "Robust Sentiment Detection on Twitter from Biased and Noisy Data" www.aclweb.org/anthology/C10-2005 Brown, Frazee, Beaver, Liu, Hoyt, Hancock, 2011: "Evolution of Sentiment in the Libyan Revolution" Blogpost: http://languagelog.ldc.upenn.edu/nll/?p=3537 Working Paper https://webspace.utexas.edu/dib97/libya-report-10-30-11.pdf

Soziolinguistik, Stil, Variabilität	Linguistics of Retweets http://danzarrella.com/retweet-linguistics.html# Bamman, Eisenstein, Schnoebelen, 2012: “Gender in Twitter: Styles, stances, and social networks” http://arxiv.org/abs/1210.4567 Schnoebelen, 2012: “Do You Smile with Your Nose? Stylistic Variation in Twitter Emoticons” http://repository.upenn.edu/pwpl/vol18/iss2/14/
Profiling	Erkenne Meinungsführer und Multiplikatoren Weng, Lim, Jiang, He, 2010: “TwitterRank: finding topic-sensitive influential twitterers” http://dl.acm.org/citation.cfm?id=1718520
IR/DR	Zanzotto, Pennacchiotti, Tsioutsiouliklis, 2011: “Linguistic Redundancy in Twitter” Proc. of EMNLP www.aclweb.org/anthology/D11-1061 Magnani, Montesi, Rossi, 2012: “Conversation retrieval for microblogging sites” http://link.springer.com/article/10.1007/s10791-012-9189-9/fulltext.html
Weitere	Semantic Role Labelling Conversation Modelling

Mögliche Projekt-/Ausarbeitungsthemen:

- Spamerkennung in Tweets
- Verbesserte Tweet-Suche (z.B. durch Synonyme)
- Tonalitätsanalyse (z.B. “2013” vor/nach Silvester)
- Übertragung eines der behandelten Themen auf deutsche Twitterdaten
- … (eigene Ideen)

Ressourcen und Links:

Twython: http://www.ling.uni-potsdam.de/~scheffler/twitter/index.html

Twitter Developer API https://dev.twitter.com/

Mögliche Quellen für weitere Literatur:
ICWSM 2012 http://www.icwsm.org/2012/
ICWSM 2011 http://www.icwsm.org/2011/
ACL Anthology http://aclweb.org/anthology-new/ (z.B. LREC, ACL conferences)
Twitter Research Bibliography http://www.danah.org/researchBibs/twitter.php (lückenhaft für CL)
TREC Microblog Track http://trec.nist.gov/pubs/trec20/t20.proceedings.html
Language Log on “Twitter Linguistics” http://languagelog.ldc.upenn.edu/nll/?p=3536

Interactive Python Tutorial http://www.learnpython.org/